

**An Evidence-Based Framework for Evaluating
Generative AI in Cultural Heritage Reconstruction:
Bamiyan, Palmyra and Bounded Generalisation**

**A Thesis Submitted for the Degree of Doctor of
Philosophy**

By

Kawsar Arzomand

Department of Computer Science

2026

Abstract

Generative AI can transform textual, photographic, archaeological and conservation records into visual reconstruction candidates for damaged or destroyed cultural heritage. This creates an opportunity to visualise aspects of heritage appearance that were previously difficult to reconstruct from dispersed evidence alone. The evaluative problem is acute where monuments can no longer be directly recorded and reconstruction depends on incomplete photographic, archaeological and conservation evidence. Under these conditions, a generated image may appear coherent while its evidential basis, prompt construction, interpretive assumptions and cultural acceptability remain insufficiently documented. The research asks how the evaluation of GenAI reconstruction candidates can be standardised before such outputs are treated as heritage reconstruction.

To address this problem, the research develops the Heritage-Aligned Reconstruction Framework (HARF) for evaluating GenAI heritage reconstruction. HARF is formalised through the Dynamic Prompt Blueprint, the Prompt Sufficiency Index, paradata, computational pre-evaluation and structured expert review. The framework is operationalised through the Western Buddha of Bamiyan, where 245 generated candidates are assessed through computational similarity, interdisciplinary screening and specialist archaeological evaluation. The Bamiyan case is extended through a community-grounded authenticity study with 62 Afghan respondents. HARF is re-instantiated at the Temple of Bel, Palmyra, to test bounded cross-site generalisation from figural sculpture to monumental architecture.

The thesis makes four contributions. It establishes an evidence-constrained workflow for GenAI heritage reconstruction; formalises HARF as an integrated AI-specific evaluation framework; shows that reconstruction validity at Bamiyan is distributed across computational, specialist and community-grounded registers; and demonstrates that HARF can be transferred to Palmyra at the level of evaluative schema without redesigning its top-level structure. The Palmyra test also shows that failure domains shift with typology, from figural micro-grammar at Bamiyan to architectural and decorative micro-grammar at Palmyra.

The contribution lies not in producing definitive reconstructions of Bamiyan or Palmyra, but in establishing how generated reconstruction candidates should be evaluated before they are treated as heritage reconstruction. HARF provides a structured procedure for assessing whether such candidates are evidence-based, documented, expert-assessed, culturally accountable and bounded in transfer.

Thanks

First and above all, I thank Allah, the Most Merciful, for giving me the strength, patience and protection to reach this point. This PhD required endurance, faith and the ability to continue even when the path felt uncertain. Alhamdulillah for the guidance, resilience and blessings that made this thesis possible.

I begin with my family, because this thesis began long before the PhD itself.

My deepest thanks are to my mother, Najiba. One of my earliest memories of education is not inside a classroom, but outside one. When I was eleven years old, in eighth grade, younger than most of my classmates and studying at an international school where only a few students could pass the three entrance examinations, she waited outside for me every day so that I would feel safe and not alone. In winter, rain and snow, she sat on a stone from seven in the morning until five in the afternoon. At the time, I did not fully understand the weight of that sacrifice. I understand it now. This thesis is built on that patience, protection and endless love.

My deepest thanks are also to my father, Prof. Akbar Jan, who made his daughters' education one of the great purposes of his life. He accepted the pain of distance from his country so that his daughters could study, grow and reach places that were not easily imaginable in the world from which we came. To the best of my knowledge, I am the first person among my relatives, on both my mother's and father's sides, to complete a PhD. This achievement is therefore not mine alone. It belongs first to my parents, who carried this dream before I knew how to name it.

I thank my sisters, Nadia, Sweeta and Mariam, who are far in distance but always close to my heart. Their support reached me across borders, time zones and difficult days. They may not always have known the details of my research, but they always understood when I needed strength, encouragement and love.

I owe a special thanks to my younger sister, Lima, who lived with the daily reality of this PhD more closely than anyone else. She saw the stress, deadlines, silence, late nights and the days when I could not easily explain what I was carrying. In her own gentle way, she brought lightness into difficult days, reminded me to laugh, prayed for me, encouraged me and stood beside me in both my lowest and happiest moments. Her presence gave me comfort, warmth and strength when I needed them most.

To my two brothers, Sameer and Shafi, thank you for being my constant support system. Whether it was midnight, whether you were tired, or whether you were fasting, you were there. You became my drivers, protectors, problem-solvers and emotional support, without ever making me feel like a burden.

I also thank Vika, my sister not by blood, but by heart. You supported me without waiting to be asked. Often, you understood my condition simply by seeing my face or hearing my voice. That kind of care and love is rare and I will always be grateful for it.

I am also grateful to my dear friends, whose patience, loyalty and kindness supported me when the work felt heavy. Their smessages, prayers, encouragement and understanding helped me through many difficult moments and I remain deeply thankful for their presence in my life.

Family and close friends may not know the technical details of the research, yet their support protects the peace, courage and strength that make sustained work possible. This thesis carries my name, but it also carries my family's sacrifices, prayers and love.

Acknowledgements

I gratefully acknowledge the financial support provided through the Warm Welcome Scholarship and British Council funding, which made it possible for me to undertake and complete this doctoral research at Brunel University London.

I am sincerely grateful to my principal supervisor, Professor Tatiana Kalganova, for her guidance, support and critical feedback throughout this research. Her supervision helped me strengthen the technical, methodological and scholarly foundations of the thesis. I also thank Dr Michael Rustell for his earlier supervision, advice and support in shaping the initial direction of the project.

I would like to thank the experts, specialists and participants who contributed to the empirical studies in this thesis. I am especially grateful to those who reviewed the Bamiyan reconstruction candidates and provided archaeological, technical and interdisciplinary judgement. I also thank the Afghan respondents who participated in the community-grounded authenticity study and shared reflections that could not have been reached through computational or expert evaluation alone.

I am grateful to colleagues, friends and staff at Brunel University London who supported my doctoral work in different ways. Their assistance, encouragement and advice have contributed to the completion of this thesis.

Declaration of Authorship

I, Kawsar Arzomand, declare that the work completed in this thesis is my original work conducted in accordance with the Code of Practice for Research Degrees. The work presented in this thesis has not been used in any other submission for an academic award.

A handwritten signature in black ink, appearing to read 'Kawsar Arzomand', with a horizontal line underneath.

Signature:

Date: 20 May 2026

Statement of Contribution to Co-authored Outputs

This thesis is submitted as a monograph thesis, not as a thesis by publication. Some chapters draw on published or submitted co-authored outputs produced during the doctoral research. In each case, the doctoral candidate led the research design, empirical work, analysis, writing and integration of the material into the thesis, unless otherwise stated.

The published and submitted outputs are listed in Section 1.8. Supervisory co-authors contributed through academic supervision, methodological advice, critical feedback and manuscript review. Chapter 6 is based on a submitted co-authored manuscript with Professor Tatiana Kalganova and Dr Manuela Ritondale. For that study, the candidate led the ethics preparation, questionnaire design, participant recruitment, data collection, quantitative and qualitative analysis, interpretation, writing and thesis integration. Professor Kalganova provided supervisory guidance and critical review. Dr Ritondale provided conceptual feedback, framing suggestions and critical review/editorial comments on the manuscript.

Ethics Approval Statement

Ethics approval for the human-participant components of this doctoral research was obtained from the Brunel University of London Research Ethics Committee under reference 50187-LR-Jan/2025-53624-2. The approval covered the expert evaluation components used in the HARF evaluation, the Bamiyan computational and specialist evaluation, the community-grounded authenticity questionnaire with Afghan respondents and the Palmyra bounded transferability questionnaire block.

All participant-facing elements were conducted through voluntary participation and informed consent. Respondents were provided with information about the study purpose, the use of their responses, anonymity, data handling and withdrawal conditions before participation. Human-participant data were processed in anonymised or pseudonymised form and stored in accordance with Brunel University research data-management requirements and UK GDPR.

Full procedural detail for the community-grounded authenticity study, including consent, anonymisation and quotation policy, is provided in Appendix D, Section S1. The expert evaluation instruments are documented in Appendix B and the relevant chapter methodology sections.

Contents

ABSTRACT	2
THANKS	3
ACKNOWLEDGEMENTS	5
DECLARATION OF AUTHORSHIP	6
STATEMENT OF CONTRIBUTION TO CO-AUTHORED OUTPUTS	7
ETHICS APPROVAL STATEMENT	8
CONTENTS	9
LIST OF FIGURES	19
LIST OF TABLES	23
LIST OF APPENDIX FIGURES AND TABLES	25
LIST OF ACRONYMS	26
CHAPTER 1	1
INTRODUCTION	1
1.1 ACCOUNTABLE GENAI HERITAGE RECONSTRUCTION	1
1.2 MOTIVATION	2
1.3 RESEARCH PROBLEM	3
1.4 AIMS, SCOPE AND BOUNDARIES	5
1.4.1 EVALUATION, NOT RECONSTRUCTION	5
1.4.2 TWO-DIMENSIONAL IMAGE OUTPUTS	6
1.4.3 FRAMEWORK SCOPE, PROMPTING AND EMPIRICAL BOUNDARIES	6
1.5 RESEARCH QUESTIONS	7
1.6 METHODOLOGICAL ARCHITECTURE	7
1.6.1 EVALUATIVE REGISTERS	8
1.7 ORIGINAL CONTRIBUTIONS	8

1.8	PUBLICATIONS AND OUTPUTS	10
1.9	STRUCTURE AND OUTLINE OF THE THESIS	11
CHAPTER 2		13
LITERATURE REVIEW		13
2.1	INTRODUCTION	13
2.2	HERITAGE STANDARDS AND EVIDENTIAL ACCOUNTABILITY	15
2.2.1	DISTINGUISHABILITY, AUTHENTICITY AND TRANSPARENCY	15
2.2.2	PARADATA AND UNCERTAINTY	17
2.2.3	IMPLICATIONS FOR GENAI EVALUATION	18
2.3	EVIDENCE-CONSTRAINED RECONSTRUCTION BEFORE GENAI	19
2.3.1	PHOTOGRAMMETRY, LASER SCANNING AND ARCHIVAL RECONSTRUCTION	19
2.3.2	HBIM AND GEOMETRIC VALIDATION	21
2.3.3	CASE-BOUND RECONSTRUCTION PRACTICE	22
2.4	GENERATIVE AI AND EVIDENTIAL-CONTROL GAP	23
2.4.1	GANS AND AI-ASSISTED RECONSTRUCTION	24
2.4.2	TEXT-TO-IMAGE AND MULTIMODAL RECONSTRUCTION	24
2.4.3	THREE-DIMENSIONAL GENERATION AS A BOUNDARY CASE	25
2.4.4	BENCHMARKS, DIAGNOSTICS AND POLICY FRAMEWORKS	26
2.4.5	TRAINING-DATA BIAS AND CULTURAL-FLATTENING	27
2.4.6	GAP 1 - EVIDENTIAL CONTROL	28
2.5	VALIDATION, METRICS AND EXPERT JUDGEMENT	30
2.5.1	EXISTING VALIDATION FRAMEWORKS IN DIGITAL HERITAGE	30
2.5.2	COMPUTATIONAL METRICS AND THEIR LIMITS	31

2.5.3	EXPERT REVIEW AS EVALUATION METHOD	32
2.5.4	GAP 2 - INTEGRATED EVALUATION	33
2.6	AUTHENTICITY, COMMUNITY AUTHORITY AND CULTURAL ACCOUNTABILITY	35
2.6.1	CULTURALLY SITUATED AUTHENTICITY	35
2.6.2	COMMUNITY AUTHORITY AND PARTICIPATORY HERITAGE	36
2.6.3	BAMIYAN AND THE STRUCTURE OF VOICE IN THE POST-2001 LITERATURE	37
2.6.4	AI ETHICS AND REPRESENTATIONAL BIAS	38
2.6.5	GAP 3 – COMMUNITY-GROUNDED AUTHENTICITY	39
2.7	TRANSFERABILITY AND BOUNDED GENERALISATION	40
2.7.1	TRANSFERABILITY IN DIGITAL HERITAGE AND COMPUTATIONAL ARCHAEOLOGY	41
2.7.2	STABLE AND VARIABLE ELEMENTS OF EVALUATION	42
2.7.3	THE TEMPLE OF BEL AS A TRANSFERABILITY CASE	43
2.7.4	GAP 4 - BOUNDED GENERALISATION	43
2.8	SYNTHESIS OF LITERATURE-DERIVED GAPS	44
2.8.1	THE FOUR GAPS AS A CUMULATIVE ARGUMENT	45
2.8.2	MAPPING THE LITERATURE	46
2.8.3	THESIS POSITION	47
CHAPTER 3		49
INITIAL EVIDENCE-CONSTRAINED GENAI RECONSTRUCTION WORKFLOW		49
3.1	INTRODUCTION	49
3.2	METHODOLOGY	51
3.2.1	DATA COLLECTION AND PREPARATION	53
3.2.2	AI IMAGE GENERATION PROCESS	53

3.2.3	IMAGE SELECTION AND ITERATIVE REFINEMENT	54
3.2.4	COMPARATIVE ANALYSIS AND PRELIMINARY SCREENING	54
3.2.5	SCHOLARLY DISCUSSION AND PRACTICAL IMPLICATIONS:	55
3.3	RESULTS AND DISCUSSION	56
3.3.1	PALMYRA (SYRIA)	60
3.3.2	MAYA SITES (MEXICO)	61
3.3.3	POMPEII (ITALY)	62
3.3.4	BUDDHAS OF BAMIIAN (AFGHANISTAN)	63
3.4	LIMITATIONS AND FUTURE DIRECTIONS	66
3.5	CONCLUSION	67
CHAPTER 4		68
HARF DEVELOPMENT AND OPERATIONALISATION		68
4.1	INTRODUCTION	68
4.2	METHODOLOGY	70
4.2.1	CASE STUDY AND EVIDENTIARY CORPUS	71
4.2.2	DYNAMIC PROMPT BLUEPRINT (DPB) ARCHITECTURE	74
4.2.3	PROMPT SUFFICIENCY ASSESSMENT	74
4.2.4	MULTIMODAL GENERATION WORKFLOW AND PLATFORM EVALUATION	75
4.2.4.1	AI MODEL EVALUATION AND COMPARATIVE ANALYSIS	76
4.3	GENERATION, PRE-SELECTION AND TRANSPARENCY	78
4.4	EXPERT VALIDATION, ETHICAL INTEGRATION AND REPRODUCIBILITY	78
4.4.1	DATA COLLECTION TOOLS AND PROCEDURES	81
4.4.2	ANALYTICAL WORKFLOW AND APPROACH	81

4.4.3	ETHICAL CONSIDERATIONS AND DATA MANAGEMENT	82
4.4.4	METHODOLOGICAL LIMITATIONS, RISK MITIGATION AND REPRODUCIBILITY	83
4.4.5	PARADATA AND METADATA CAPTURE	84
4.5	RESULTS	84
4.5.1	EXPERT-EVALUATION OF THE DPB AND PROMPT STRUCTURE	85
4.5.2	QUANTITATIVE PSI AND REFINEMENT OUTCOMES	87
4.5.3	CROSS-SITE EVALUATION	88
4.5.4	EVALUATION OF ETHICAL AND TRANSPARENCY MEASURES	89
4.5.5	NOVELTY, CONTRIBUTION AND SCHOLARLY SIGNIFICANCE	90
4.6	LIMITATIONS AND FUTURE WORK	91
4.7	CONCLUSION	92
CHAPTER 5		94
COMPUTATIONAL AND EXPERT EVALUATION AT BAMIIYAN		94
5.1	INTRODUCTION	94
5.2	METHODOLOGY	95
5.2.1	CORPUS GENERATION AND ITERATIVE REFINEMENT	95
5.2.2	IMPLEMENTATION PARAMETERS AND ITERATIVE CYCLES	96
5.2.3	COMPUTATIONAL STRUCTURAL EVALUATION	99
5.2.4	EXPERT EVALUATION DESIGN	101
5.3	RESULTS	102
5.3.1	BASELINE GENERATIVE BEHAVIOUR	102
5.3.2	CORPUS REFINEMENT AND PLATFORM BEHAVIOUR	104
5.3.3	CANDIDATE SELECTION AND STRUCTURAL EVALUATION	107

5.3.4	FORCED-CHOICE COMPARISON RESULTS	111
5.3.5	FULL-FIGURE RECONSTRUCTION SCORES (IMAGES 1–11; FIGURE 30)	111
5.3.5.1	COMPARISON WITH ALGORITHMIC STRUCTURAL SIMILARITY	112
5.3.5.2	MEAN APPLICABLE-CRITERIA SCORES ACROSS ALL CANDIDATES (K=18)	114
5.3.5.3	CRITERION-LEVEL MEANS ACROSS CANDIDATES	115
5.3.5.4	INTER-RATER AGREEMENT BY CRITERION	116
5.4	DISCUSSION	117
5.4.1	DIVERGENCE BETWEEN ALGORITHMIC SIMILARITY AND EXPERT EVALUATION	117
5.4.2	RECURRENT ANATOMICAL AND DRAPERY INSTABILITIES	118
5.4.3	ITERATIVE REFINEMENT LIMITS	119
5.4.4	EXPERT RECOMMENDATIONS AND METHODOLOGICAL IMPLICATIONS	120
5.5	LIMITATIONS	121
5.6	CONCLUSIONS	122
CHAPTER 6		124
COMMUNITY-GROUNDED AUTHENTICITY AT BAMİYAN		124
6.1	INTRODUCTION	124
6.2	BAMİYAN IN THE LITERATURE: POST-2001 DOCUMENTATION AND THE QUESTION OF AUTHENTICITY	126
6.3	METHODOLOGY	128
6.4	RESULTS	130
6.5	A DESIGN ORIENTATION FOR THE AI-MEDIATED WORKFLOW	134
6.6	DISCUSSION, LIMITATIONS AND FUTURE WORK	136
6.7	CONCLUSION	138
CHAPTER 7		139
BOUNDED CROSS-SITE GENERALISATION OF THE TEMPLE OF BEL, PALMYRA		139

7.1	INTRODUCTION	139
7.2	METHODOLOGY	141
7.2.1	ADAPTING THE DYNAMIC PROMPT BLUEPRINT TO PALMYRA	141
7.2.2	EVALUATION INSTRUMENT	142
7.2.3	METHODOLOGICAL JUSTIFICATION FOR FRAMEWORK-LEVEL EVALUATION	145
7.2.4	RESPONDENTS AND ANALYTICAL STRATEGY	145
7.3	RESULTS	146
7.3.1	ARCHITECTURAL FIDELITY	146
7.3.2	ITERATIVE REFINEMENT	148
7.3.3	COMPARISON WITH THE PHOTOGRAPHIC RECORD	149
7.3.4	PROMPT-KEYWORD SUFFICIENCY	149
7.4	DISCUSSION	150
7.5	LIMITATIONS	152
7.6	CONCLUSION	152
CHAPTER 8		154
DISCUSSION AND CONCLUSION		154
8.1	HARF AND THE RESEARCH QUESTIONS	154
8.2	OVERALL EVALUATION OF THE FRAMEWORK	157
8.2.1	PLATFORM VOLATILITY AND REPRODUCIBILITY	157
8.2.2	REFERENCE-MATERIAL DEPENDENCY	157
8.2.3	EXPERT-PANEL SCOPE	158
8.2.4	COMMUNITY-LAYER SCOPE	158
8.2.5	COMPUTATIONAL-PIPELINE SCOPE	158

8.2.6	TWO-DIMENSIONAL OUTPUT SCOPE	158
8.3	FUTURE WORK	159
8.4	CLOSING STATEMENT	160
APPENDIX A. DYNAMIC PROMPT BLUEPRINT (DPB): SITE-AGNOSTIC SCHEMA		161
A.1.	DPB SCHEMA SPECIFICATION	161
A.2.	IMPLEMENTATION NOTES	161
A.3.	ABSTRACT EXAMPLE (NON-SITE-SPECIFIC)	162
APPENDIX B. PHASE I EXPERT EVALUATION INSTRUMENT (BAMIYAN CASE STUDY)		163
B.1	SCOPE AND DESIGN	163
B.2	PARTICIPANT PROFILE	163
B.3	EVALUATION SCALE	163
B.4	STRUCTURAL FIDELITY AND SURFACE REALISM (IMAGES 1–7)	164
B.5	ICONOGRAPHIC AND STYLISTIC ACCURACY (IMAGES 8–17)	164
B.6	ENVIRONMENTAL PLAUSIBILITY (IMAGES 18–23)	164
B.7	ITERATIVE RECONSTRUCTION ASSESSMENT	165
B.8	PROMPT SUFFICIENCY	165
B.9	USE OF RESPONSES	165
APPENDIX C. BAMIYAN CANDIDATE CORPUS, VISUAL STIMULI AND PARADATA		166
C.9	DPB-DERIVED BAMIYAN PROMPT RECORD	169
C.9.1	SCOPE AND STATUS	169
C.9.2	PROMPT RECORD	169
C.9.2.1	INITIAL PAIRED WESTERN/EASTERN BUDDHA RECONSTRUCTION	169
C.9.2.2	INITIAL WESTERN BUDDHA VERSION	170
C.9.2.3	INITIAL EASTERN BUDDHA VERSION	170
C.9.2.4	STANDING BUDDHA WITH EARLY MATERIAL AND INSCRIPTION DESCRIPTORS	171
C.9.2.5	HYPER-REALISTIC WESTERN BUDDHA WITH EARLY COLOUR DESCRIPTORS	171
C.9.2.6	BAMIYAN VALLEY AND CAVE CONTEXT	171
C.9.2.7	WESTERN BUDDHA WITH DIMENSIONAL AND MATERIAL ANCHORS	171
C.9.2.8	EASTERN BUDDHA WITH DIMENSIONAL ANCHORS	172
C.9.2.9	BAMIYAN VALLEY SETTING WITH BOTH STATUES	172
C.9.2.10	UNIFIED WESTERN/EASTERN BUDDHA	172
C.9.2.11	REVISED UNIFIED WITH DISTANCE RELATION	173
C.9.2.12	WESTERN BUDDHA FOCUSED VERSION	174
C.9.2.13	EASTERN BUDDHA FOCUSED VERSION	174
C.9.2.14	FULL STANDING WESTERN BUDDHA CORRECTION	175
C.9.2.15	HEAD MORPHOLOGY	175

C.9.2.16 RIGHT-HAND	176
C.9.2.17 LEFT-HAND	176
C.9.2.18 TORSO AND ROBE	176
C.9.2.19 FEET AND TOES	176
C.9.2.20 FRONT VIEW	176
C.9.2.21 SIDE VIEW	177
C.9.2.22 BACK VIEW	177
C.9.2.23 45-DEGREE VIEW	177
C.9.2.24 NICHE AND STATUE	177
C.9.2.25 NICHES AND SURROUNDING CAVES	177
C.9.2.26 PANORAMIC BAMİYAN VALLEY	177
C.9.2.27 RIGHT HAND, STANDING POSTURE	178
C.9.2.28 LEFT HAND, STANDING POSTURE	178
C.9.2.29 RIGHT HAND FROM MULTIPLE ANGLES	178
C.9.2.30 BACK AND SIDE VIEWS OF BOTH HANDS	178
C.9.2.31 BACK VIEW FROM BUDDHA PERSPECTIVE	179
C.9.2.32 BACK SIDE OF RIGHT HAND	179
C.9.2.33 SANDSTONE CLIFFS AND MONASTIC CAVES	179
C.9.2.34 EVIDENCE-CONSTRAINED WESTERN BUDDHA	180
C.9.2.35 STRUCTURED EXPERT-INFORMED EXAMPLE	180
C.9.2.36 CULTURAL VISUALISATION	181
C.9.2.37 FULL-BODY PHOTOREALISTIC ARCHIVAL-STYLE	181
C.9.2.38 CONDENSED DETAIL-RETENTION VERSION	182
C.9.2.39 MASTER FULL-FIDELITY VERSION	182
C.9.2.40 REFERENCE-FUSION INSTRUCTION	183
C.9.2.41 LOCALISED CORRECTION INSTRUCTION	185
APPENDIX D. SUPPLEMENTARY MATERIALS FOR CHAPTER 6 (COMMUNITY-GROUNDED AUTHENTICITY AT BAMİYAN)	186
S1. ETHICS APPROVAL, CONSENT PROCEDURE AND ANONYMISATION	186
S2. QUESTIONNAIRE INSTRUMENT	187
BLOCK A — DEMOGRAPHIC CONTEXT	187
BLOCK B — CONNECTION AND MEMORY	187
BLOCK C — INTERPRETATIONS OF DESTRUCTION	188
BLOCK D — RECONSTRUCTION PREFERENCES AND ETHICS	188
BLOCK E — INTANGIBLE AND SENSORY DIMENSIONS	188
BLOCK F — OPEN REFLECTIONS	189
S3. SAMPLE COMPOSITION (FULL DEMOGRAPHIC BREAKDOWN)	189
S4. FULL SINGLE-SELECT ITEMS WITH WILSON 95% CONFIDENCE INTERVALS	191
S5. MULTI-SELECT ITEM FREQUENCIES WITH 95% CIS	191

S6. FULL SUB-GROUP CONTRASTS	194
S7. SAMPLE VS POPULATION REPRESENTATIVENESS	195
S8. HAZARA SUB-SAMPLE BREAKDOWN (N = 13)	195
S9. THEMATIC CODING SCHEME	196
S10. FILES AND REPRODUCTION RECIPE	197
REFERENCE	198

List of Figures

Figure 1. Author-generated diagram of the thesis architecture showing the relationship between the research problem, literature-derived gaps, research questions, contribution chapters and bounded generalisation test. 10

Figure 2. Five-phase workflow used in this chapter; diagram produced for the thesis. _____ 52

Figure 3. Flowchart illustrating the five-phase exploratory workflow; diagram produced for the thesis with AI-generated illustrative thumbnails. _____ 52

Figure 4. Reference photographs of the surviving ruins of Palmyra, including the Temple of Bel, the Baalshamin Temple and the ancient theatre. Source: Ko Hon Chiu Vincent/UNESCO. _____ 60

Figure 5. AI-generated exploratory reconstruction candidates for sections of Palmyra, produced using Midjourney and DALL-E. _____ 60

Figure 6. Reference photographs of Maya archaeological sites in Mexico. Source: Community Tours Sian Ka'an and Ko Hon Chiu Vincent. _____ 61

Figure 7. AI-generated exploratory reconstruction candidate of the architectural layout of Chichen Itza, produced using Midjourney. _____ 61

Figure 8. AI-generated exploratory reconstruction candidates showing architectural features of Chichen Itza, produced using Midjourney. _____ 62

Figure 9. Reference photographs of the ruins of Pompeii. Source: J. Frias Velatti and Limes Media/Tim Schnarr. _____ 62

Figure 10. AI-generated exploratory reconstruction candidates showing Pompeii architecture and interiors, produced using Midjourney and DALL-E. _____ 63

Figure 11. Western Buddha of Bamiyan: (a) reference photograph of the Western Buddha in its niche. Source: James Gordon/Wikimedia Commons; (b–c) AI-generated reconstruction candidates produced using Midjourney/DALL-E. _____ 63

Figure 12. Comparison of image-quality metrics for ten Western Buddha candidate images, labelled Images 0–9. _____ 64

Figure 13. Per-image metric panel for Image 3, showing the reference image, AI-generated candidate image, SSIM map, difference image, pixel-difference histogram and metric values. _____ 65

Figure 14. Per-image metric panel for Image 9, showing the reference image, AI-generated candidate image, SSIM map, difference image, pixel-difference histogram and metric values. _____ 65

Figure 15. Author-generated overview diagram of the Heritage-Aligned Reconstruction Framework (HARF) and its six interconnected stages for the reconstruction of the Western Buddha of Bamiyan. _____ 71

Figure 16. Pre-2001 view of the Western Buddha of Bamiyan showing the niche architecture, drapery contours and surrounding monastic caves. Source: Khan Academy, UNESCO. _____ 72

Figure 17. Author-generated demographic profile of the Phase I expert panel ($n = 33$), showing country distribution, educational attainment, years of professional experience and familiarity with the Buddhas of Bamiyan. _____ 79

Figure 18. AI-generated reconstruction candidates evaluated in Phase I. (a) structural reconstruction candidate (Image 1 - left), (b) iconographic detail candidate (Image 13 - middle), (c) environmental context candidate (Image 21 - right) _____ 80

Figure 19. Author-generated chart showing expert ratings of baseline prompt sufficiency ($n = 33$). _____ 85

Figure 20. Author-generated chart showing domains prioritised by experts for prompt enrichment ($n = 33$). The distribution visualises which categories were most frequently prioritised for refinement in AI-guided heritage visualisation. _____ 86

Figure 21. Author-generated workflow diagram showing the empirical evaluation pipeline for the Bamiyan reconstruction case study. The workflow summarises the staged evaluation process applied to the Western Buddha reconstruction corpus. An initial generation corpus ($k = 245$ images; Sora = 187, Gemini = 58) underwent interdisciplinary expert screening ($k = 23$ images; $n = 33$ experts) followed by iterative prompt refinement and regeneration. The refined candidate set ($k = 18$ images) was subsequently evaluated in Phase II specialist assessment ($n = 5$) using a structured applicable-criteria rubric, producing the final analytical dataset comprising 11 full-figure reconstructions, 3 drapery studies and 4 head studies. Note: k denotes the number of images; n denotes the number of experts. _____ 97

Figure 22. Author-defined ROI annotation used for computational structural evaluation. The statue ROI (red) and niche ROI (green) enable independent measurement of morphological similarity and architectural alignment relative to the archival reference. _____ 99

Figure 23. Representative geometric registration to the archival reference. The figure combines the archival photograph (left), an AI-generated reconstruction candidate prior to registration (centre), and an author-generated registration overlay (right). The overlay illustrates the two-dimensional alignment procedure applied before structural similarity measurement. _____ 100

Figure 24. Archival photograph of the Giant Buddha at Bamiyan © UNESCO (left), with AI-generated baseline reconstruction (centre) and AI-generated refined reconstruction (right). _____ 103

Figure 25. Structurally inconsistent AI-generated reconstructions presenting architectural and contextual deviations. All images were generated without DPB slot constraints. (a) Incorporation of Petra-like carved façade geometry within the niche (far left); (b) Addition of a temple-style podium and secondary monumental figures (second from left); (c) Introduction of dense foreground vegetation not evidenced in archival documentation (third from left); (d) Symmetrical niche redesign and halo-like sculptural additions (far right). _____ 103

Figure 26. Phase I baseline AI-generated reconstruction candidates (Images 1–7) generated using OpenAI DALL-E. Each image was produced from a single-shot prompt without DPB slot constraints and iterative refinement. No post-generation editing was applied. These images formed the first part of the Phase I questionnaire screening set and were used to assess baseline generative behaviour before DPB-guided refinement. The outputs demonstrate structural and iconographic inconsistencies characteristic of unconstrained generative prompting. _____ 104

Figure 27. AI-generated iterative DPB-guided refinement of head morphology, hand articulation and drapery across three successive generations. Iteration 1 was generated using an initial DPB-structured prompt. Iterations 2 and 3 were produced within the same OpenAI ChatGPT image-generation session through progressive prompt modification, introducing increasingly specific morphological and iconographic constraints for facial features, hand gestures and drapery articulation. The underlying prompt architecture remained constant, while constraint granularity was incrementally increased. All outputs are direct model generations without manual pixel-level editing. _____ 104

Figure 28. AI-assisted reference-conditioned refinement of Buddha head morphology using Sora. A reference image of a comparable Gandhāran Buddha head was used to guide generation through successive prompt refinements describing cranial height, bun form and hair wave structure. The sequence shows the progressive adjustments leading to the final generated head. _____ 105

Figure 29. AI-assisted targeted regional refinement of drapery articulation using the Gemini API. Targeted regional edits were applied through the Gemini API to correct robe folds and garment layering. Successive iterations progressively stabilised the wet-fold structure, garment layering and under-robe visibility. _____ 105

Figure 30. Final set of reconstruction candidates retained for expert evaluation. Images 1–11 were generated through AI workflows, while Image 12 (right) represents a manually corrected reconstruction produced in Adobe Photoshop. _____ 106

Figure 31. Author-produced masking illustration from the AI-assisted refinement workflow. Highlighted regions indicate areas selected for targeted prompt refinement, including hair morphology, hand articulation, niche openings behind the feet and foreground architectural elements. These regions guided the iterative correction process during AI image generation. These masks were used during generation refinement and were not part of the computational evaluation pipeline. _____ 107

Figure 32. Author-generated charts. Candidate selection and structural evaluation across analysis phases. (a) Phase I - Interdisciplinary Expert Screening (IES) selections across evaluation domains. Bars show the frequency with which reconstruction candidates were selected by respondents ($n = 33$) across the three screening domains: structural plausibility ($k=7$; Images 1–7), iconographic authenticity ($k=10$; Images 8–17) and environmental context ($k=6$; Images 18–23). Multiple selections were permitted within each domain. (b) Structural similarity metrics for Phase II full-figure reconstruction candidates. The figure reports statue-region SSIM, niche-region SSIM, proportional similarity and the resulting composite structural similarity score used to rank the eleven reconstruction candidates relative to the archival reference image. (c) Algorithmic structural ranking of Phase II full-figure reconstructions. Higher values indicate stronger geometric correspondence with the archival reference photograph. _____ 109

Figure 33. AI-generated iconographic reconstruction candidates used during Phase I — Interdisciplinary Expert Screening (IES), including head morphology, hand gesture and drapery studies (Images 8–17). _____ 109

Figure 34. AI-generated environmental context reconstruction candidates used during Phase I — Interdisciplinary Expert Screening (IES), illustrating alternative interpretations of the Bamiyan cliff architecture and surrounding cave landscape (Images 18–23). _____ 109

Figure 35. Author-generated chart showing the relationship between expert SSE mean scores and AI composite similarity scores for the eleven full-figure reconstruction candidates. Panel (a) reports the relationship between expert and computational scores; panel (b) compares expert SSE mean scores and AI composite similarity scores across the eleven candidates. _____ 114

Figure 36. Author-generated chart based on Phase II — Site-Specific Specialist Evaluation (SSE) mean applicable-criteria scores for the 18 candidates. Mean specialist scores are reported on a 0–1 scale for head studies, drapery studies and full-figure reconstructions. _____ 115

Figure 38. Author-generated heat map based on Phase II SSE criterion-level scores across candidates. Darker cells represent higher mean specialist scores (0–1). Grey cells indicate criteria not applicable to a given candidate subset. _____ 116

Figure 39. Author-generated inter-rater agreement chart across SSE evaluation criteria. Criterion-wise agreement was calculated from the five Phase II specialist ratings using quadratic-weighted Cohen's κ and Gwet's AC2, with 95% confidence intervals. N/A entries were excluded from each criterion-specific calculation. The dashed line at 0.60 is used as a visual reference threshold for interpretive comparison. _____ 117

Figure 40. Author-generated chart based on expert identification of reconstruction instabilities across iterative refinement stages based on the reconstruction outputs illustrated in Figure 27. The chart compares the

frequency with which experts identified three diagnostic features, head details, hand gestures and robe drapery, as problematic across Iterations 1 (Preliminary), 2 (Intermediate) and 3 (Final). The vertical axis represents the number of expert observations. _____ 120

Figure 41. Author-generated chart showing the modes through which respondents reported a connection to the Bamiyan Buddhas (multi-select; N = 62). Respondents could select more than one mode, so totals exceed N by design. Counts and percentages are shown alongside each bar. _____ 131

Figure 42. Author-generated chart showing reconstruction support and aspects respondents prioritised for reconstruction (N = 62). Panel A shows responses to the single-select item on whether the Buddhas should be reconstructed. Panel B shows multi-select responses on which aspects should be reconstructed; respondents could select more than one aspect, so totals exceed N by design. Counts and percentages are shown alongside each bar. _____ 132

Figure 43. Architectural and structural-accuracy stimuli: the ten AI-generated Temple of Bel reconstruction candidates evaluated by the panel. AI-generated evaluation stimuli, not presented as final reconstructions; platform metadata, model identifiers, prompt version, seed values and refinement-cycle records are recorded in the project paradata in line with HARF and the London Charter [16], [30]. _____ 143

Figure 44. Iterative-refinement stimuli for Temple of Bel relief iconography. Top left: reference photograph of a Palmyra relief (Francesco Bandarin/UNESCO, 2007); remaining panels: AI-generated iterations used as evaluation stimuli. _____ 144

Figure 45. Comparison of documented Temple of Bel ruins and AI-generated reconstruction stimuli. Image 11: Pre-2015 reference photograph of the Temple of Bel ruins (Getty Images via BBC News). Image 12: Pre-2015 reference photograph of the Temple of Bel ruins (Ian Plumb, CC BY 2.0). Images 13–14: AI-generated evaluation stimuli. _____ 144

Figure 46. Iteration judged most historically accurate for Temple of Bel relief iconography. The modal response was “None of the iterations are historically accurate” (56.2%; 95% CI 39.3–71.8%), indicating that iterative refinement did not resolve relief-iconographic concerns. Author-generated chart based on Palmyra questionnaire responses. _____ 148

Figure 47. Categories identified as most divergent from the documented Temple of Bel ruins. Bars denote the proportion of respondents flagging each category; error bars denote Wilson 95% confidence intervals. Surface, relief and roofline detail concentrate at the top of the distribution. Author-generated chart based on Palmyra questionnaire responses. _____ 149

List of Tables

<i>Table 1. Evaluation functions in selected GenAI and AI-mediated heritage reconstruction literature</i>	28
<i>Table 2. Validation approaches relevant to GenAI heritage reconstruction</i>	33
<i>Table 3. Literature-derived gaps and thesis response</i>	46
<i>Table 4. Overview of Historical Heritage Sites: Characteristics and Conditions</i>	56
<i>Table 5. Technological Approaches in Heritage Site Reconstruction</i>	58
<i>Table 6. Platform-selection matrix for the retained HARF generation workflow. The table summarises platform roles during DPB prompt development and the basis for retaining or excluding each system from the main Bamiyan workflow. It reports workflow-selection judgements, not benchmark scores.</i>	76
<i>Table 7. Summary of model versions and generation parameters used in the HARF workflow. Settings apply to all experiments unless otherwise indicated.</i>	77
<i>Table 8. Summary of methodological risks, their relevance to reproducibility and cultural fidelity, the mitigation strategies applied and the residual severity after mitigation.</i>	83
<i>Table 9. Thematic clusters of expert-suggested prompt refinements, with representative quotations illustrating specific areas for improvement. n = number of experts out of 33 who identified that theme during the qualitative analysis.</i>	86
<i>Table 10. Cross-Site Substitution of DPB Structural Tiers</i>	88
<i>Table 11. Formation and use of the Bamiyan evaluation datasets</i>	97
<i>Table 12. Expert rationales for structural-domain preferences during IES (Images 1–7; Figure 26). Key reasons and common feedback cited by respondents (n = 33).</i>	110
<i>Table 13. Mean SSE scores for full-figure reconstruction candidates (Images 1–11; Figure 30). Values represent the arithmetic mean of ratings assigned independently by five specialists using the SSE rubric (0 = absent/incorrect, 0.5 = partially correct/ambiguous, 1 = correct/clearly satisfied). Candidate-level means were calculated as the arithmetic mean across the ten applicable criteria. The three highest-ranking candidates (Images 2, 1 and 11) are highlighted for visual clarity. Note: N/A entries were excluded from criterion-specific means.</i>	112
<i>Table 14. Spearman correlations between computational similarity metrics and expert SSE scores.</i>	112
<i>Table 15. Comparison of expert and algorithmic rankings for full reconstruction candidates (Images 1–11; Figure 30). Human rank is based on the candidate-level mean SSE score across ten criteria. AI rank is based on the composite structural similarity score reported in Figure 32. Rank difference is calculated as the absolute difference between human and algorithmic positions. Agreement categories are defined as: Match (difference = 0), Close (difference = 1), Moderate divergence (difference = 2–3) and Large divergence (difference ≥ 4).</i>	113
<i>Table 16. Post-2001 reconstruction approaches to the Bamiyan Buddhas, the institutional teams involved and what each approach addresses and does not address.</i>	127

<i>Table 17. Headline proportions with Wilson 95% confidence intervals (N = 62).</i>	130
<i>Table 18. Sub-group contrasts on core items</i>	134
<i>Table 19. Cross-site instantiation of the Dynamic Prompt Blueprint for the Temple of Bel, with Bamiyan referents shown for comparability. The schema is preserved; the slot contents are replaced with site-specific descriptors drawn from documentary sources.</i>	141
<i>Table 20. Likert ratings of the AI reconstructions of the Temple of Bel (n = 32). Counts denote the number of respondents in each scale category. Wilson 95% confidence intervals are reported for the proportion rated ≥ 4.</i>	147
<i>Table 21. Three-layer model of HARF transferability and typology-specific instability across Bamiyan and Palmyra.</i>	151

List of Appendix Figures and Tables

<i>C. 1. Hybrid refinement workflow combining Sora base generation with Gemini (Google AI Studio) regional masked editing to correct localised anatomical and drapery distortions while preserving global statue proportions.</i>	166
<i>C. 2. Gemini-only generations (prompt-only; excluded). Gemini generations produced from text prompts without Sora inputs. Outputs often exhibited high surface realism but introduced non-evidenced architectural insertions and/or proportion drift and were therefore excluded from the candidate pool.</i>	166
<i>C. 3. Google AI Studio API-based masked editing workflow. It shows general API testing on non-heritage images to verify regional editing functionality.</i>	166
<i>C. 4. Google AI Studio API-based masked editing It illustrates heritage-specific application to Bamiyan reconstruction outputs.</i>	166
<i>C. 5. Initial Sora-generated outputs produced from a single-shot prompt. Observed failures include mirrored hand gestures, incorrect mudra interpretation, exaggerated cranial proportions and drapery folds inconsistent with Gandhāran “wet-fold” conventions.</i>	167
<i>C. 6. Sora outputs under iterative text-only prompt revision. Sora generations produced by repeatedly revising the textual prompt to correct specific failures (posture, facial morphology, drapery structure). Despite successive revisions, several outputs retained structural or stylistic instability (e.g., residual hand errors, inconsistent drapery grammar, or niche-context distortions)</i>	167
<i>C. 7. Expert observations on strengths and inaccuracies identified during Phase I — Interdisciplinary Expert Screening (IES), covering both statue reconstruction outputs (Images 1–7; Figure 26) and environmental context reconstructions (Images 18–23, Figure 34).</i>	168
<i>C. 8. Provenance of the 18 shortlisted candidates, showing subset (full figure, drapery, head), platform of origin (Sora, Gemini/Google, hybrid) and targeted merges/corrective edits.</i>	169

List of Acronyms

ACRONYM	FULL FORM
2D	Two-dimensional
2D-ERT	Two-dimensional Electrical Resistivity Tomography
3D	Three-dimensional
3D GIS	Three-dimensional Geographic Information System
3DMM	Three-dimensional Morphable Model
AC2	Gwet's AC2 agreement coefficient
AD	Anno Domini
AI	Artificial Intelligence
AIGC	AI-Generated Content
API	Application Programming Interface
AR	Augmented Reality
BC	Before Christ
BIM	Building Information Modelling
BRIEF	Binary Robust Independent Elementary Features
cGANs	Conditional Generative Adversarial Networks
CI	Confidence Interval
CIs	Confidence Intervals
CLIP	Contrastive Language–Image Pre-training
CSV	Comma-Separated Values
DAFA	Délégation Archéologique Française en Afghanistan
dB	Decibel
DK	Do not know / Don't know
DPB	Dynamic Prompt Blueprint
EMF	Electromagnetic Field
ERT	Electrical Resistivity Tomography
FAIR	Findable, Accessible, Interoperable and Reusable
FAST	Features from Accelerated Segment Test
GAN	Generative Adversarial Network
GANs	Generative Adversarial Networks
GDPR	General Data Protection Regulation
GenAI	Generative Artificial Intelligence
GIS	Geographic Information System
HARF	Heritage-Aligned Reconstruction Framework

HBIM	Historic Building Information Modelling
HD	High Definition
ICOMOS	International Council on Monuments and Sites
IES	Interdisciplinary Expert Screening
IFVA	International Forum of Virtual Archaeology
IQR	Interquartile Range
ISPRS	International Society for Photogrammetry and Remote Sensing
JSON-LD	JavaScript Object Notation for Linked Data
LMM	Large Multimodal Model
LPIPS	Learned Perceptual Image Patch Similarity
LRM	Large Reconstruction Model
MAE	Mean Absolute Error
MSE	Mean Squared Error
N/A	Not applicable
NeRF	Neural Radiance Field
ORB	Oriented FAST and Rotated BRIEF
PSI	Prompt Sufficiency Index
PSNR	Peak Signal-to-Noise Ratio
RANSAC	Random Sample Consensus
ROI	Region of Interest
ROIs	Regions of Interest
RQ	Research Question
SD	Standard Deviation
SPACH	Society for the Preservation of Afghanistan's Cultural Heritage
SSE	Site-Specific Specialist Evaluation
SSIM	Structural Similarity Index Measure
TLS	Terrestrial Laser Scanning
TOF	Time-of-Flight
UAV	Unmanned Aerial Vehicle
UNESCO	United Nations Educational, Scientific and Cultural Organization
VR	Virtual Reality
WebGIS	Web-based Geographic Information System
XLSX	Microsoft Excel Open XML Spreadsheet

Chapter 1

Introduction

Generative AI has made images of destroyed monuments easier to produce than to justify as heritage reconstruction. The difficulty is not only whether an output looks plausible, but whether it can be traced to evidence, examined by specialists and assessed against the cultural meanings associated with the damaged or destroyed site. This thesis addresses that difficulty through the development and validation of the Heritage-Aligned Reconstruction Framework, using Bamiyan as the main development and validation case and Palmyra as a bounded cross-site generalisation test.

1.1 Accountable GenAI Heritage Reconstruction

Image-based reconstruction has always depended on more than visual completion. Where a monument survives only through fragments, photographs, drawings, surveys or conservation records, reconstruction depends on the quality, distribution and interpretive limits of the surviving record. It is therefore an argument about what the evidence permits. The deliberate destruction of the sixth- to seventh-century CE Buddhas of Bamiyan in March 2001 and the explosive demolition of the Temple of Bel at Palmyra in August 2015 make this issue especially acute. In both cases, the destroyed elements cannot be directly re-recorded. Reconstruction depends instead on uneven records assembled from pre-destruction photography, measured drawings, photogrammetric campaigns, archaeological monographs and conservation reports (Flood, 2002; Grün et al., 2004; Petzet, 2009; Schmidt-Colinet, 2019; Toubekis et al., 2017; Wahbeh et al., 2016). Under these conditions, reconstruction is not the recovery of a lost object. It is the production of a visual claim about an object that no longer exists in its destroyed form.

Digital heritage has developed established ways of controlling such claims. Photogrammetry, laser scanning, archival reconstruction and Historic Building Information Modelling treat reconstruction as an evidence-bound activity in which measurement, archival material, comparative scholarship and interpretive decisions are documented. The Venice Charter, the Nara Document on Authenticity, the London Charter for the Computer-Based Visualisation of Cultural Heritage and the Seville Principles

provide the standards against which reconstruction should be assessed: documented, inferred and conjectural elements should remain distinguishable; authenticity is culturally situated; interpretive decisions should be recorded through paradata; and uncertainty should be communicated rather than concealed (ICOMOS, 1964, 1994; López-Menchero Bendicho & Grande, 2017). These standards do not prescribe a specific technology. They require a discipline of accountability.

Generative AI changes the practical conditions under which that discipline operates. Diffusion-based text-to-image systems, multimodal models and large-scale generative pipelines can produce visually coherent reconstruction candidates from textual, photographic and multimodal inputs ((Arzomand et al., 2024; Delgado-Martos et al., 2026; Shih, 2025; Wei et al., 2026)). This is significant for heritage research because reconstruction from textual and archival evidence from chronicles, inscriptions, period descriptions, conservation reports and pre-destruction photographs, has historically required long, expert-led modelling work and skilled illustration. Generative systems make it practically feasible to translate such evidence into visual reconstruction candidates at a speed and scale that were not previously available, which is precisely why a framework for evaluating those candidates as heritage claims has become necessary. In conventional digital heritage workflows, the relation between source evidence, modelling decision and final output can usually be traced through measured construction and recorded interpretation. In prompt-based generation, a single image may combine documentary evidence, prompt wording, stochastic sampling and model-internal priors learned from large generic training corpora. The finished image does not disclose which source, or process has shaped which part of the output.

The consequence is that visual coherence can conceal evidential weakness. A generated reconstruction may present documented features, plausible inference and unsupported model addition with the same visual confidence. Recent studies have begun to identify this divergence between visual plausibility and evidential adequacy, while broader surveys of generative digital heritage note that evaluation remains largely perceptual or visual rather than evidentiary (Foka & Griffin, 2024a; Wei et al., 2026; Xu et al., 2025). The standards of digital heritage remain relevant, but they need to be operationalised for outputs produced through prompt-based generation.

1.2 Motivation

The motivation for this thesis follows directly from that gap. Generative systems can now produce heritage reconstruction candidates quickly, at scale and with high visual finish, but heritage evaluation operates on different terms. A reconstruction candidate must be assessed against the evidence from

which it claims authority, the interpretive decisions through which it was produced and the forms of expert and community judgement that make it acceptable or unacceptable as heritage representation. Existing generative heritage studies have shown that plausible images can be produced (Arzomand et al., 2024; Delgado-Martos et al., 2026; Shih, 2025; Wei et al., 2026). That finding is not sufficient for destroyed and contested heritage. Without a record of source selection, prompt construction, model choice, refinement and output filtering, the image cannot be audited as a reconstruction claim. Without expert review, archaeological, stylistic and proportional errors may remain hidden beneath visual polish. Without community-grounded assessment, especially where destruction is tied to memory, absence and contested representation, technical plausibility may be mistaken for cultural acceptability.

Bamiyan and Palmyra are used in this thesis because they test different parts of that evaluative demand. Bamiyan is the development and deep-validation case. It allows the framework to be built from documentary and archaeological evidence, tested through computational and specialist assessment and extended to community-grounded authenticity through Afghan respondent data. Palmyra has a different function. It is not a second full validation case and is not treated as one. It tests whether the evaluative schema developed through Bamiyan can be re-instantiated for a different heritage typology: monumental architecture rather than figural sculpture.

The thesis therefore does not ask whether generative AI can make convincing images of Bamiyan or Palmyra. It asks how such images should be evaluated before they are treated as heritage reconstruction. HARF is developed in response to that question: not as a reconstruction engine, but as an evaluation framework for assessing whether GenAI reconstruction candidates are evidentially constrained, expert-assessed, culturally accountable and transferable within stated limits.

1.3 Research Problem

The research problem is the absence of a reproducible, evidence-based, expert-informed and culturally accountable framework for evaluating generative AI reconstruction of destroyed and contested cultural heritage. The problem is methodological. It concerns four linked issues: evidential control, evaluation procedure, the relation between computational, expert and community-grounded assessment and transferability across heritage typologies.

The first issue is evidential control. Existing GenAI heritage studies show that plausible reconstruction candidates can be generated, but the evidential pathway from source material to output is rarely documented as a structured and auditable procedure. The multi-site text-to-image work shows what unstructured prompting can produce [10]. Later single-site studies [12] and pipeline-based

architectural reconstruction work [11] extend the technical reach without formalising the documentary record behind each output. Recent reviews of generative practice in digital heritage similarly note that evaluation remains predominantly visual rather than evidentiary (Arzomand et al., 2024; Delgado-Martos et al., 2026; Shih, 2025; Wei et al., 2026; Xu et al., 2025). Prompt construction, model selection, seed choice, refinement cycles and output filtering are not neutral technical settings; they are interpretive decisions that shape the historical claim made by the generated image. Where those decisions are not recorded, the generated output cannot be examined as a reconstruction claim.

The second issue is the absence of an AI-specific evaluation procedure. Digital heritage already contains structured validation traditions, but those traditions were developed for other techniques and remain divided across separate methodological settings. The London Charter and the Seville Principles set the evidentiary expectations of distinguishability, paradata and uncertainty communication (Denard, 2009; López-Menchero Bendicho & Grande, 2017), while photogrammetric and HBIM workflows provide measurable validation for geometric and semantic correspondence (Nikolakopoulou & Koutsabasis, 2020; Pietroni, 2021; Pietroni & Ferdani, 2021). AI-specific work has begun to address adjacent problems: the Time Travel benchmark scores large multimodal models on curated historical content (Ghaboura et al., 2025) and bias-diagnostic work identifies how representational distortions can propagate into later training corpora (Foka et al., 2025; A. Tiribelli et al., 2024). These contributions identify components of evaluation but do not yet integrate them into a single reproducible procedure linking evidence anchoring, prompt sufficiency, paradata, computational pre-evaluation and structured expert review.

The third issue is the separation of computational, expert and community-grounded assessment. Image-similarity metrics can identify departures from documented reference imagery but cannot adjudicate stylistic, iconographic or culturally specific accuracy. Structured expert assessment can address disciplinary plausibility but does not on its own settle questions of cultural legitimacy, especially where destroyed heritage is bound to community memory, ongoing custodianship and acceptable representation. The literature reviewed in Chapter 2 treats these registers largely as separate methodological settings: bias-diagnostic work shows the field-level risks of AI-mediated representational distortion (Foka & Griffin, 2024a, 2024b), decolonial and participatory scholarship insists on community authority in heritage representation (Risam, 2018; Stobiecka, 2020) and trustworthy-AI work in heritage articulates the ethical demands without supplying a workflow (S. Tiribelli et al., 2024). What is missing is an operational procedure for bringing these registers into the same evaluation workflow.

The fourth issue is transferability across heritage typologies. Most GenAI heritage reconstruction studies remain case-specific. The multi-site text-to-image work (Arzomand et al., 2024) and the

architectural pipeline study (Delgado-Martos et al., 2026) demonstrate technical reach across contexts, but neither test whether an evaluation procedure remains valid when the heritage typology changes. Comparative work in computational archaeology emphasises that any method developed within one site condition needs explicit transferability evidence before it can be reused (Bevan & Lake, 2016) and adjacent reviews note the same problem for AI-mediated cultural-heritage applications (Cui & Wu, 2025; X. Yang et al., 2020). A framework developed within one heritage typology cannot be assumed to remain valid when applied to a different typology with a different evidential record, different expert criteria and different failure domains; transferability must be tested empirically across cases.

The four issues cannot be resolved by improving generative capability alone. They require an evaluation procedure that organises evidence, computational assessment, expert judgement, community-grounded authenticity and bounded cross-site testing within a single accountable workflow.

1.4 Aims, Scope and Boundaries

The aim of this thesis is to develop and validate an evidence-based GenAI evaluation framework for cultural heritage reconstruction and to test its bounded generalisation across heritage typologies. The Heritage-Aligned Reconstruction Framework integrates evidence-anchored prompting, prompt sufficiency assessment, paradata recording, computational pre-evaluation and structured expert review within a reproducible procedure. It is developed and deeply validated through the Western Buddha of Bamiyan and tested for bounded transferability through the Temple of Bel at Palmyra.

1.4.1 Evaluation, Not Reconstruction

The thesis is not primarily concerned with producing visually convincing reconstructions of the Bamiyan Buddhas or the Temple of Bel. Generative outputs are used throughout the contribution chapters, but they are treated as evaluation stimuli rather than as final reconstructions. No claim is made that any generated image of the Western Buddha of Bamiyan or the Temple of Bel is a historically accurate reconstruction of the destroyed monument. The reconstructive question is therefore subordinated to the evaluative question.

1.4.2 Two-Dimensional Image Outputs

Evaluation is restricted to two-dimensional, image-based generative AI reconstruction. HARF is not extended in this thesis to text-to-3D, volumetric reconstruction, Neural Radiance Fields, Gaussian splatting or immersive heritage environments. Those methods are relevant future extensions, but they raise additional evidentiary, spatial and experiential validation questions that fall outside the empirical scope of this thesis.

1.4.3 Framework Scope, Prompting and Empirical Boundaries

The thesis does not determine the correct policy on physical reconstruction at either case site. Decisions on whether the Buddhas of Bamiyan or the Temple of Bel should be physically rebuilt belong to the relevant cultural, governmental and community authorities. The contribution made here is to the evaluation of AI-mediated visualisations of these sites, not to the adjudication of reconstruction policy.

All image generation in the empirical chapters follows a structured prompting procedure rather than free-form prompt writing. The Dynamic Prompt Blueprint, formalised in Chapter 4, defines the prompt slots through which documentary, photographic, archaeological and architectural evidence is translated into generation instructions. The Prompt Sufficiency Index is used to assess whether each prompt instance is sufficiently specified before the resulting output is treated as valid for evaluation. This forms part of HARF's internal procedural standardisation. The thesis does not claim to establish field-wide standards for prompt-based heritage reconstruction. It defines and tests the prompt structure used within HARF, documents prompt decisions through paradata and provides worked prompt examples in the relevant appendices. At Bamiyan, the schema is populated with figural, Gandhāran, niche and surface-condition evidence. At Palmyra, the same top-level structure is re-instantiated with architectural evidence from the Temple of Bel.

The empirical reach of the thesis is also bounded. Community-grounded evaluation is conducted for Bamiyan through primary respondent data from Afghan participants; no equivalent community study is conducted for Palmyra. Computational similarity is applied within the Bamiyan case and is not extended to Palmyra, where the transferability test is questionnaire-based. Expert assessment is interdisciplinary at Phase I and specialist at Phase II for Bamiyan, while the Palmyra block is completed by a subset of the interdisciplinary panel rather than by a dedicated Palmyrene specialist panel. These boundaries define the conclusions that can be drawn from each case.

1.5 Research Questions

The research is structured by four questions.

RQ1: How can generative AI heritage reconstruction be constrained by documentary, photographic, archaeological and architectural evidence so that source material, prompt construction, model decisions and output selection form an auditable record?

RQ2: What integrated evaluation procedure can assess generative AI heritage reconstructions through evidence-anchored prompting, prompt sufficiency assessment, paradata recording, computational pre-evaluation and structured expert review?

RQ3: How do computational similarity, structured expert judgement and community-grounded authenticity assess different dimensions of validity in AI-mediated reconstruction at Bamiyan?

RQ4: To what extent can an evaluation framework developed through a figural Gandhāran case be re-instantiated for a monumental architectural case without redesigning its top-level schema and where do typology-dependent failure domains emerge?

RQ1 is addressed at the level of an initial evidence-constrained reconstruction workflow in Chapter 3 and formalised within the Heritage-Aligned Reconstruction Framework in Chapter 4. RQ2 is addressed by the framework developed in Chapter 4 and is empirically validated through Chapter 5. RQ3 is addressed through paired computational and expert evaluation in Chapter 5 and through community-grounded assessment in Chapter 6, each register being treated as a distinct source of evaluative evidence. RQ4 is addressed through bounded cross-site testing at the Temple of Bel in Chapter 7.

1.6 Methodological Architecture

The methodological architecture is built around the progression from evidential constraint to evaluative validation. The initial reconstruction workflow establishes the limits of unstructured prompting and informal comparison. HARF then formalises the evaluative procedure by integrating evidence-anchored prompting, prompt sufficiency assessment, paradata recording, computational pre-evaluation and structured expert review. The framework is subsequently tested across computational and expert registers at Bamiyan, extended to community-grounded authenticity through Afghan respondent evidence and re-instantiated at Palmyra to examine bounded transferability across heritage typologies.

1.6.1 Evaluative Registers

The term “register” is used here to distinguish forms of evaluation that produce different kinds of evidence. Computational similarity identifies visible departures from reference material but cannot adjudicate iconographic, historical or cultural plausibility. Expert judgement addresses archaeological, stylistic and proportional validity. Community-grounded assessment addresses authenticity, absence and acceptable representation from the standpoint of affected heritage communities. These registers are connected through HARF, but they do not perform the same evaluative work.

Three principles organise the method. Evidence precedes generation. Evaluation is distributed across distinct registers. Transferability is tested rather than assumed. These principles connect the research questions to the heritage standards against which GenAI reconstruction is assessed.

1.7 Original Contributions

The thesis makes four original contributions.

Contribution 1 — An evidence-constrained AI reconstruction workflow for cultural heritage:

The first contribution is an initial evidence-constrained workflow linking source selection, prompt construction, image generation and expert comparison. Its value lies not in establishing AI-generated images as reliable reconstructions, but in showing why unstructured prompting and informal visual comparison are insufficient for heritage reconstruction. This contribution is developed in Chapter 3 and addresses the evidential-control problem that motivates RQ1.

Contribution 2 — The Heritage-Aligned Reconstruction Framework:

The second contribution is HARF, a standardised AI-specific evaluation framework combining the Dynamic Prompt Blueprint, Prompt Sufficiency Index, paradata recording, computational pre-evaluation and structured expert review. HARF distinguishes visual plausibility from evidential plausibility and supplies the methodological structure for the empirical chapters. This contribution is developed in Chapter 4 and provides the framework response to RQ2.

Contribution 3 — A multi-register evaluation of AI-mediated reconstruction at Bamiyan:

The third contribution is an empirical account of reconstruction validity across computational, expert and community-grounded registers. It combines computational and expert evaluation of 245 Bamiyan reconstruction candidates with primary respondent evidence from 62 Afghan participants on authenticity, absence and acceptable reconstruction. The Bamiyan evaluation shows that image similarity, specialist judgement and community-grounded authenticity assess different aspects of

validity and cannot be collapsed into a single score or criterion. This contribution is developed across Chapters 5 and 6 and addresses RQ3.

Contribution 4 — A bounded cross-site transferability test of HARF:

The fourth contribution is the re-instantiation of HARF for a typologically different heritage case: the Temple of Bel at Palmyra. The Palmyra test examines whether the evaluative schema can transfer beyond its Bamiyan development case while identifying typology-dependent failure domains. It shows that the framework can be re-applied at the level of evaluative schema while the dominant failure domains shift from figural micro-grammar at Bamiyan to architectural and decorative micro-grammar at Palmyra. HARF generalises at the level of evaluative schema, evidentiary logic, prompt-structure transfer, paradata discipline and failure-domain diagnosis. It does not prove that AI outputs are historically accurate across all heritage sites. This contribution is developed in Chapter 7 and addresses RQ4. Figure 1 summarises the thesis architecture by mapping the main research problem to the literature-derived gaps, research questions, contribution chapters, bounded generalisation test and associated research outputs.

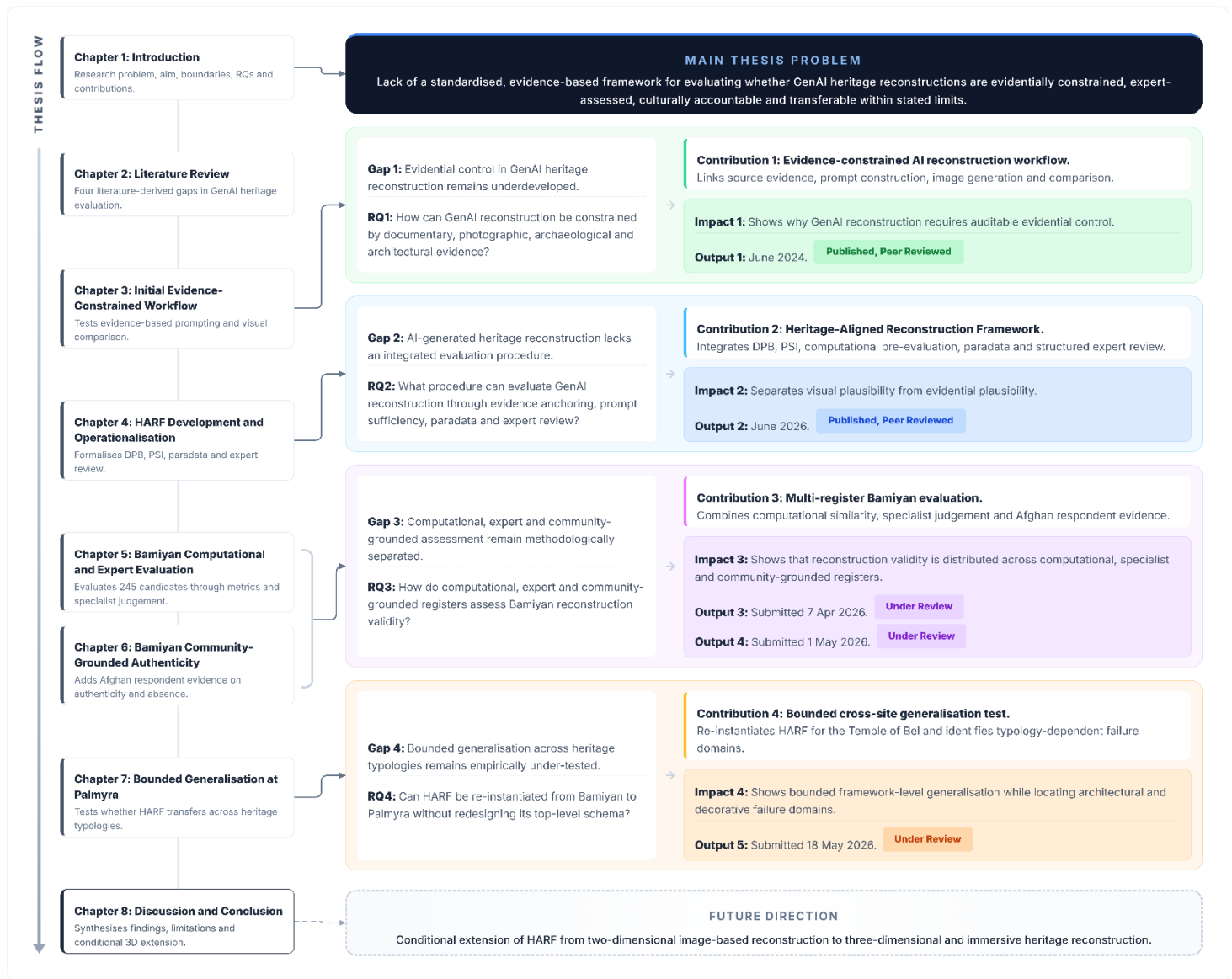


Figure 1. Author-generated diagram of the thesis architecture showing the relationship between the research problem, literature-derived gaps, research questions, contribution chapters and bounded generalisation test.

1.8 Publications and Outputs

The following publications and submitted manuscripts arose from the doctoral research. They are listed as research outputs and do not define the thesis structure. Where relevant, chapters acknowledge their relationship to these outputs and the front-matter statement records the candidate's contribution to co-authored work.

Output 1: Arzomand, K., Rustell, M. and Kalganova, T. (2024) 'From ruins to reconstruction: harnessing text-to-image AI for restoring historical architectures', *Challenge Journal of Structural Mechanics*, 10(2), pp. 69–85. doi:10.20528/cjsmec.2024.02.004. Published, peer-reviewed. (Arzomand et al., 2024)

Output 2: Arzomand, K., Kalganova, T. and Rustell, M. (2026) 'HARF: A human–AI collaborative framework for cultural heritage reconstruction with expert-guided multi-platform generative AI and systematic prompt engineering', *Digital Applications in Archaeology and Cultural Heritage*, 42, e00554. doi:10.1016/j.daach.2026.e00554. Published, peer-reviewed. (Arzomand et al., 2026)

Output 3: Arzomand, K. and Kalganova, T. 'Evaluating generative AI reconstructions of the Bamiyan Buddhas: computational similarity and expert archaeological assessment'. Submitted to *Journal of Cultural Heritage*. Under review. (Arzomand & Kalganova, 2026b)

Output 4: Arzomand, K., Kalganova, T. and Ritondale, M. 'Community-grounded authenticity at Bamiyan: Afghan perspectives on AI-mediated heritage reconstruction'. Submitted to *International Journal of Heritage Studies*. Under review. (Arzomand & Kalganova, 2026a)

Output 5: Arzomand, K. and Kalganova, T. 'Generalising HARF beyond Bamiyan: a bounded cross-site evaluation of the Temple of Bel, Palmyra'. Submitted to *Heritage* (MDPI). Under review. (Arzomand & Kalganova, 2026c)

1.9 Structure And Outline of the Thesis

The argument proceeds through eight chapters.

Chapter 1 introduces the research problem, aim, boundaries, research questions, methodological architecture, original contributions and research outputs.

Chapter 2 reviews the literature on heritage reconstruction standards, pre-GenAI reconstruction workflows, generative AI heritage reconstruction, validation, expert evaluation, community-grounded authenticity and cross-site transferability. It identifies the four literature-derived gaps that structure the thesis.

Chapter 3 presents the initial evidence-constrained AI reconstruction workflow. It shows why unstructured prompting and informal expert comparison are insufficient for accountable heritage reconstruction.

Chapter 4 develops HARF as the methodological framework of the thesis. It defines the Dynamic Prompt Blueprint, Prompt Sufficiency Index, paradata protocol, computational pre-evaluation pipeline and structured expert review procedure.

Chapter 5 operationalises the computational and expert registers of HARF at the Western Buddha of Bamiyan. It evaluates 245 reconstruction candidates through image-quality and feature-matching metrics, interdisciplinary expert screening, specialist assessment and inter-rater reliability analysis. The chapter identifies where algorithmic similarity scores and expert judgement diverge.

Chapter 6 operationalises the community-grounded register at Bamiyan. It analyses respondent evidence from 62 Afghan participants on authenticity, absence, reconstruction preferences and acceptable representation.

Chapter 7 tests bounded cross-site transferability at the Temple of Bel, Palmyra. It re-instantiates the Dynamic Prompt Blueprint for an architectural typology and identifies the shift from figural micro-grammar failure at Bamiyan to architectural and decorative failure domains at Palmyra.

Chapter 8 synthesises the findings against the four literature-derived gaps, research questions and contributions. It states the consolidated limitations of the thesis and identifies future work, including the conditional extension of HARF to three-dimensional generative outputs.

Chapter 2

Literature Review

2.1 Introduction

This chapter reviews the literature necessary to establish the central problem addressed by this thesis: the absence of a consistent and reproducible evidence-based and expert-informed evaluation framework for generative artificial intelligence (GenAI) reconstruction of destroyed and contested cultural heritage. Cultural accountability is treated here as one of the evaluation problem's core dimensions. Reconstruction validity cannot be established through visual plausibility or technical correspondence alone; it also depends on the cultural standpoint from which authenticity, loss and acceptable reconstruction are assessed. The chapter is not a general survey of digital heritage or generative AI. Recent surveys already map those broader fields (Colace et al., 2025; Paolanti et al., 2026; Spennemann, 2024; Xu et al., 2025). This review uses them only where they help define the evaluation problem addressed by the thesis. It is organised around four research gaps addressed through the Heritage-Aligned Reconstruction Framework (HARF), developed and validated through the Bamiyan case and tested for bounded cross-site generalisation through Palmyra.

The review proceeds from three claims. The first is that the central problem is no longer generation alone, but the evaluation of generated reconstruction candidates against heritage-specific standards of evidence, expertise and cultural legitimacy. Recent work has shown that generative systems can produce visually coherent reconstruction candidates from textual, photographic and multimodal inputs (Arzomand et al., 2024; Delgado-Martos et al., 2026; Shih, 2025; Wei et al., 2026). What remains unresolved is whether such candidates can be evaluated as historically plausible, evidentially constrained and culturally legitimate heritage reconstructions. The second is that evaluation in this domain cannot be reduced to one type of evidence. Computational similarity, expert archaeological judgement and community-grounded authenticity each assess different dimensions of validity. The third is that a framework developed for one heritage typology cannot be assumed to transfer to another without empirical testing. Most reconstruction studies remain case-specific and the literature provides limited guidance on cross-typology generalisation.

The chapter proceeds in four movements. Sections 2.2 and 2.3 establish the standards and pre-GenAI reconstruction practices against which AI-mediated reconstruction must be assessed. Section 2.4 turns to GenAI heritage reconstruction, including text-to-image, multimodal and text-to-3D work and identifies the evidential-control gap. Section 2.5 reviews validation and expert-grounded evaluation, showing why computational and disciplinary assessment must be structured but cannot exhaust reconstruction validity. Sections 2.6 and 2.7 address cultural authenticity and transferability, before Section 2.8 synthesises the four literature-derived gaps and positions the thesis contribution. Each substantive section contributes to the cumulative construction of the four gaps. The published and submitted works are positioned as responses to problems established through the literature.

The literature reviewed in this chapter was identified through a targeted, problem-led search rather than a systematic review. Searches were conducted in ScienceDirect, Web of Science and Google Scholar and were supplemented by relevant heritage-sector sources, including UNESCO, ICOMOS and digital-heritage charter material. The search combined heritage terms, including cultural heritage reconstruction, digital heritage, virtual archaeology, paradata and authenticity, with computational terms, including generative AI, text-to-image, diffusion models, GANs, text-to-3D, image similarity, benchmarking and evaluation. Site-specific searches were also conducted for Bamiyan and Palmyra.

Foundational conservation standards, digital-heritage charters and core methodological sources were included without date restriction, because they remain central to evidential accountability, authenticity, paradata and reconstruction validation. Contemporary literature on GenAI, AI evaluation, computational reconstruction and AI-mediated heritage practice was prioritised from 2020 to May 2026, reflecting the period most directly relevant to the recent application of deep generative models to heritage reconstruction. Earlier studies were included where they were methodologically foundational or directly relevant to Bamiyan, Palmyra, computational image comparison, photogrammetry, HBIM or digital reconstruction practice.

Studies were included where they addressed at least one of the following: AI-mediated or generative heritage reconstruction; evidence documentation, paradata or authenticity in digital heritage; computational validation or image-similarity assessment; expert evaluation of reconstruction outputs; community-grounded or participatory heritage assessment; or transferability across heritage contexts. Studies were excluded where they concerned general AI image generation without heritage relevance, or heritage visualisation without an evaluative, evidentiary or AI-specific dimension. Tables 1 and 2 were compiled to show how existing work distributes evaluation across separate functions: evidence control, paradata, computational comparison, expert judgement, benchmarking, community authority and transferability.

2.2 Heritage Standards and Evidential Accountability

Digital reconstruction of cultural heritage is not a technical activity governed by engineering tolerances alone. It is an interpretive activity governed by established standards of documentation and accountability. Those standards have developed over six decades through four principal charters and a substantial methodological literature on paradata, provenance and reconstruction uncertainty. Together, they define the evidentiary baseline against which digital, virtual and AI-mediated reconstruction must be assessed. This section identifies the standards that digital heritage reconstruction must satisfy and shows that, although those standards provide clear principles for documentation, transparency and uncertainty, they have not yet been translated into an operational procedure for generative-AI outputs.

2.2.1 Distinguishability, Authenticity and Transparency

The Venice Charter (ICOMOS, 1964) provides a foundational principle for the evidentiary discipline required in reconstruction. Its relevance for GenAI evaluation lies in its insistence that restoration must not falsify historical evidence and that additions or replacements must remain distinguishable from the original material. Article 9 requires indispensable additional work to remain distinct from the architectural composition and to bear a contemporary stamp; Article 12 requires replacements of missing parts to integrate harmoniously while remaining distinguishable from the original. In digital reconstruction, this principle becomes a requirement that documented, inferred and conjectural elements should not be visually collapsed into a single authoritative image. For GenAI reconstruction, the problem is sharper: a generated image can present speculative content with the same visual confidence as evidence-based content. Evaluation must therefore ask not only whether an output appears plausible, but whether its evidentiary status is visible and traceable.

The Nara Document on Authenticity (ICOMOS, 1994) extends the evidentiary question raised by Venice by treating authenticity as more than material fabric alone. Authenticity may reside in form and design, materials and substance, use and function, traditions and techniques, location and setting, language and spirit or feeling. Nara also rejects a universal test of authenticity. Article 11 provides that judgements concerning heritage values and the credibility of information sources may differ between cultures and even within the same culture. This is directly relevant to AI-mediated reconstruction of sites such as Bamiyan and Palmyra. A reconstruction may correspond visually to archival material or to external expert expectations yet still fail to reflect the cultural meanings through which the site is

understood by connected communities. Evaluation therefore cannot be limited to technical similarity or archaeological plausibility; it must also ask whose standards of authenticity are being applied.

The London Charter for the Computer-Based Visualisation of Cultural Heritage (Denard, 2009) brings these principles into the specific context of digital visualisation. Its importance lies less in the fact that it concerns digital heritage than in the discipline it imposes on the relationship between evidence, interpretation and output. The Charter requires the aims, sources, methods and evidential basis of a visualisation to be made explicit and it treats evaluation as part of responsible visualisation practice rather than as an optional supplement. Its central methodological contribution is the concept of *paradata*: the record of the interpretive decisions through which evidence is transformed into a visual reconstruction. Metadata describes the digital artefact; *paradata* explains how and why that artefact took its final form. For GenAI reconstruction, this distinction is critical. Prompt design, model selection, seed choice, iteration history and output filtering are not merely technical settings; they are interpretive decisions that affect the historical claim made by the generated image. Unless those decisions are recorded, an AI-generated reconstruction cannot be audited as a heritage reconstruction. It remains only a visual output.

The Seville Principles (López-Menchero Bendicho & Grande, 2011, 2017) develop the London Charter for the specific context of virtual archaeology. Their importance for this thesis lies in the insistence that virtual reconstruction is not justified by visual coherence alone. It must be interdisciplinary, historically rigorous, transparent in its evidential basis and explicit about uncertainty. In archaeological reconstruction, this means that documented elements, inferred elements and conjectural completions should not be presented as if they had the same evidential status. For GenAI reconstruction, the requirement is particularly important because the generated image can normalise uncertainty: missing, weakly evidenced or speculative details may appear as visually complete as documented ones. The Seville Principles therefore provide the immediate methodological bridge between digital heritage standards and the evaluation problem addressed in this thesis.

Taken together, the Venice Charter, the Nara Document, the London Charter and the Seville Principles establish the standards that any AI-mediated heritage reconstruction must satisfy: distinguishability, evidential traceability, transparency about inference, explicit treatment of uncertainty and culturally situated authenticity. They do not, however, specify how these standards should be implemented when reconstruction is produced through prompt-based generative systems. HARF addresses this procedural gap by embedding these standards within the generation, documentation and evaluation of GenAI reconstruction candidates.

2.2.2 Paradata and Uncertainty

Paradata is the principal means by which transparency in digital heritage reconstruction becomes auditable. In virtual heritage practice, paradata has been established as a methodological requirement because records of reconstruction reasoning can be structured, archived and associated with the resulting visualisation (Bentkowska-Kafel & Denard, 2012). The distinction from metadata is essential. Metadata describes the finished digital artefact, including its date, format, authorship and software environment. Paradata records the interpretive process through which that artefact was produced, including the sources consulted, inferences made, alternatives rejected and uncertainties retained. It therefore differs from metadata both in content and in timing: metadata describes the result, whereas paradata records decisions made during reconstruction. In this sense, paradata performs a function comparable to citation, linking a visual claim to the evidence and reasoning on which it depends.

Subsequent literature has refined this principle while exposing its practical difficulty. Paradata has often been theorised more strongly than it has been implemented, with documentation practices in digital heritage remaining inconsistent and selective (Huvila, 2022a). The problem is not simply the absence of documentation, but its integration into the workflow. Paradata is difficult to reconstruct once modelling decisions have already been made. If documentation is not built into the process from the outset, the reasoning behind a reconstruction is easily lost or reduced to retrospective description. The same requirement appears in reproducible archaeological reporting, where transparency depends not only on final explanation, but on recording the decisions, tools and processes through which research outputs are produced (Marwick, 2017, 2022).

Practical implementation depends on explicit workflow design. The 3D reconstruction work at Copán demonstrates that systematic paradata recording can be built into the reconstruction process when documentation is treated as part of modelling practice rather than retrospective annotation (Richards-Rissetto & von Schwerin, 2017). The Copán case is methodologically instructive because it shows that the difficulty of paradata recording is not inherent to paradata as a concept but arises when documentation is separated from the routine decisions of reconstruction practice. Built into the workflow, paradata supports auditability; added after completion, it becomes retrospective description.

A related literature addresses the visual communication of uncertainty. Uncertainty in digital reconstruction should be treated as a visible attribute of the reconstructed object, not confined to accompanying documentation (Pietroni, 2021; Pietroni & Ferdani, 2021). This follows from the distinguishability requirement in the Venice Charter and the transparency principle in the London Charter: documented, inferred and conjectural elements should not be presented as if they carried

the same evidential weight. Uncertainty-visualisation schemes that map evidential confidence onto colour or transparency in three-dimensional models extend this principle into visual form (Apollonio et al., 2018). The specific visual convention is less important for this thesis than the principle it demonstrates: a reconstruction must preserve the connection between visual form, evidential basis and interpretive confidence.

For GenAI reconstruction, these requirements apply directly. Prompt construction, model selection, seed choice, refinement cycles and output filtering are not neutral technical settings; they are interpretive decisions that shape the historical claim made by the generated image. At the same time, a generated image can present an apparently complete visual field even where different parts of the output rest on different degrees of evidence. Weakly evidenced details, unsupported model additions and historically grounded features may therefore appear with the same visual authority. Existing paradata and uncertainty-visualisation practices were not designed for prompt-based generation, but their underlying demand remains applicable: the reconstruction process must be recorded in a form that allows the output to be examined, challenged and, where possible, reproduced, while the evidential status of its elements remains visible.

2.2.3 Implications for GenAI evaluation

The literature reviewed in this section establishes a clear evidentiary standard for digital heritage reconstruction. Reconstruction should distinguish original, inferred and conjectural elements; authenticity should be culturally situated; interpretive decisions should be documented through paradata; and uncertainty should remain visible rather than hidden by the coherence of the final visualisation. These principles provide the standard against which AI-mediated reconstruction must be assessed.

The difficulty is that these standards were developed for workflows in which the relationship between source evidence, interpretive decision and rendered output remains traceable through human modelling practice. Prompt-based GenAI complicates that relationship. The Venice Charter, the Nara Document, the London Charter and the Seville Principles predate text-to-image, text-to-3D and large generative-model reconstruction. They do not specify how prompt construction, model choice, seed selection, refinement cycles or output filtering should be recorded as paradata. Nor do they explain how evidence-based and conjectural elements can be distinguished within a generated image whose visual features may derive from source evidence, prompt design, stochastic generation and model-internal priors.

The issue, therefore, is not the absence of heritage standards. It is the absence of an operational procedure for applying those standards to GenAI reconstruction. This section establishes the evidentiary basis against which the following sections assess GenAI reconstruction, validation practice and cultural accountability. The HARF workflow developed in Chapter 4 responds to this problem by translating these standards into evidence-anchored prompting, paradata logging, computational pre-evaluation and expert review.

2.3 Evidence-Constrained Reconstruction Before GenAI

Reconstruction from incomplete evidence did not begin with generative AI. Archaeology, architectural history and digital heritage have long worked with damaged, fragmentary and partially documented sites through photogrammetry, laser scanning, archival image analysis and Historic Building Information Modelling. These methods show that reconstruction is not simply visual production. It involves measurement, archival evidence and interpretation and its evidential strength depends on what survives, what was documented before loss and what must be inferred.

Photogrammetry, laser scanning and HBIM operate under evidential conditions different from those of prompt-based image generation. They operate through surviving fabric, survey data, archival comparison and structured parametric modelling, whereas GenAI can produce visually coherent candidates without preserving a stable relation between source evidence, inference and output. Their relevance to this review lies in the prior reconstruction tradition they establish: reconstruction can be evidence-constrained, but only where its sources, assumptions and uncertainties remain traceable. Where material survives, reconstruction can be constrained by measured fabric, survey data and material traces. Where material has been destroyed, reconstruction depends on archival records, comparative evidence and expert interpretation. The problem addressed in this thesis is therefore not the absence of reconstruction methods, but the absence of a transferable procedure for evaluating two-dimensional GenAI reconstruction candidates against evidence, inference and uncertainty.

2.3.1 Photogrammetry, laser scanning and archival reconstruction

Photogrammetry and laser scanning have provided digital heritage with established methods for recording surviving fabric and producing metric documentation of damaged or partially surviving monuments. Standard workflows combine image acquisition, control-point networks, multi-view reconstruction, dense matching, mesh generation and validation against survey or reference data (Maiwald et al., 2017; Remondino, 2011). These methods are strongest where the object or monument survives in sufficient form to be measured. In such cases, the reconstruction remains tied

to material evidence and the assessment concerns the correspondence between the digital model and recorded remains.

Destroyed heritage raises a different problem. Photogrammetry can record remaining fragments, niches, architectural traces and surrounding landscape, but it cannot directly recover a lost object once the object has disappeared. The missing form must be reconstructed from pre-destruction photographs, measured drawings, survey records and comparative material. Retrospective photogrammetry from archival images can establish dimensional information, but its reliability depends on the completeness, calibration and viewpoint diversity of the surviving photographic record (Grün et al., 2004; Maiwald et al., 2017). The limitation is evidential as much as technical: the reconstruction is constrained by the record from which it is inferred.

The Western Buddha of Bamiyan illustrates this constraint. Post-destruction documentation records the cliff, the niches and surviving fragments, but it cannot directly reconstruct the lost sculpted body. The dimensional framework for the figure has therefore depended on pre-2001 photography and survey material, including photogrammetric reconstruction from archival imagery (Grün et al., 2004). Prior conservation work has developed the documentary record of the niches and surrounding fabric, including virtual environments used to examine stabilisation, partial reconstruction and anastylosis before physical intervention (Petzet, 2009; Toubekis et al., 2017). Remote-sensing work has extended the record to the wider Bamiyan cultural landscape (Hammer et al., 2018).

The pre-destruction record supplies the visual and dimensional basis against which any figural reconstruction must be judged. Photographic campaigns and measured surveys recorded proportional relationships, drapery rhythms and niche morphology, while Japanese cultural missions in the 1970s produced detailed drawings and architectural surveys of the niche geometry (Petzet, 2009; Toubekis et al., 2017; Yamauchi & Bunka, 2012). Art-historical scholarship on Bamiyan and Gandhāran Buddhism supplies the stylistic and iconographic context for evaluating the figure, including robe structure, sculptural lineage and Buddhist visual conventions (Klimburg-Salter, 1989; Rienjang & Stewart, 2023).

The Bamiyan record is uneven. It is stronger for dimensions, niche geometry and broad proportional relationships than for damaged or poorly documented figural details. Drapery articulation, cranial morphology, hand gesture and surface treatment remain interpretive domains. These are precisely the features for which visual generation cannot be accepted without expert evaluation.

The Bamiyan case therefore shows the strength and limit of archival reconstruction. The available record can constrain dimensions, proportions and broad visual structure, but it cannot remove the interpretive work required for damaged or poorly documented figural features. This record functions

as an evidential baseline against which two-dimensional GenAI reconstruction candidates can be assessed. Such features require expert evaluation rather than visual generation alone.

2.3.2 HBIM and Geometric Validation

A second strand of evidence-constrained reconstruction has developed around Historic Building Information Modelling. HBIM adapts building information modelling to the geometric and semantic irregularities of historic structures, organising the digital model through structured object libraries linked to documentary sources and conservation records (Costantino et al., 2023; Murphy et al., 2013). Scan-to-HBIM workflows extend this approach by using terrestrial laser scanning or photogrammetric point clouds as input, segmenting the captured geometry and converting it into parametric heritage objects (Quattrini et al., 2015; X. Yang et al., 2020).

This literature has produced a more formal validation culture than many visual reconstruction workflows. Geometric accuracy can be assessed by comparing the HBIM model with the underlying point cloud or survey-grade reference data, while semantic accuracy can be assessed through the classification of architectural elements. These procedures make parts of the validation process measurable, reproducible and reportable. Recent reviews nevertheless identify persistent difficulties in this culture. Pocobelli et al. document the gap between HBIM information modelling in heritage practice and the level of automation reached in mainstream BIM (Pocobelli et al., 2018), while Radanovic et al. report the same gap from a geospatial perspective by examining how survey-grade data are translated into HBIM-ready geometry (Radanovic et al., 2020). Yang et al. extend the diagnosis to semantic enrichment, showing that the classification of architectural elements remains uneven across heterogeneous historic typologies (X. Yang et al., 2020). More recent reviews then connect this to interoperability across institutional contexts: Fernandes Dionizio and Dezen-Kempter examine the workflow integration problem (Fernandes Dionizio & Dezen-Kempter, 2025) and Cui and Wu review how AI-mediated tools are entering HBIM practice without yet supplying the validation procedures that heritage application demands (Cui & Wu, 2025).

The strength of HBIM validation is also its boundary. It is strongest where the object of assessment is geometric or semantic correspondence. Deviation analysis can show whether a model follows a point cloud and semantic checks can show whether architectural elements have been classified consistently. They do not determine which historical phase is being represented, how absent elements have been inferred, how uncertainty has been marked, or how survey data have been combined with archival and documentary sources. Those decisions remain matters of conservation reasoning, documentary interpretation and expert judgement, even where the geometric model is well controlled.

The same boundary applies to computational assessment in GenAI image reconstruction. Image similarity, feature matching or proportional comparison can help identify candidates that depart from known dimensions, structural outlines or visible reference features. They cannot establish whether generated details are historically plausible in style, iconography, material treatment or conjectural completion. Computational assessment can therefore support evaluation where reference material exists, but it cannot replace expert judgement about historical plausibility. This limitation becomes important for the GenAI literature reviewed next, where image-based comparison is often treated as stronger evidence than heritage reconstruction can sustain.

2.3.3 Case-Bound Reconstruction Practice

Across photogrammetric, archival and HBIM workflows, destroyed-heritage reconstruction is usually developed within the evidential conditions of a particular site. Each project assembles its own documentary record, resolves missing or uncertain elements and evaluates the result through the expertise available to that project. This proximity to site-specific evidence is a strength of the pre-GenAI tradition. It is also the reason its procedures are difficult to transfer without further formalisation.

The Temple of Bel at Palmyra illustrates the point. Reconstructions of the temple have drawn on professional and crowdsourced panoramic imagery to produce three-dimensional models, while later work has developed immersive environments for interaction with the reconstructed site (Rihani, 2023; Wahbeh et al., 2016). Yet particular architectural features remain open to revision. The uncertainty concerns not only the overall form of the temple, but also specific architectural and decorative elements for which the photographic or comparative record remains incomplete. The temple's crenellations, for example, have been reconsidered through comparative Hellenistic-Roman evidence, showing that even familiar reconstructions of well-documented monuments may rest on contestable assumptions (Schmidt-Colinet & Seigne, 2022). A dense visual or geometric model does not remove the need to account for inference. The result is not a single self-validating reconstruction, but a set of reconstruction claims whose evidential basis must be assessed.

Bamiyan shows the same case-bound logic from a different direction. Virtual reconstructions of the niches have been used to examine restoration alternatives and communicate uncertainty before physical intervention, while remote-sensing studies have extended the documentary record to the wider cultural landscape (Hammer et al., 2018; Toubekis et al., 2017). These works organise available evidence and make restoration scenarios examinable, but their procedures remain tied to the specific conservation questions for which they were produced.

The case-bound character of this tradition is not a methodological failure. It reflects the evidential specificity of heritage reconstruction. A destroyed Buddhist figure, a partially surviving architectural sanctuary and a documented cultural landscape do not present the same evidential problems. The difficulty arises when prompt-based generative systems enter this setting. Such systems can produce many visually plausible candidates, but the output does not show how archival evidence, inference, uncertainty and expert judgement have shaped the result.

The case-bound nature of this practice also explains why transferability cannot be assumed. A procedure developed for a destroyed Buddhist figure does not automatically transfer to a monumental architectural sanctuary, because the evidential categories and likely failure domains differ. The transferability question is developed in Section 2.7.

Pre-GenAI reconstruction practice therefore provides strong models of site-specific documentation, modelling and expert review. It does not provide a standardised procedure for evaluating prompt-based generative outputs against documentary evidence, expert judgement and uncertainty. The next section examines whether the GenAI heritage reconstruction literature has supplied such a procedure.

2.4 Generative AI and Evidential-Control Gap

The use of generative AI in cultural heritage reconstruction has expanded rapidly in recent years. The relevant literature includes GAN-based image synthesis and restoration, diffusion-based text-to-image systems such as Stable Diffusion, DALL·E, Midjourney and Imagen, multimodal and video-generation systems, including Sora and Gemini, that can be used within visual reconstruction workflows and emerging text-to-3D, neural radiance field and Gaussian-splatting methods. These systems can produce visually plausible reconstruction candidates from textual, photographic or multimodal inputs. The unresolved question is whether such candidates can be evaluated as historically defensible reconstructions, not simply as coherent visual outputs.

This section reviews the GenAI heritage literature against the evaluative requirements established in Sections 2.2 and 2.3: evidential traceability, documentation of inference, expert assessment, uncertainty and transferability beyond a single case. The gap addressed here concerns evidential control. Existing GenAI heritage studies demonstrate generation capacity, but they rarely provide a structured procedure for showing how archival sources, prompt decisions, model behaviour, iteration and output selection shape the reconstruction.

2.4.1 GANs and AI-Assisted Reconstruction

Earlier GAN work in heritage focused mainly on image enhancement, conditional synthesis and restoration. Conditional GAN architectures have been used to generate architectural elevations from colour-pattern inputs and to enhance degraded heritage photographs through super-resolution (Navarro-Mateu et al., 2021; Nayak & Balabantaray, 2021). These applications show that generative models can support visual recovery where the archival record is degraded or where controlled image synthesis is required. Their evaluation conventions remain closer to computer vision than to heritage reconstruction. Perceptual quality, image clarity and super-resolution metrics do not establish whether the generated or enhanced output remains evidentially faithful to the source material.

The ARCHER project represents a more developed AI-assisted reconstruction pipeline. Recent work describes a sequence of methods including GAN-based restoration of Greek temples, transformer-supported text and image processing and three-dimensional reconstruction using Gaussian Splatting and SuGaR for Romanesque-Mudéjar churches (Delgado-Martos et al., 2026). The value of ARCHER lies in its attempt to build an integrated AI workflow across several architectural cases, rather than treating generation as an isolated image task.

Its evaluation, however, remains primarily technical. The reported criteria concern reconstruction quality, automation efficiency and pipeline performance. These criteria are appropriate for assessing the functioning of an AI system, but they do not document the evidential route from archival source to generated output. Nor do they show how uncertainty is represented, how interpretive decisions are recorded, or how expert judgement enters the assessment of historical plausibility. The GAN and AI-assisted reconstruction literature therefore establishes technical feasibility, but leaves evidential accountability underdeveloped.

2.4.2 Text-To-Image and Multimodal Reconstruction

Text-to-image and multimodal systems have moved GenAI heritage work closer to the reconstruction problem. An early multi-site feasibility study used Midjourney and DALL-E to generate visual reconstruction candidates for ten historical sites, including Bamiyan and Palmyra, from iteratively refined prompts derived from historical documentation (Arzomand et al., 2024). The outputs were assessed through visual comparison with archival photographs and a limited set of image-quality metrics applied to the Bamiyan Giant Buddha, including SSIM, MSE, PSNR and MAE. The study demonstrated that text-to-image systems can produce plausible visual candidates where documentary material is available. It also exposed the limits of unstructured prompting and image-

quality metrics as an evaluation procedure. Expert validation, prompt formalisation and the systematic treatment of evidential uncertainty remained unresolved.

Comparable patterns appear in recent single-site and multimodal studies. Text-to-image feasibility work demonstrates visually coherent reconstruction outputs but relies mainly on visual comparison or limited expert inspection (Shih, 2025). Multimodal AIGC work on Qing-dynasty interior canopies identifies the divergence between visual plausibility and evidential adequacy but does not convert that distinction into a transferable evaluation procedure (Wei et al., 2026). Cultural-cognitive approaches emphasise culturally aware prompting without specifying how generated outputs should be assessed across cases (Li et al., 2025). Work on AI perception and Etruscan art raises the problem of culturally specific visual interpretation but does not provide a reconstruction-evaluation framework (Forte, 2024). Review work on generative technologies in digital museums similarly records the growth of GenAI heritage applications while showing that evaluation remains largely visual, perceptual or experience-based (Xu et al., 2025).

The pattern is consistent. Text-to-image and multimodal systems can generate plausible heritage images across different contexts, but the evaluation of those images remains under-structured. Visual comparison, image-quality metrics and limited expert inspection do not show, on their own, how documentary evidence, prompt design, model behaviour, iteration and output selection have shaped the final reconstruction candidate. The literature therefore demonstrates generation capacity while leaving evidential control unresolved.

2.4.3 Three-Dimensional Generation as a Boundary Case

Three-dimensional generation intensifies the evidential problem already visible in image-based reconstruction. Text-to-3D research has moved through several methodological generations: optimisation-based approaches such as Score Distillation Sampling (Lin et al., 2022; Poole et al., 2022; Z. Wang et al., 2023), multi-view diffusion (R. Liu et al., 2023; Shi et al., 2024), feed-forward image-to-3D models (Tochilkin et al., 2024; Xu et al., 2024) and large reconstruction models such as LRM (Hong et al., 2024). NeRF and Gaussian-splatting methods have also entered heritage documentation and reconstruction, where they offer strong rendering quality and, in some settings, reduced acquisition demands. Their value for historical reconstruction still depends on how far the generated geometry can be checked against photogrammetric, laser-scanning or archival reference material (Clini et al., 2024; Kerbl et al., 2023; Murtiyoso & Grussenmeyer, 2023). These methods expand what can be generated, reconstructed or rendered, but they do not resolve how evidence, inference and uncertainty should be represented within the generated object.

Evaluation in this literature remains largely model-centred. CLIP similarity, Chamfer distance, F-score, multi-view consistency, PSNR, LPIPS and user-preference studies can assess semantic alignment, geometric correspondence and rendered-view quality; benchmarks such as T3Bench and GPTEval3D add more structured comparison across models and prompts (Y. He et al., 2024; T. Wu et al., 2024). These measures are useful for comparing systems. They do not record the evidential basis of a generated heritage object, the interpretive decisions embedded in prompt construction, or the distinction between documented and conjectural elements. Moving into three dimensions therefore increases technical complexity before the evidential problem has been solved.

Three-dimensional generation extends the unresolved evidential problem from image-based plausibility to spatial reconstruction. Generated geometry, textures, missing volumes and interactive views would need to be traceable to material, archival or comparative sources, with conjectural elements and uncertainty recorded as paradata. The empirical chapters address the two-dimensional image case because the relation between source material, prompt construction, generated candidate, computational filtering, expert judgement and community-grounded authenticity can be examined within a controlled validation design. Chapter 8 returns to the extension of HARF into three-dimensional, volumetric and immersive reconstruction, where additional criteria are required for scale, volumetric consistency, viewpoint stability, spatial coherence, interaction and the separation of documented from conjectural content.

2.4.4 Benchmarks, diagnostics and policy frameworks

Recent benchmark, diagnostic and policy work has brought more structure to the evaluation of GenAI heritage systems, but it works at a different level from reconstruction assessment. The Time Travel Benchmark assesses large multimodal models on historical and cultural artefacts by scoring outputs against curated reference data across multiple historical domains (Ghaboura et al., 2025). Its contribution is cross-model evaluation of historical content. It does not assess a particular reconstruction against the evidence for a particular monument, nor does it document how sources, prompts, model behaviour and output selection produce a reconstruction candidate.

Bias-diagnostic work addresses a different weakness. The bias-loop framework developed by Foka and Griffin identifies how representational distortions can persist when AI-generated cultural content is absorbed into later training corpora (Foka et al., 2025; Foka & Griffin, 2024a, 2024b). This explains why heritage distortion cannot be treated as a temporary artefact of immature models. Repeated generation may reinforce culturally narrow or historically inaccurate representations. The framework

is valuable as a diagnosis of field-level risk, but it does not give a procedure for deciding whether one reconstruction is evidentially grounded, historically plausible or culturally accountable.

Policy and ethical literature set the terms for responsible use but remain largely principle-based. Spennemann maps the implications of generative AI for heritage practice and identifies trustworthy use as an open methodological problem (Spennemann, 2024). Paolanti et al. articulate a trustworthy-AI framework for cultural heritage applications, emphasising accountability, transparency and explainability (Paolanti et al., 2026). Colace et al. extend the discussion to copyright and intellectual-property questions that arise when AI-generated heritage content is produced and distributed (Colace et al., 2025). Westenberger and Farmaki consider how AI-mediated representation interacts with heritage protection more broadly, including the safeguarding of vulnerable sites (Westenberger & Farmaki, 2025). These concerns are necessary, but they do not specify how a reconstruction workflow should record prompt decisions, model choice, paradata, output filtering or expert review. The literature therefore recognises the need for responsible GenAI heritage practice but still leaves the project-level evaluation problem unresolved.

2.4.5 Training-Data Bias and Cultural-Flattening

Generative models are trained on uneven image–text corpora. Because those corpora are largely assembled from web-scraped material, language, geography, platform visibility, digitisation history and aesthetic convention shape the material available for training. Studies of machine-vision datasets show that such corpora reproduce the values and exclusions of their source material, including Western and English-language dominance (Birhane et al., 2023). In heritage reconstruction, this imbalance affects the way models complete missing, damaged or weakly specified details.

The consequence is not only generic bias. Where particular iconographic, material or architectural traditions are poorly represented, generated outputs may fall back on more familiar visual conventions: generic ornament, simplified surface texture, anachronistic material treatment, or weakened local iconographic specificity. Chang et al. document this fallback pattern in the broader content-generation literature, where culturally specific visual features are systematically replaced by more frequent conventions present in the training distribution (Chang et al., 2023). Nyaaba et al. extend the diagnosis to educational and cultural-content applications, showing how regional and minority traditions are most exposed to this form of distortion (Nyaaba et al., 2024). The trustworthy-AI work by S. Tiribelli et al. frames the problem as a cultural-flattening risk specific to generative heritage applications (S. Tiribelli et al., 2024), while A. Tiribelli et al. examine the ethical implications for representation and authority (A. Tiribelli et al., 2024). Such errors are substantive because they

affect the details through which historical plausibility is judged: ornament, gesture, material finish, symbolic meaning and regional style. The risk is greatest where the documentary record is incomplete, since visual fluency can make unsupported completion appear credible.

This directly affects evaluation. A GenAI reconstruction cannot be assessed only through the visual plausibility produced by the model. That plausibility may reflect the training distribution as much as the historical evidence. The generated image must therefore be checked against documentary sources, specialist interpretation and cultural context. The standards established in Section 2.2, including distinguishability, evidential traceability and culturally situated authenticity, help prevent model plausibility from being treated as historical validity. They require the evaluator to ask what is documented, what is inferred, what remains speculative and whose cultural standards are being used to judge authenticity.

2.4.6 Gap 1 - Evidential Control

Table 1 compares the GenAI heritage reconstruction studies reviewed in this section against the dimensions that a heritage-specific evaluation framework should integrate structured prompt design, paradata recording, expert evaluation, computational filtering against evidence, community input and explicit transferability testing.

Table 1. Evaluation functions in selected GenAI and AI-mediated heritage reconstruction literature

Literature focus	Representative studies	Year	What is evaluated	What remains unresolved
Image generation	(Arzomand et al., 2024; Navarro-Mateu et al., 2021; Shih, 2025)	2021 - 2025	Visual plausibility	Evidence trail
Image enhancement	(Nayak & Balabantaray, 2021)	2021	Image quality	Heritage validity
Multimodal generation	(Wei et al., 2026)	2026	Visual-semantic fit	Auditability
Prompting frameworks	(Li et al., 2025)	2025	Prompt logic	Empirical validation
AI perception	(Forte, 2024)	2024	Visual interpretation	Reconstruction relevance
3D / architectural pipelines	(Delgado-Martos et al., 2026)	2026	Technical output	Transferability
Historical benchmarks	(Ghaboura et al., 2025)	2025	Model capability	Reconstruction workflow

Bias diagnostics	(Foka et al., 2025; Foka & Griffin, 2024b, 2024a)	2024-2025	Representation risk	Evaluation procedure
Trustworthy-AI heritage	(Paolanti et al., 2026; Spennemann, 2024)	2024-2026	Ethical principles	Operational method

Across the studies compared in Table 1, evaluation remains tied to the primary output of each method. GAN and reconstruction-pipeline studies assess visual restoration or system performance; benchmarks score model responses against curated historical data; diagnostic and policy work identifies risks, principles and conditions for responsible use. Text-to-image and multimodal studies move closest to the reconstruction problem, but their assessment still depends mainly on visual comparison, limited metrics or restricted expert inspection. The stages that produce the reconstruction claim remain weakly connected: source selection, prompt construction, model configuration, iteration, output filtering and paradata are rarely treated as a single evidential process. A reconstruction candidate may be visually persuasive, technically coherent or highly scored by image-quality metrics, yet still fail to identify the documented, inferred and model-produced elements in the image. In heritage reconstruction, that distinction is fundamental. It determines whether the output can be examined as a reconstruction claim.

This pattern defines the first gap addressed by the thesis. GenAI heritage studies show that damaged or destroyed heritage can be visualised through generative systems, but they do not yet provide a structured procedure for explaining how evidence and model-mediated decisions shape the final reconstruction. The relevant heritage standards remain applicable, but their operational form for prompt-based generation remains underdeveloped.

Chapter 3 develops this gap through the published multi-site text-to-image study, where pre-HARF prompting, visual comparison and limited computational metrics proved insufficient for assessing reconstruction validity (Arzomand et al., 2024). Chapter 4 then addresses it methodologically through HARF, which formalises evidence-anchored prompting, prompt sufficiency assessment, paradata logging and computational pre-evaluation as components of a reproducible evaluation workflow (Arzomand et al., 2026).

2.5 Validation, Metrics and Expert Judgement

The GenAI literature reviewed in Section 2.4 shows that plausible reconstruction candidates can be generated without a corresponding procedure for assessing their evidential status. Digital heritage has a longer validation tradition, developed through photogrammetry, HBIM, uncertainty visualisation, expert review and, more recently, AI benchmarking. Geometric comparison, image-similarity metrics, structured expert assessment, uncertainty communication and benchmark scoring nevertheless remain divided across separate methodological settings and do not yet form a reproducible workflow for assessing AI-generated reconstruction candidates from evidence selection, through computational filtering, to expert judgement. The problem is therefore not the absence of validation methods. It is the absence of an AI-specific evaluation framework that integrates those methods into a single heritage reconstruction workflow.

2.5.1 Existing Validation Frameworks in Digital Heritage

Digital heritage already contains established validation practices, but they were not designed for prompt-based generation. Photogrammetric and HBIM validation, discussed in Section 2.3, provide reproducible procedures for geometric and semantic correspondence, including deviation analysis against survey-grade reference data, classification accuracy for segmented architectural elements and interoperability testing through standardised exchange formats. These procedures are strongest where the object of assessment is measurable geometry or semantic classification. They do not settle the evidential questions raised by reconstruction from incomplete evidence: which historical state is being represented, how missing elements have been inferred and how uncertainty has been documented.

Structured expert assessment provides a different form of validation. Expert evaluation has been combined with uncertainty visualisation in archaeological reconstruction and applied in mixed expert-user evaluation of cultural heritage virtual experiences (Nikolakopoulou & Koutsabasis, 2020; Pietroni, 2021; Pietroni & Ferdani, 2021). These studies show that expert and user assessment can be organised through defined criteria. Their focus, however, is usually visualisation quality, interpretation, communication or experience. They do not provide a workflow for deciding whether an AI-generated reconstruction is evidentially traceable, historically plausible and procedurally reviewable. Heritage monitoring and change-detection work provides another validation context. Agapiou's study of UNESCO World Heritage properties uses optical and radar satellite data to monitor dynamic

environmental and anthropogenic change, but this is a monitoring framework rather than a reconstruction-evaluation protocol (Agapiou, 2021).

AI-specific validation remains narrower. Benchmarking, bias-diagnostic and trustworthy-AI work provide useful components: curated historical reference data, diagnosis of representational distortion and principles of accountable practice (Foka et al., 2025; Foka & Griffin, 2024a, 2024b; Ghaboura et al., 2025; Paolanti et al., 2026). These contributions identify what may need to be assessed, where assessment may fail and what responsible use should require. They do not combine evidential traceability, computational filtering, paradata recording and expert review into a reproducible procedure for evaluating a specific AI-generated heritage reconstruction.

The London Charter and Seville Principles remain evidentiary expectations rather than validation procedures. They require documentation of evidence, inference and uncertainty, but do not specify how those requirements should be implemented for prompt-based GenAI outputs. Existing validation work therefore supplies components of evaluation, not the integrated framework required for AI-mediated reconstruction.

These approaches leave the evaluation task distributed across separate registers. Digital heritage standards specify duties of evidence, inference, transparency and uncertainty. Computational metrics and AI benchmarks test limited forms of visual correspondence, model behaviour or representational distortion. Trustworthy-AI and bias-focused work identify risks around accountability and responsible use. None of these approaches, however, provides a procedure for assessing an individual GenAI reconstruction candidate from source selection through prompt construction, paradata recording, computational filtering and structured expert review. HARF addresses this procedural gap by linking these stages within a single evaluation workflow for two-dimensional GenAI heritage reconstruction.

2.5.2 Computational Metrics and Their Limits

Image-similarity metrics provide a technical screening layer for image-based heritage evaluation, but reconstruction validity depends on evidence, interpretation and expert judgement as well as visual correspondence. The Structural Similarity Index Measure assesses similarity through luminance, contrast and structural correspondence and is more sensitive to perceptual differences than pixel-level error measures (Zhou Wang et al., 2004). Mean Squared Error, Peak Signal-to-Noise Ratio and Mean Absolute Error provide pixel-level fidelity indicators. Feature-based methods extend this set toward photometric and geometric robustness: Rublee et al. introduced the Oriented FAST and Rotated BRIEF feature detector as a computationally efficient alternative to earlier descriptors (Rublee et al., 2011); Fischler and Bolles established the RANSAC consensus-filtering procedure that is

routinely applied to such feature correspondences to suppress outliers (Fischler & Bolles, 1981); and Dalal and Triggs introduced the Histogram of Oriented Gradients descriptor, which captures local edge orientation distributions and remains widely used where feature-level correspondence is required across archival imagery of variable photometric quality (Dalal & Triggs, 2005).

These metrics are useful within their scope. They can identify candidates that depart from known dimensions, structural outlines or visible reference features, especially where archival reference imagery has a documented evidential status. Their limitation is that they do not assess semantic, iconographic or culturally specific accuracy. Two images can score highly on SSIM while differing in iconographic detail, stylistic convention or material treatment. Two images can score lower on SSIM because of stylistic variation that an expert would accept as legitimate. Feature-based methods are more robust to photometric variation than pixel-level metrics, but they share the same boundary: feature correspondence is not iconographic correspondence.

Computational metrics can therefore support filtering, but not final judgement. They can narrow the set of candidates requiring expert assessment by identifying outputs that depart visibly from documented evidence. They cannot establish historical plausibility, nor can they determine whether a reconstruction is evidentially and culturally defensible. A framework that relies on metrics alone risks accepting culturally specific errors as acceptable visual variation. A framework that ignores computational filtering risks admitting candidates that conflict with known reference material. Both layers are needed, but they perform different evaluative functions.

2.5.3 Expert Review as Evaluation Method

Expert review is present in digital heritage evaluation, but it is often reported as a judgement rather than structured as a method. Four questions remain inconsistently addressed: who counts as an expert in each heritage context, what experts are asked to assess, how disagreement is recorded and how expert judgement interacts with computational scores. Where expert review is included, the criteria, instruments and reliability of the assessment are not always specified.

Methods for assessing inter-rater agreement are well established in adjacent fields. Cohen's kappa, Krippendorff's alpha, Gwet's agreement coefficients and intraclass correlation are used in clinical research, content analysis and other rater-based studies to measure agreement under different assumptions (Cohen, 1960; K. Gwet, 2012; Krippendorff, 2004; Shrout & Fleiss, 1979). Their use in digital heritage evaluation remains uneven. Structured protocols, where reported, often use Likert-scale ratings across predefined criteria and summarise results through means and standard deviations.

Inter-rater reliability is reported less consistently and methodological issues such as skewed rating distributions receive limited attention.

This matters for AI-mediated reconstruction because generated outputs can be produced in volume. Computational filtering is only useful if the candidates that pass through it are then assessed through a consistent expert protocol. Without such a protocol, expert disagreement cannot be examined, the relation between computational and expert assessment cannot be diagnosed and the evaluation process cannot be reproduced by another research team.

Recent work has begun to identify a divergence between visual plausibility and evidential adequacy in GenAI heritage outputs (Wei et al., 2026), while review literature notes that evaluation in generative digital heritage often remains perceptual, experiential or visually oriented rather than evidentiary (Xu et al., 2025). Chapter 5 develops this problem through paired computational and expert assessment of 245 Bamiyan reconstruction candidates, including inter-rater reliability for the expert protocol (Arzomand & Kalganova, 2026b). The wider literature recognises the risk of divergence between technical assessment and disciplinary judgement, but has not yet established its structure, magnitude or domain-specificity within a controlled GenAI heritage reconstruction workflow.

2.5.4 Gap 2 - Integrated Evaluation

Table 2 compares the validation approaches reviewed in this section against the dimensions an AI-specific heritage evaluation framework would need to organise pre-generation evidence control, computational pre-evaluation, structured expert review, paradata recording and integration across registers.

Table 2. Validation approaches relevant to GenAI heritage reconstruction

Approach	Domain	Studies	Year	Evaluated	Strength	Unresolved for GenAI reconstruction
London Charter / Seville Principles	Digital heritage	(Denard, 2009; López-Mencheró Bendicho & Grande, 2011, 2017)	2009, 2017	Evidence status	Accountability principles	Operational workflow
Photogrammetry / HBIM	Geometric reconstruction	(Pocobelli et al., 2018; Radanovic et al., 2020; Remondino,	2020 – 2021	Spatial accuracy	Metric validation	Prompt process

		2011; X. Yang et al., 2020)				
Digital reconstruction uncertainty	Visualisation practice	(Pietroni, 2021; Pietroni & Ferdani, 2021)	2021	Uncertainty handling	Expert judgement	AI generation
Heritage virtual experience	User experience	(Champion et al., 2021; Nikolakopoulou & Koutsabasis, 2020)	2025	Perception / immersion	Structured feedback	Heritage authority
Historical LMM benchmarks	AI evaluation	(Ghaboura et al., 2025)	2025	Model capability	Ground-truth scoring	Reconstruction workflow
GenAI heritage bias	AI heritage critique	(Foka et al., 2025; Foka & Griffin, 2024a)	2024 – 2025	Representation risk	Bias diagnosis	Evaluation procedure
Trustworthy-AI heritage	Cultural AI	(Paolanti et al., 2026; Spennemann, 2024; S. Tiribelli et al., 2024)	2024 – 2026	Ethical principles	Governance framing	Empirical method

Validation is strongest where the object of assessment is narrow. Photogrammetric and HBIM procedures can test geometric or semantic correspondence. Image metrics can filter candidates against visible reference features. Expert-assessment and uncertainty frameworks can structure disciplinary judgement, while AI-specific benchmarks can score model behaviour against curated historical data. Trustworthy-AI and bias-diagnostic work identify principles and risks for cultural-heritage AI, but they do not operationalise those principles as a reconstruction-evaluation workflow.

The limitation is therefore not the absence of validation methods. It is their separation by method and object of assessment. Existing approaches provide useful components, but not a reproducible procedure that links pre-generation evidence control, prompt sufficiency, computational filtering, paradata recording and structured expert judgement in the assessment of a generated reconstruction candidate.

This defines the second literature-derived gap: the absence of an integrated AI-specific evaluation framework for GenAI heritage reconstruction. Chapter 4 addresses this gap through HARF, which formalises evidence-anchored prompting, prompt sufficiency assessment, paradata logging and computational pre-evaluation. Chapter 5 then tests the framework through computational and

specialist evaluation of the Bamiyan reconstruction candidates. A further register, distinct from but connected to computational and expert validation, concerns cultural authenticity. Where destroyed heritage is also a matter of custodial memory, community authority and contested meaning, reconstruction validity cannot be settled by similarity metrics or expert archaeological judgement alone. That register is examined in Section 2.6.

2.6 Authenticity, Community Authority and Cultural Accountability

Computational comparison and expert archaeological judgement can assess whether a reconstruction candidate corresponds to evidence and disciplinary expectations. They cannot determine, on their own, whether the reconstruction is culturally legitimate for a destroyed or contested site. For such sites, reconstruction is also a claim about memory, loss, absence and belonging. Its validity depends not only on what can be measured or archaeologically defended, but also on the cultural standpoint from which authenticity is judged.

This section reviews the literature that makes cultural legitimacy part of the evaluation problem. It moves from post-Nara authenticity theory to community authority and participatory heritage, then to the post-2001 Bamiyan literature and the representational ethics of AI-mediated reconstruction. The section develops the third literature-derived gap: existing GenAI heritage evaluation does not yet integrate computational assessment, expert judgement and community-grounded authenticity within a single evaluation framework.

2.6.1 Culturally Situated Authenticity

Section 2.2 established that authenticity in heritage reconstruction cannot be reduced to material correspondence. In destroyed or contested heritage, a reconstruction accepted by conservation professionals or archaeological experts may still be judged differently by state authorities, religious communities, displaced groups or local custodians. The disagreement concerns more than visual accuracy. It can concern which historical state is being privileged, whether absence should remain visible, whether completion is desirable and whose memory of the site is being authorised. Evaluation must therefore specify whose account of authenticity is being applied.

Post-Nara heritage scholarship treats authenticity as a judgement formed through conservation practice, institutional recognition, public memory and social use. Authenticity remains connected to evidence, but evidence enters reconstruction through selection, interpretation and authorisation (Harrison, 2012; Jones, 2010; Smith, 2006). Reconstruction does not merely reproduce a known past.

It gives priority to particular sources, frames the relation between damage and repair and assigns authority to some forms of knowledge over others. These choices are not external to the reconstruction claim; they constitute part of it.

Critical heritage studies explain how such choices become normalised within heritage practice. Authorised heritage discourse often gives greater weight to expert, institutional and state-recognised values, especially where heritage is understood through monuments, material fabric and conservation procedure (Smith, 2006; Waterton & Watson, 2013). Community memory lived attachment and local or diasporic meanings can be recognised without determining the final account. In difficult and contested heritage, this hierarchy carries greater force because the site is not only a damaged object but also a focus of loss, identity, displacement and unresolved meaning (Logan & Reeves, 2008; Macdonald, 2008).

AI-mediated reconstruction can obscure this imbalance. A generated image can follow parts of the archival record and satisfy archaeological expectations while presenting the site in a form that connected communities do not recognise as acceptable, respectful or authentic. The failure is not always visible as a historical error. It can lie in the treatment of absence, in the assumption that visual completion is desirable, or in the closure of a contested history into a single image. Similarity metrics cannot identify that failure. Expert judgement remains necessary, but it cannot by itself establish cultural legitimacy where authenticity depends on memory, custodianship and acceptable reconstruction.

Community-grounded authenticity is therefore part of reconstruction validity, not a separate ethical qualification after computational and expert validation. An AI-generated heritage reconstruction has to be assessed not only for evidential traceability and archaeological plausibility, but also for whether its account of authenticity recognises the communities for whom the site remains culturally meaningful.

2.6.2 Community Authority and Participatory Heritage

Participatory heritage scholarship shows how community authority can be incorporated into heritage practice. Community-based archaeology and participatory heritage work argue that descendant, local and custodial communities should not be positioned only as audiences for heritage interpretation. They can hold knowledge that affects how a site is valued, how damage is understood and what forms of reconstruction are considered legitimate (Atalay, 2012; Silberman, 2012). Where communities have been excluded from institutional heritage management, the assessment of authenticity, value and

representation cannot rest only with the institutions that have historically controlled those processes (Risam, 2018; Stobiecka, 2020).

The relevance for reconstruction evaluation lies in the evidence that participation makes available. Archival records and expert interpretation can address material form, stylistic plausibility and historical consistency. Community-held knowledge can address custodial attachment, inherited memory, local naming, symbolic meaning and the cultural acceptability of reconstruction. These are not substitutes for archaeological evidence. They address a different part of the evaluation problem. A reconstruction process that excludes them narrows the assessment before the generated image is judged.

Digital heritage has begun to translate participation into practice through co-creation, co-design and community involvement in interpretation, design and evaluation (Champion et al., 2021; Nikolakopoulou & Koutsabasis, 2025). This work shows that participation can be organised through defined methods. Its usual focus, however, remains visitor experience, interpretive design, digital engagement or interactive heritage communication. It does not yet provide a procedure for assessing AI-generated reconstructions of destroyed heritage against community-grounded authenticity.

For GenAI reconstruction, community input cannot be treated as a response gathered after a finished image has already been selected. It must shape the criteria by which reconstruction candidates are judged. Participatory heritage scholarship therefore supports the need for community authority in GenAI evaluation, but it has not yet translated that authority into a structured assessment procedure for generated reconstruction candidates.

2.6.3 Bamiyan and the Structure of Voice in the Post-2001 Literature

The destruction of the Bamiyan Buddhas in March 2001 generated sustained international scholarship on cultural loss, reconstruction, intervention and the future of the niches and surrounding landscape (Flood, 2002; Lostal & Vasconcelos Vilaça, 2015). The UNESCO symposium volume remains a major recent consolidation of this debate (Nagaoka, 2020). Its contributors examine anastylosis, partial reconstruction, non-reconstruction and stabilisation and address what authenticity, memory and intervention mean where the original material has been almost entirely lost (Holtorf, 2020; Jokilehto, 2020; Rössler, 2020).

The limitation lies less in the care of that scholarship than in its structure of voice. Afghan perspectives appear in the post-2001 literature, but often through institutional processes, international experts, secondary citation or externally organised conservation debates. Hazara perspectives require particular attention because Bamiyan is not only an archaeological or Buddhist art-historical site. It is

also embedded in local history, identity and memory. Existing scholarship therefore does not fully answer how Afghan and Hazara respondents themselves understand authenticity, loss, absence and acceptable reconstruction at Bamiyan.

Contextual scholarship supports this point. Studies of Hazara history and identity establish the political and cultural distinctness of the Hazara as a community connected to Bamiyan (Monsutti, 2005; Mousavi & Canfield, 1999). Naming traditions around the figures, including Salsal and Shahmama, show that the statues have also entered local narrative and memory, not only Buddhist, archaeological or international conservation discourse (Singh, 2022). For AI-mediated reconstruction, this matters because a visually plausible and archaeologically constrained image may still fail to engage the cultural meanings through which Bamiyan is understood by communities connected to the site.

The post-2001 Bamiyan literature therefore frames the cultural-authenticity problem but does not resolve it empirically. It identifies reconstruction as a question of memory, loss and acceptable intervention, yet gives limited direct evidence of how Afghan and Hazara participants judge AI-mediated reconstruction. Chapter 6 addresses that gap through primary respondent data from 62 Afghan participants, examining cultural connection, authenticity, attitudes to reconstruction, the representation of absence and the risk that AI-generated images oversimplify cultural meaning (Arzomand & Kalganova, 2026a).

2.6.4 AI Ethics and Representational Bias

The AI ethics literature adds a representational dimension to the authenticity problem developed above. Bias in GenAI heritage reconstruction is not only a matter of inaccurate visual detail. It can affect which heritage forms are easily generated, which forms of loss are visually smoothed and which cultural meanings are simplified through the model's learned visual conventions. The bias-loop argument discussed in Section 2.4.4 explains how AI-generated cultural content can reinforce future distortions when it re-enters training data. The training-data problem discussed in Section 2.4.5 explains why culturally specific and under-represented heritage is vulnerable to generic visual completion. The ethical issue is therefore inseparable from evaluation: the reconstruction process must be able to identify and challenge these distortions before they are presented as heritage images.

Digital-colonial and critical digital-heritage scholarship gives this problem a wider frame. Uneven digitisation, platform visibility, language dominance and institutional control affect which cultural materials become computationally available and which communities are able to contest the outputs produced from them (Risam, 2018; Stobiecka, 2020). In reconstruction, this matters because the generated image can appear neutral while drawing authority from unequal archives and uneven data

infrastructures. A reconstruction may be technically coherent while reproducing exclusions already present in the sources, platforms or datasets from which the model has learned.

Recent AI-ethics work in cultural heritage identifies related concerns around accountability, explainability, representational fairness, cultural harm and the responsibilities of heritage professionals using generative systems (D. Ocón et al., 2026; S. Tiribelli et al., 2024). These contributions are valuable because they define the ethical conditions under which GenAI heritage work should proceed. Their limitation for the present evaluation problem is operational. They do not specify how an individual reconstruction project should test for representational responsibility, how community authority should enter that assessment, or how cultural accountability should be combined with computational and expert evaluation.

The ethical literature therefore strengthens, but does not close, the gap identified in this section. It shows why GenAI reconstruction must be assessed for more than technical correctness or historical plausibility. It does not yet provide a workflow for judging whether a generated reconstruction is culturally accountable to the communities for whom the heritage remains meaningful.

2.6.5 Gap 3 – Community-Grounded Authenticity

The literature reviewed in Sections 2.5 and 2.6 shows that GenAI heritage reconstruction requires more than one form of assessment. Computational comparison can identify departures from reference material where suitable evidence exists. Expert archaeological judgement can assess historical plausibility, stylistic coherence and the handling of inference. Community-grounded authenticity addresses a different question: whether the reconstruction is acceptable, recognisable and culturally legitimate for those connected to the heritage.

These forms of assessment are not interchangeable. A reconstruction can score well against image metrics and still fail expert review. It can satisfy expert expectations and still conflict with community understandings of loss, absence or acceptable reconstruction. It can be culturally meaningful while remaining weakly documented or visually inconsistent with archival evidence. Such divergence is not a methodological inconvenience. It is part of what evaluation must be able to diagnose.

The third literature-derived gap lies in this lack of integration. Digital heritage standards establish evidentiary principles. Validation literature provides computational and expert-assessment tools. Participatory and authenticity literature establishes the need for community authority. AI-ethics literature identifies risks of bias, cultural flattening and representational harm. These literatures are developed in their own terms, but they do not yet specify how computational filtering, structured

expert review and community-grounded authenticity should be combined within one evaluation workflow for GenAI heritage reconstruction.

Chapter 5 addresses the computational and expert components of this problem through paired computational and structured expert evaluation of 245 Bamiyan reconstruction candidates. Chapter 6 addresses the community-grounded component through primary respondent data from 62 Afghan participants, examining authenticity, cultural connection, attitudes to reconstruction, the representation of absence and the risk of AI-mediated simplification. Together, those chapters test the central claim developed in this section: reconstruction validity cannot be established through isolated assessments; it requires a framework capable of identifying alignments and tensions between computational similarity, expert archaeological judgement and community-grounded authenticity.

2.7 Transferability and Bounded Generalisation

Sections 2.4 to 2.6 established the components required for evaluating GenAI heritage reconstruction: evidential control, computational filtering, structured expert review and community-grounded authenticity. The remaining question concerns the scope of such a framework. A procedure developed through one heritage case cannot be treated as generalisable only because it works within that case. Figural sculpture, monumental architecture and archaeological landscapes present different source records, visual grammars, expert criteria and failure domains.

This is particularly important for GenAI reconstruction because generation can create an appearance of generality. A model may produce visually plausible candidates across several sites, but multi-site generation does not establish transferability of evaluation. The relevant test is whether the evaluative structure can be re-instantiated when the heritage object changes whether evidence can be reorganised, prompts constrained, outputs filtered, expert criteria adjusted and failure domains diagnosed without redesigning the framework itself.

This section reviews how transferability and generalisation are treated in digital heritage, computational archaeology and GenAI heritage reconstruction. It identifies a gap in the empirical testing of cross-typology transfer and develops the fourth literature-derived gap: existing GenAI heritage research does not yet show how an evaluation framework developed through one heritage typology can be tested, in bounded form, against another.

2.7.1 Transferability In Digital Heritage and Computational Archaeology

Transferability has been addressed unevenly in digital heritage. Photogrammetric and laser-scanning workflows transfer comparatively well across heritage contexts because their primary object of assessment is geometric correspondence. Acquisition conditions vary by scale, access, surface condition and image coverage, but the validation logic remains broadly stable: reconstructed geometry is assessed against reference data, control points or survey accuracy (Maiwald et al., 2017; Remondino, 2011). The method changes with the site, but the evidential claim remains centred on measurable form.

HBIM shows the boundary of this technical transfer. Geometric validation can be carried across projects through point-cloud comparison and deviation analysis, but semantic and interpretive structures are less portable. Object libraries, classification schemes, levels of detail, uncertainty representation and documentation practices depend on architectural typology and institutional context. Recent reviews identify continuing difficulties in semantic enrichment, automation across heterogeneous architectural forms, interoperability and uncertainty representation (Cui & Wu, 2025; X. Yang et al., 2020). The technical layer transfers more readily than the interpretive layer.

Computational archaeology makes the problem of generalisation more explicit. Computational models are built within particular datasets, spatial scales, chronological ranges and recording conditions. Their transfer to a new context cannot be inferred from performance in the development context alone. Re-application exposes assumptions in the original data and modelling procedure and failure can be methodologically informative rather than merely negative (Bevan & Lake, 2016). This discipline has not been applied consistently to evaluation frameworks for digital heritage reconstruction.

Most reconstruction studies remain site-specific or typology-specific. Their evaluation procedures are usually tested within the same evidential setting in which they were developed. Multi-site application does not change this unless transferability is itself examined. A multi-site text-to-image study reported plausible reconstruction candidates across several heritage sites, including Bamiyan and Palmyra. It established multi-site generation capacity but did not test transferability at the level of evaluation-framework design (Arzomand et al., 2024). The ARCHER project applies an integrated AI pipeline across architectural cases, including Greek temples and Romanesque-Mudéjar churches, but its reported evaluation concerns pipeline functioning and consistency, not the transferability of an evaluation framework as an evidential procedure (Delgado-Martos et al., 2026).

Recent GenAI benchmarking and diagnostic work addresses a different level of generalisation. Benchmarks such as the Time Travel Benchmark compare model behaviour across curated historical

content, while bias-diagnostic work examines field-level patterns of distortion (Foka & Griffin, 2024a, 2024b; Ghaboura et al., 2025). These approaches are useful for cross-model assessment and risk diagnosis. They do not test whether an evaluation framework developed for one heritage typology can be applied to another typology with different evidence categories, visual failure modes and expert criteria. Cross-model testing is therefore not a substitute for cross-typology validation.

2.7.2 Stable And Variable Elements of Evaluation

Transferability in GenAI heritage reconstruction depends on what is expected to transfer. Heritage typology changes the evidential record, the visual features requiring assessment and the criteria through which experts judge plausibility. A framework may retain the same evaluative structure across cases, but the specific evidence categories and likely failure domains will not remain the same.

The Western Buddha of Bamiyan and the Temple of Bel at Palmyra illustrate this distinction. For a destroyed figural sculpture such as the Western Buddha, the unstable features concern drapery articulation, cranial morphology, hand gesture, surface treatment and the iconographic conventions associated with Gandhāran Buddhist sculpture. These are features for which archival evidence is uneven and expert interpretation remains necessary. For a monumental architectural structure such as the Temple of Bel, the evaluation problem shifts. The relevant instability points concern architectural proportion, column and capital morphology, relief carving, surface treatment, inscriptional detail, roofline elements and architectural features for which the photographic or comparative record is incomplete. Schmidt-Colinet and Seigne's reconsideration of the temple's crenellations shows that even familiar elements of the canonical reconstruction can remain interpretively contested (Schmidt-Colinet & Seigne, 2022).

The evidential categories that an evaluation framework must engage therefore vary by typology, even where the high-level evaluation logic remains stable. A framework may transfer at the level of evidence anchoring, prompt sufficiency, paradata recording, computational filtering, structured expert review and failure diagnosis. It cannot be expected to transfer at the level of specific failure modes. A reconstruction candidate from a different typology may fail in different domains while still being assessed through the same procedural structure. A cross-site test must therefore distinguish what remains stable from what has to be re-specified for the new heritage object.

2.7.3 The Temple of Bel as a Transferability Case

The Temple of Bel provides a suitable context for testing bounded transferability because it differs from Bamiyan in typology, evidential structure and likely failure domains. Earlier reconstruction work has produced digital models and immersive representations of the temple, while later scholarship has re-examined specific architectural claims, including the treatment of crenellations (Rihani, 2023; Schmidt-Colinet & Seigne, 2022; Wahbeh et al., 2016). These studies are important reconstruction and interpretation work. They do not test whether an evaluation framework developed through a figural heritage case can be re-instantiated for a monumental architectural case.

The Temple of Bel has a sufficient archival and scholarly record to support evidence-anchored prompting and expert assessment. It is also typologically distinct from the Western Buddha of Bamiyan. Its primary representational risks concern architectural proportion, decorative relief, column and capital morphology, surface treatment, inscriptional features and contested or incompletely documented architectural elements. The site therefore allows the framework to be tested under different evidential and visual conditions without making an unrestricted claim of general applicability.

Chapter 7 conducts this test through a bounded re-instantiation of HARF at the Temple of Bel. It uses structured expert assessment from 32 respondents and identifies typology-dependent instability points specific to architectural reconstruction (Arzomand & Kalganova, 2026c). The transferability claim is deliberately limited. It is not that HARF is universally valid across all heritage contexts. It is that the framework's evidence-anchoring logic, prompt-sufficiency structure, paradata procedure and failure-diagnosis process can be re-instantiated from a figural typology to an architectural typology, while the specific failure domains shift with the decorative, material, inscriptional and structural characteristics of the new case.

2.7.4 Gap 4 - Bounded Generalisation

Transferability in GenAI heritage evaluation cannot be inferred from successful application within a single case, or from the ability of a model to generate outputs across several sites. Multi-site generation, pipeline application and cross-model benchmarking demonstrate technical reach across contexts. They do not establish whether an evaluation procedure remains valid when applied to a different heritage typology with different evidential records, expert criteria and failure domains.

The fourth literature-derived gap is therefore the absence of empirically tested cross-typology transferability in GenAI heritage evaluation frameworks. Existing work does not yet show how a framework developed through one heritage typology can be re-instantiated for another while

distinguishing between stable and variable elements of evaluation. Evidence anchoring, prompt sufficiency, paradata recording, computational filtering, structured expert review and failure diagnosis are procedural elements. Source materials, prompt descriptors, reference features, expert criteria and instability domains have to be re-specified for each typology. Without this distinction, generalisation remains a claim about application, not a tested property of the evaluation procedure.

Chapter 7 addresses this gap through the bounded re-instantiation of HARF at the Temple of Bel, Palmyra. The chapter does not use Palmyra to prove that AI-generated reconstructions are historically defensible across heritage sites in general. It tests whether a framework developed through the Western Buddha of Bamiyan can be re-instantiated for a monumental architectural case while preserving its evaluative structure.

At Bamiyan, instability is concentrated around figural and iconographic features. At Palmyra, it shifts towards architectural proportion, decorative relief, surface treatment, inscriptional detail and roofline form. These differences are central to the transferability claim. Generalisation, in this thesis, does not mean that the same reconstruction problems recur across typologies. It means that the evaluation framework remains capable of identifying the evidential demands and failure domains produced by a different heritage object.

2.8 Synthesis of Literature-Derived Gaps

The literature reviewed in this chapter demonstrates the feasibility of producing visually coherent heritage reconstruction candidates through generative systems. It also shows that visual coherence does not make those outputs accountable as heritage reconstructions. Accountability requires an evidential pathway from source material to generated image, a validation procedure that combines computational filtering with expert judgement, a means of addressing cultural legitimacy where connected communities hold custodial claims and a test of whether the procedure remains valid beyond the case in which it is developed.

These requirements define four literature-derived gaps. The first concerns evidential control over GenAI heritage outputs. The second concerns the absence of a standardised AI-specific evaluation procedure. The third concerns the separation of computational, expert and community-grounded assessment. The fourth concerns the lack of empirical testing of evaluation-framework transferability across heritage typologies.

2.8.1 The Four Gaps as a Cumulative Argument

Heritage reconstruction standards provide the evidentiary baseline for the chapter. The Venice Charter, Nara Document, London Charter and Seville Principles establish requirements of distinguishability, transparency, paradata, uncertainty and culturally situated authenticity. These standards define the conditions under which reconstruction should be accountable, but they do not specify how prompt construction, model selection, iteration, output filtering or AI-generated uncertainty should be handled in GenAI reconstruction.

Pre-GenAI reconstruction practice shows that reconstruction can be evidence-constrained and interpretively disciplined. Photogrammetry, archival reconstruction, HBIM and expert review can organise evidence within particular sites and support measurable forms of validation. Their strength is also their boundary. These workflows remain tied to the evidential conditions of specific cases and geometric validation cannot determine whether a reconstruction is historically plausible, culturally legitimate or transferable as an evaluation procedure.

GenAI heritage studies expose the first gap. They demonstrate that plausible reconstruction candidates can be generated, but they rarely provide a structured account of how source material, prompt decisions, model behaviour, iteration and output selection shape the final image. Evidential control is weakest at the point where the generated output becomes a reconstruction claim.

The validation and expert-evaluation literature demonstrate the second gap. Computational metrics, structured expert protocols and AI-specific benchmarks each address part of the evaluation problem. They do not yet form a single AI-specific workflow linking evidence control, prompt sufficiency, computational filtering, paradata and structured expert review. The difficulty is not the absence of evaluation methods, but their separation across different methodological settings.

The authenticity, community-authority and AI-ethics literature exposes the third gap. Reconstruction validity cannot be settled by computational similarity or expert archaeological judgement where destroyed heritage is also bound to memory, absence, custodianship and acceptable representation. Community-grounded authenticity is recognised as necessary, but it has not been operationalised within the same evaluation structure as computational and expert assessment.

The transferability literature exposes the fourth gap. A framework developed through one heritage typology cannot be assumed to generalise because it works in that typology. Transferability has to be tested across cases with different evidence bases, expert criteria and failure domains. For GenAI heritage reconstruction, generalisation is a property of the evaluation procedure to be examined, not an assumption attached to the framework after development.

Together, these gaps establish the requirement for HARF: auditable evidential control, integration of computational and expert assessment, community-grounded authenticity where cultural legitimacy is at stake and bounded transferability across heritage typologies. The framework is developed through Bamiyan and tested for bounded transfer through Palmyra.

2.8.2 Mapping The Literature

Table 3 shows that the thesis gap is cumulative rather than isolated. Heritage standards define the accountability requirements, but not their operational form for prompt-based generation. Pre-GenAI reconstruction workflows provide evidential and geometric discipline but remain largely case-bound. GenAI heritage studies demonstrate visual feasibility, but not a structured evidential pathway from source material to output. Validation, expert review, authenticity and transferability are each developed in separate literatures but are not integrated into a single GenAI reconstruction-evaluation procedure.

Table 3. Literature-derived gaps and thesis response

Literature area	Indicative sources	What the literature establishes	What remains unresolved	Thesis response
Heritage reconstruction standards	(Apollonio et al., 2018; Bentkowska-Kafel & Denard, 2012; Denard, 2009; Huvila, 2022a; ICOMOS, 1964, 1994; López-Menchero Bendicho & Grande, 2011, 2017; Pietroni, 2021)	Evidence, paradata, uncertainty and authenticity are core requirements.	GenAI operationalisation	Gaps 1–3
Pre-GenAI reconstruction workflows	(Grün et al., 2004; Murphy et al., 2013; Pocobelli et al., 2018; Radanovic et al., 2020; Remondino, 2011; Toubekis et al., 2017; Wahbeh et al., 2016; X. Yang et al., 2020)	Evidence-constrained reconstruction and geometric validation are possible.	Case-bound evaluation	Gaps 1–2
GenAI heritage reconstruction	(Arzomand et al., 2024; Delgado-Martos et al., 2026; Forte, 2024; Li et al., 2025; Navarro-Mateu et al., 2021; Nayak &	GenAI can produce visually plausible reconstruction candidates.	Weak evidential pathway	Gap 1, Chs. 3–4

	Balabantaray, 2021; Shih, 2025; Wei et al., 2026; Xu et al., 2025)			
Validation and expert evaluation	(Foka & Griffin, 2024a; Ghaboura et al., 2025; Nikolakopoulou & Koutsabasis, 2020; Paolanti et al., 2026; Pietroni, 2021; Pietroni & Ferdani, 2021; Zhou Wang et al., 2004)	Metrics, benchmarks and expert judgement provide partial evaluation tools.	No integrated procedure	Gap 2, Chs. 4–5
Authenticity and community authority	(Flood, 2002; Foka & Griffin, 2024b; Holtorf, 2020; Jokilehto, 2020; Lostal & Vasconcelos Vilaça, 2015; Monsutti, 2005; Nagaoka, 2020; D. Ocón et al., 2026; Rössler, 2020; Silberman, 2012; Singh, 2022; Smith, 2006; Stobiecka, 2020; S. Tiribelli et al., 2024)	Authenticity is culturally situated and representation can cause harm.	Community register absent	Gap 3, Chs. 5–6
Transferability and bounded generalisation	(Arzomand et al., 2024; Bevan & Lake, 2016; Cui & Wu, 2025; Delgado-Martos et al., 2026; X. Yang et al., 2020)	Transferability matters in computational archaeology and digital heritage.	Cross-typology testing absent	Gap 4, Ch. 7

Across these strands, the unresolved problem is not the absence of relevant principles, methods or case studies. It is their lack of integration into a procedure capable of evaluating AI-generated heritage images as reconstruction claims.

2.8.3 Thesis Position

This thesis contributes to the evaluation of GenAI heritage reconstruction. Existing GenAI heritage work is primarily oriented towards producing, restoring or visualising heritage content. Existing digital heritage validation work assesses geometric, experiential or communicative outcomes through

methods largely developed outside prompt-based AI reconstruction. Evaluation, in the form required for AI-mediated reconstruction of culturally specific heritage, has been recognised as a problem but has not yet been delivered as a procedure. This research addresses that problem through the Heritage-Aligned Reconstruction Framework. HARF provides a standardised, evidence-based, expert-informed and culturally accountable procedure for assessing whether GenAI heritage reconstruction candidates are evidentially traceable, historically plausible, expert-assessed, community-grounded and capable of bounded transfer across heritage typologies.

Chapter 3 reports the multi-site text-to-image study showing the limits of unstructured prompting, visual comparison and limited computational metrics (Arzomand et al., 2024). Chapter 4 develops HARF as the operational response, formalising evidence-anchored prompting, prompt sufficiency assessment, paradata logging, computational pre-evaluation and structured expert review (Arzomand et al., 2026). Chapter 5 tests the computational and expert components through paired assessment of 245 Bamiyan reconstruction candidates, including inter-rater reliability for the expert protocol (Arzomand & Kalganova, 2026b). Chapter 6 develops the community-grounded component through primary respondent data from 62 Afghan participants, addressing cultural connection, authenticity, attitudes to reconstruction, the representation of absence and the risk of AI-mediated simplification (Arzomand & Kalganova, 2026a). Chapter 7 tests bounded transferability through re-instantiation of HARF at the Temple of Bel, Palmyra, with 32 expert respondents and typology-dependent failure domains (Arzomand & Kalganova, 2026c).

The thesis does not argue that GenAI can recover destroyed heritage automatically, that visual plausibility is equivalent to historical validity, or that a framework developed through Bamiyan transfers universally to all heritage contexts. It argues that GenAI heritage reconstruction can be evaluated through a framework that makes evidence, prompt construction, model-mediated decisions, computational filtering, expert judgement, community-grounded authenticity and bounded transferability examinable. Chapter 3 therefore begins with the multi-site text-to-image study as the empirical precursor to HARF, before Chapter 4 develops the formal evaluation framework required by its limitations.

Chapter 3

Initial Evidence-Constrained GenAI Reconstruction Workflow

This chapter presents the first empirical stage of the thesis. It tests whether textual, photographic, archaeological and architectural evidence can be translated into visual reconstruction candidates through text-to-image generation. Its role is deliberately limited. The chapter does not claim that generated images constitute definitive heritage reconstructions. It examines what an evidence-informed prompt workflow can achieve before the more formal evaluation procedure introduced in Chapter 4.

The chapter is based on the published article *From Ruins to Reconstruction: Harnessing Text-to-Image AI for Restoring Historical Architectures*. In the thesis sequence, its function is to establish the practical starting point for HARF. It shows that GenAI can make damaged or destroyed heritage easier to visualise from dispersed evidence, including textual and archival sources. It also shows that visual plausibility, prompt construction and informal comparison are not sufficient grounds for treating such outputs as accountable heritage reconstruction.

3.1 Introduction

The architectural design process is crucial across various disciplines, from engineering and software development to the arts and humanities and it is significantly shaped by the fusion of traditional methods and contemporary technological advancements. Recent innovations in AI, particularly text-to-image systems such as Midjourney, DALL-E and Stable Diffusion, mark a substantial evolution in the design and heritage conservation arenas, expanding the use of natural-language prompting in design and heritage visualisation (Brisco et al., 2023; Kenig et al., 2023). This evolution reflects a broader movement to integrate cutting-edge AI tools within the architectural design process, underscoring the transition from conventional methods to more advanced, future-oriented approaches. The advent of 3D and 2D modelling tools, including Building Information Modelling (BIM) software such as ArchiCAD, AutoCAD and SketchUp, has expanded architects' ability to visualise conceptual designs, while the application of AI techniques, ranging from conditional Generative Adversarial Networks (cGANs) to deep learning and machine learning algorithms, has initiated a new era in the digital preservation and

reconstruction of historical landscapes and artefacts (Navarro-Mateu et al., 2021). These technological advancements are reshaping methodologies in archaeological research and conservation efforts, particularly as the architectural and cultural heritage sectors confront profound losses inflicted by conflicts and wars. Considering such destruction, AI has become increasingly relevant, not only in attempting to preserve and reconstruct cultural heritage, but also to facilitate a deeper understanding of historical contexts (Kadhim & Abed, 2023; Lee & Chang, 2022).

The destruction of cultural heritage sites due to conflicts, natural disasters and human activities poses a substantial challenge to preservation. Traditional reconstruction methods often rely on incomplete or damaged visual records, leading to reconstructions that may lack accuracy and authenticity. Recent advancements in text-to-image technology have introduced tools such as Midjourney and DALL-E, which have expanded the generation of high-quality photorealistic images from simple text prompts (Adetayo, 2024; Becker & Laycock, 2023; Kenig et al., 2023). Models such as Stable Diffusion, DALL-E and Midjourney are notable in the realm of text-to-image generation (Cobb, 2023; Huang et al., 2023; Lyu et al., 2022). The impact of these models is significant, influencing various fields such as neuroscience research, education, libraries and architectural visual culture (Adetayo, 2024; Becker & Laycock, 2023; Steinfeld, 2023). These tools have made it possible for users to create fully rendered images promptly, showcasing the potential of AI in creative endeavours (Newton & Dhole, 2023).

AI-generated imagery offers a way of producing visual reconstruction candidates from textual descriptions, historical records and available visual references (Nichol et al., 2021; Ramesh et al., 2022a). In the context of damaged or destroyed heritage, this capacity is useful because evidence is often dispersed across photographs, archaeological reports, architectural descriptions and conservation records. The value of text-to-image generation at this stage is exploratory: it can test how far evidence-informed prompts produce visually plausible candidates, but it cannot by itself establish historical accuracy, structural fidelity or heritage validity. Dense multi-image reconstruction, photogrammetry, laser scanning, UAV-based documentation and related three-dimensional methods remain important sources of spatial and documentary evidence in digital heritage (Rihani, 2023; Wahbeh et al., 2016). Generative adversarial networks have also been used in heritage image enhancement and super-resolution, where their value lies in improving visual quality rather than resolving reconstruction validity (Nayak & Balabantaray, 2021). These methods provide the wider methodological context; the empirical workflow in this chapter remains limited to two-dimensional text-to-image generation and exploratory comparison.

This study contributes an initial evidence-informed workflow for translating historical, architectural and archaeological descriptions into two-dimensional GenAI reconstruction candidates. Textual prompts were developed from documented site features and available references, then iteratively

refined to examine how generated images responded to historically specified constraints. The value of the workflow lies in making the generation process examinable: sources were converted into prompts, prompts were adjusted in response to outputs and candidate images were compared with available visual and documentary evidence. This establishes the practical potential of text-to-image generation for heritage visualisation while also showing that generated candidates require subsequent evaluation through documented prompting, computational screening and expert review. The workflow was applied to the Buddha of Bamiyan as one of the chapter's principal heritage examples, using available historical, architectural and visual references to guide two-dimensional text-to-image generation (Grün et al., 2004; Hammer et al., 2018). The generated images are treated here as reconstruction candidates derived from available source material. They are used to examine the early text-to-image workflow before the structured evaluation procedure developed in Chapter 4 (Brisco et al., 2023; Kenig et al., 2023; Navarro-Mateu et al., 2021)(Kadhim & Abed, 2023; Lee & Chang, 2022), (Bharati, 2023; Orengo et al., 2015).

3.2 Methodology

This chapter used a five-phase workflow to test whether documented heritage features could be translated into two-dimensional GenAI reconstruction candidates. The workflow comprised data collection, prompt construction, image generation, candidate selection and comparative analysis. Text-to-image platforms, including Midjourney and DALL-E, were used to generate visual candidates from prompts informed by historical, architectural and archaeological evidence. The workflow did not implement three-dimensional reconstruction, physical restoration or formal expert validation. Natural-language prompting is relevant here because it allows design and visualisation instructions to be expressed directly through text, a capability already discussed in architectural and design contexts (Hsu et al., 2022; Ko et al., 2023). Figure 2 summarises the five workflow phases.



Figure 2. Five-phase workflow used in this chapter; diagram produced for the thesis.

Figure 3 presents a detailed flowchart of the five-phase workflow for AI-driven heritage site reconstruction. The process begins with collection and organisation of textual descriptions, architectural details and historical records and available visual references. These materials informed the construction of prompts for Midjourney and DALL-E. The generated outputs were then reviewed through iterative visual comparison with available references and prompts were adjusted where the outputs departed from documented site features. The comparative phase provided preliminary screening. The final phase considered the implications of this exploratory workflow for heritage visualisation and for the structured evaluation procedure developed in Chapter 4.

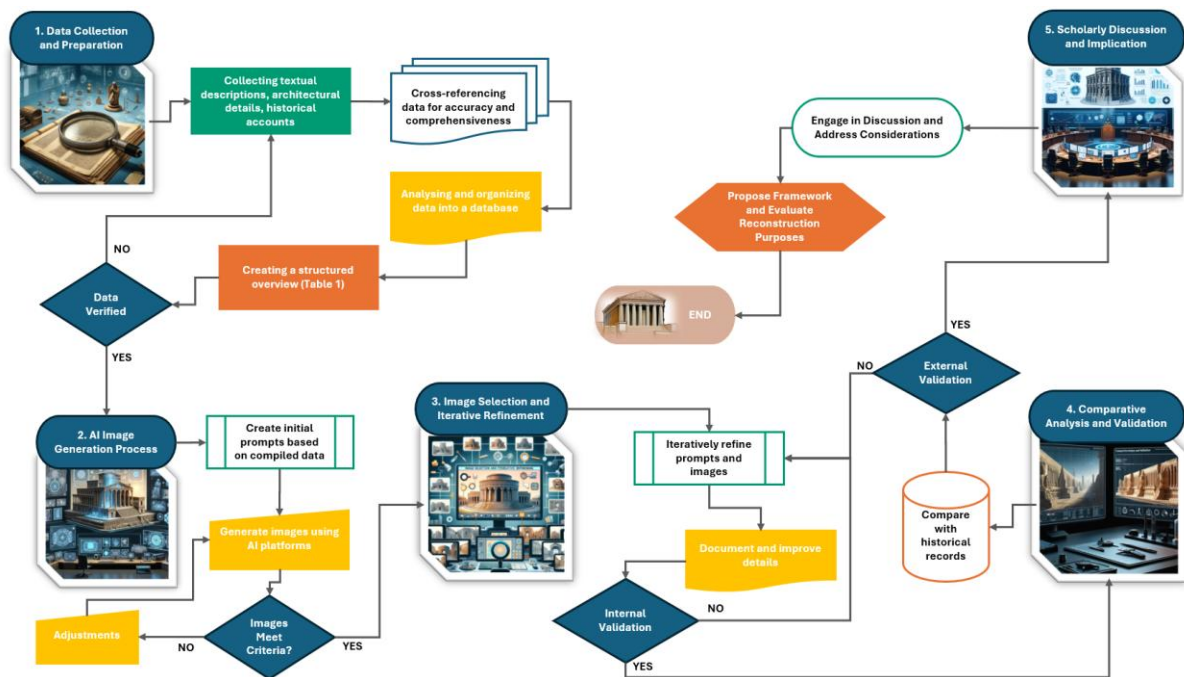


Figure 3. Flowchart illustrating the five-phase exploratory workflow; diagram produced for the thesis with AI-generated illustrative thumbnails.

3.2.1 Data Collection and Preparation

The initial phase involves the collection and analysis of textual descriptions, architectural details, historical accounts and existing visual records of selected heritage sites (Remondino, 2011). This extensive dataset is crucial for generating detailed prompts that provide a thorough historical context to guide the image-generation process. A diverse range of historical sources, including books, academic journals, architectural records and digital archives, was gathered to support prompt construction. The incorporation of official photographs from UNESCO and other reputable institutions improved the dataset by providing reference material for visual comparison with the generated candidates. The data collection process involved manually reading historical sources and articles to identify keywords related to the original construction of the heritage sites. These keywords are used to create comprehensive textual prompts for the AI generation process. The systematic extraction and organisation of specific phrases and terminology indicative of heritage sites' original construction methods were documented in a database, which was categorised by heritage site, historical period, architectural style and other relevant factors to facilitate their effective use in prompt generation.

3.2.2 AI Image Generation Process

Employing AI platforms such as Midjourney and DALL-E, this study implements a systematic approach to convert textual prompts into visual reconstruction candidates (Gualandi et al., 2021; Horn et al., 2022). Despite extensive research efforts aimed at uncovering photographic records to inform and compare candidate outputs, certain heritage sites lack definitive visual records, presenting a significant challenge for generating plausible visual candidates. Considering these constraints, the methodology relies more on prompt engineering, utilising detailed textual descriptions derived from available historical documentation and expert insights. However, it is important to note that the AI-generated images produced under these conditions served as initial approximations rather than exact replicas of the original structures. The textual prompts derived from the compiled data aimed to represent the architectural and cultural aspects of the targeted heritage sites. Multiple iterations and a variety of prompts are utilized to address the multifaceted nature of each site, facilitating a broad spectrum of visual outputs (Kadhim & Abed, 2023). Utilising the information presented in Table 4, initial prompts were formulated to encapsulate the fundamental architectural features, historical contexts and cultural significance of each heritage site. For instance, the Buddhas of Bamiyan are described in terms of their monumental statues, niches and surrounding features before they were destroyed. These prompts were iteratively fine-tuned based on AI-generated images. Each iteration

involved adjusting the textual descriptions to increase the specificity and visual coherence of the generated images. This process has been thoroughly documented to track modifications and advancements. AI platforms generate multiple images based on refined prompts and capture diverse perspectives and details. The generated images were assessed by using historical records and visual references. For heritage sites with restricted photographic records, the methodology relies on extensive textual descriptions and expert insights to guide the AI. For instance, for Palmyra, where many structures have been damaged or destroyed, archaeological reports and historical texts provide necessary information.

3.2.3 Image Selection and Iterative Refinement

The image selection and refinement phase examined how generated outputs changed in response to prompt revision. Candidate images were reviewed against available photographs, site descriptions and architectural information, with attention to visible form, material features, setting, scale and unsupported additions. Outputs that departed substantially from documented features, or that produced generic architectural forms with little relation to the site, were used to inform further prompt revision rather than being treated as final candidates.

Prompt refinement involved adjusting the wording of the prompt and regenerating images. The revisions focused on making site-specific descriptors clearer, including material descriptions, architectural form, surrounding setting and viewpoint (Oppenlaender, 2022), (Xie et al., 2023). This process improved the visual coherence of some outputs, but it did not establish their historical or cultural adequacy. At this stage, the work remained an internal comparison process.

The purpose of this phase was to produce a more coherent candidate set for the comparative stage that followed. It helped identify visibly unsuitable or weakly evidenced outputs and kept a basic record of how changes in prompt wording affected the generated images. Image selection and iterative refinement therefore functioned as a preliminary filtering step within the exploratory workflow.

3.2.4 Comparative Analysis and Preliminary Screening

This phase compared generated candidates with available historical sources, photographs and scholarly interpretations to identify visible correspondences and departures (George, 2022). The comparison focused on documented site features, including architectural form, visible materials, spatial setting, major structural elements and evident omissions. It was used to examine how far the outputs reflected the descriptors supplied through the prompts and where the images introduced unsupported or weakly evidenced features.

For the exploratory site examples, the comparison remained qualitative. Figure 5 presents Palmyra candidates generated from prompts referring to major architectural elements such as the Temple of Bel, while Figure 10 presents Pompeii candidates generated from prompts focused on visible Roman architectural and interior features. These outputs were compared with available photographs and descriptions to identify broad visual correspondence, omissions and unsupported additions. Prompt refinement was then used to adjust visible features where necessary.

3.2.5 Scholarly Discussion and Practical Implications:

The final phase considered the implications of the exploratory workflow for heritage visualisation, education and public interpretation (Bharati, 2023; Orengo et al., 2015). The generated images were treated as candidate visualisations. The discussion focused on the relation between source material, prompt wording, image output and preliminary comparison, rather than on the production of verified reconstructions.

Prompt engineering formed the main link between documentary evidence and image generation. Historical, architectural and archaeological descriptors were selected from available sources and converted into prompts for Midjourney and DALL-E (Oppenlaender, 2022). Subsequent revisions tested whether more specific wording produced clearer visual correspondence with known site features, materials and settings (Rombach et al., 2022; Xie et al., 2023). The prompt sequence therefore became part of the evidence trail: it recorded how textual material was selected, compressed and translated into visual form.

This phase also clarified the limit of the Chapter 3 workflow. A prompt may produce a coherent image, but that image may still conceal uncertainty, simplify competing interpretations, or present conjectural features as visually complete. Prompt refinement and documentary comparison therefore provide only preliminary screening. They do not provide the full evaluative structure required for heritage reconstruction claims. That problem is taken forward in Chapter 4 through HARE.

Two forms of comparison were used in this chapter. The first was documentary comparison, in which candidates were compared with available photographs, historical records and site descriptions to identify visible correspondences and departures. The second was metric comparison, applied to the Western Buddha candidates through SSIM, MSE, PSNR and MAE against the available reference image. Documentary comparison supported preliminary screening, while metric comparison provided image-level indicators of visual correspondence within the Bamiyan image set (Figure 12, Figure 14).

3.3 Results and Discussion

This section reports the outputs of the exploratory text-to-image workflow. The section first sets out the site context and technological background used to support prompt construction, before presenting the generated image candidates. Table 4 summarises the selected heritage sites, including their period, principal features, cause of damage, current condition and reference sources. Its purpose is to organise the historical and site-specific information used in the workflow. It should not be read as evidence that the AI system reconstructed missing architectural data or resolved uncertain features.

Table 4. Overview of Historical Heritage Sites: Characteristics and Conditions

Site	Country	Period Founded	Historical Period	Main Features	Destruction Year	Area	Causes of Damage	Current Condition	Damaged Parts	Key References
Palmyra	Syria	3rd millennium BC	Roman Empire, Islamic Caliphate	Temple of Bel, Great Colonnade	2015	Varies	Extremist group activities	Partially ruined	Temple of Bel, other structures	UNESCO , Britannica , (Elcheikh, 2019; Raja & Seland, 2022)
Buddhas of Bamiyan	Afghanistan	6th -7th century CE	Gandhara period	Two monumental statues	2001	105 hectares	Extremist group activities	Destroyed, niches remain	Two monumental statues	UNESCO , (Flood, 2002; Grün et al., 2004; Toubekis et al., 2017)
Ancient City of Nimrud	Iraq	13th century BC	Assyrian Empire	Palaces, temples, colossal statues	2015	Over 360 hectares	Extremist group activities	Ruined	Palaces, temples, statues	UNESCO , (Pollegioni et al., 2015; Reade, 2002)
Pompeii	Italy	6th-7th century BC	Roman Empire	Urban infrastructure, frescoes	79 AD	66 hectares	Natural disaster (Volcanic Eruption)	Ruins preserved	Entire city	UNESCO , Britannica , (Fulford & Wallace-Hadrill, 1999; Scorrano et al., 2022; Senatore et al., 2014; Wotzlaw et al., 2022)
Hampi	India	14th century AD	Vijayanagara Empire	Temples, royal complexes	1565	4187, 24 hectares	Historical Conflict	Ruins	Temples, marketplaces	UNESCO , (Knell, 2022; Pakkeerappa & Thomas, 2006; Powell, 2018; Rajangam & Sundar, 2021;

										Schettino, 2016)
Parthenon (Acropolis of Athens)	Greece	447-432 BC	Classical Greece	Doric temple dedicated to Athena	1687	69.5 by 30.9 m (228 by 101 ft)	Military Bombardment	Partially ruined	Structure, sculptures	UNESCO , Britannica , (Angouri et al., 2017; Fincham, 2012; Hamilakis, 2002)
Temples of Thebes	Egypt	1500 - 30 BC	New Kingdom of Egypt	Karnak and Luxor temples	Various periods	7,390 ha with a buffer zone of 444 ha	Natural and Human Impact	Partially ruined	Karnak, Luxor temples	UNESCO , Britannica , (Aubry et al., 2009; Mahmoud et al., 2019; Manning, 2012)
The Great Wall	China	7th century BC and onwards	Various Chinese dynasties	Fortification system	Ongoing	21,196 km (total length)	Erosion and Human Impact	Partially ruined	Sections over time	UNESCO , Britannica , (Cao et al., 2023; Shelach-Lavi et al., 2020; M. Yang, 2017)
Maya Sites	Mexico	2000 BC to 16th century AD	Pre-Columbian Maya civilization	Temples, pyramids	Post 9th century AD	331,397 ha, surrounded by a buffer zone of 391,788 ha	Sociopolitical Decline	Ruins	Several city structures	UNESCO , Britannica , (Chase et al., 2011; Scherer, 2007)
Angkor Wat	Cambodia	Early 12th century	Khmer Empire	Temple complex	Not specified	Over 162 hectares	Neglect and Decay	Partially restored	Temples, infrastructure	UNESCO , Britannica , (Carter et al., 2019; Chen, 2021; Chen et al., 2016; Evans et al., 2013)

Table 4 provides the contextual basis for prompt development. The listed features, damage histories and reference sources helped identify descriptors used in prompt formulation, including monument type, material, setting, architectural form and visible condition. The table also shows that the selected examples could not be treated as a uniform reconstruction set. They differ in date, material condition, damage history, available documentation and surviving visual evidence. Table 4 therefore functions as contextual and organisational evidence for prompt construction, not as a validation dataset.

Table 5 records documentation and reconstruction technologies previously applied to the selected sites by other projects. For Palmyra, previous work has used public-domain imagery, professional panoramic imagery, crowdsourced photographs and 3D photogrammetric processes to support digital representation of the Temple of Bel and related heritage material (Wahbeh et al., 2016), (Denker, 2017; Pan et al., 2020). For Maya sites, photogrammetry, reality-based documentation tools and the MayaArch3D project show the role of 3D modelling, orthoimage generation and WebGIS in analysing ancient Maya architecture and landscapes (Pierrot-Deseilligny et al., 2011)-(von Schwerin et al., 2013). Pompeii is represented through work on heritage information systems and parametric approaches to Roman architectural components (Mazzaglia, 2021), (Sbrogiò, 2022). For Bamiyan, historical photography, photogrammetry from archival imagery, 3D modelling of the present condition, laser recording and virtual reconstruction projects demonstrate the importance of source images and spatial documentation in prior reconstruction work (Maiwald et al., 2017; Spallone et al., 2022; Toubekis et al., 2017)-(Bevilacqua et al., 2019; Heikkinen, 2009). The Ancient City of Nimrud illustrates the use of 3D GIS, CityGML and archaeological modelling for urban reconstruction and spatial analysis (Remondino, 2011), (Biljecki et al., 2015), (Gröger & Plümer, 2012).

Table 5. Technological Approaches in Heritage Site Reconstruction

Site	Author	Photogrammetry	3D GIS	CityGML	Parametric Approaches	3D Reconstruction Modelling	EMF	3DMM	2D-ERT	TOF Cameras	Point Cloud	Deep Learning Methods	MayaArch3D	Reality-based 3D Documentation	Information System	3D WebGIS	Crowdsourcing	Panoramic Imagery	Remote Sensing Data	3D laser-scanning	Historical Photography	VR Technologies
Palmyra	(Denker, 2017; Pan et al., 2020; Wahbeh et al., 2016)	✓				✓												✓				
Maya Sites	(Pierrot-Deseilligny et al., 2011; von Schwerin et al., 2013)	✓											✓			✓						
Pompeii	(Mazzaglia, 2021; Sbrogiò, 2022)					✓										✓						
Buddhas of Bamiyan	(Bevilacqua et al., 2019; Heikkinen, 2009; Maiwald et al., 2017; Spallone et al., 2022)	✓				✓														✓	✓	✓
Ancient City of Nimrud	(Biljecki et al., 2015; Gröger & Plümer, 2012; Remondino, 2011)	✓	✓	✓																		
Hampi	(Keyvanfar et al., 2022; Natampally, 2014)	✓				✓																
Parthenon	(Debevec, 2004; Hu et al., 2013)																			✓		

Temples of Thebes	(Bennoui-Ladraa & Chennaoui, 2018; Berto et al., 2021; Hu et al., 2013; Meister et al., 2021)	✓					✓	✓	✓											
The Great Wall	(Bassier et al., 2020; Zeng & Jin, 2023; Zhou et al., 2022)									✓	✓	✓								
Angkor Wat	(Shishido et al., 2017; X. Wang et al., 2020)	✓			✓												✓		✓	

Table 5 therefore situates the exploratory workflow in relation to established digital heritage methods. These methods are relevant because they show how heritage reconstruction depends on source material, documentation quality, spatial evidence and evidential control. They were not implemented or validated in this chapter. The empirical work reported here remains confined to two-dimensional text-to-image generation and preliminary comparison. The table provides background for understanding the evidential standards surrounding digital reconstruction, while the outputs in this chapter remain candidate visualisations rather than products of a validated reconstruction pipeline.

The relevance of AI to heritage visualisation also needs to be read against wider debates in architectural design and education. Traditional architectural practice remains grounded in material knowledge, cultural continuity, sustainability and inherited design values (T. Liu et al., 2019). AI-based design tools introduce a different mode of visual production, in which prompts, model priors and generated outputs mediate the relation between design intention and visual form (Hosseini et al., 2016). This shift raises ethical and methodological concerns, particularly around authorship, human creativity, evidential control and the preservation of traditional design judgement (Hegazy & Saleh, 2023). These concerns are also reflected in architectural education, where AI tools are increasingly discussed as part of changing design pedagogy rather than as replacements for disciplinary expertise (Sadek et al., 2024). In this chapter, that debate provides context for the exploratory use of text-to-image generation, while the generated outputs remain candidates requiring further evaluation.

The figures below present the image outputs from the exploratory workflow. Figure 4, Figure 6 and Figure 9 provide reference photographs for Palmyra, Maya sites and Pompeii. Figure 5, Figure 7, Figure 8 and Figure 10 show the corresponding AI-generated candidates. These outputs are included to show the range of visual forms produced from historically informed prompts across different heritage contexts. They are used as exploratory visualisations, not as verified reconstructions.

Figure 11 has a narrower role. It presents the reference photograph of the Western Buddha in its niche alongside candidate images generated for the Bamiyan case. These Bamiyan candidates form the image set later assessed through SSIM, MSE, PSNR and MAE. The section therefore distinguishes the broader exploratory examples from the focused Bamiyan image-comparison exercise.

The following prompts are reproduced as generation inputs from the exploratory workflow; their restoration-oriented wording should not be read as claims that the resulting images achieved historical accuracy.

3.3.1 Palmyra (Syria)

Prompt: a newly reconstructed Palmyra, where the essence of its ancient glory is captured in a structure that stands proud and unblemished under the clear sky. At its heart, a grand edifice, inspired by the Temple of Bel, showcases a symmetrical front view with a grand entrance. The architecture boasts complete and pristine columns rising to support a perfectly restored entablature, capped with a pediment that marries timeless artistry with contemporary innovation. This masterpiece gleams in the sunlight, reflecting a blend of historical reverence with the pinnacle of modern reconstruction. The surrounding environment is meticulously maintained, featuring a well-kept street leading to the temple's imposing steps, inviting visitors into a space where history and the present are seamlessly intertwined. Capture the essence of both ancient grandeur and modern architectural achievements, highlighting the harmony between past and future Figure 4 and Figure 5.

(a)



(b)



(c)



Figure 4. Reference photographs of the surviving ruins of Palmyra, including the Temple of Bel, the Baalshamin Temple and the ancient theatre. Source: Ko Hon Chiu Vincent/UNESCO.

(a)



(b)



(c)



Figure 5. AI-generated exploratory reconstruction candidates for sections of Palmyra, produced using Midjourney and DALL-E.

3.3.2 Maya Sites (Mexico)

Prompts:

- a. Chichen Itza in its heyday, focusing on the Pyramid of Kukulcan and the Great Ball Court. Show the precise architectural and astronomical alignment of the pyramid, the ornate carvings and the surrounding temples and sacred cenotes. The lush Yucatan jungle should frame the site, highlighting the Maya's connection with their environment. realistic, architectural design Figure 6.
- b. A detailed 3D model description of the complete architectural layout of Chichen Itza, focusing on the overall arrangement of the ancient Maya city's buildings and structures. Describe the spatial relationships between the step-pyramids, temples, ball courts, observatory and other key architectural features, providing information on their relative sizes, orientations and distances from one another. Emphasize the architectural aesthetics and the visual impact of Chichen Itza's layout, allowing for a comprehensive representation of this ancient Maya city's architectural grandeur Figure 7.

(a)



(b)



(c)



Figure 6. Reference photographs of Maya archaeological sites in Mexico. Source: Community Tours Sian Ka'an and Ko Hon Chiu Vincent.

(a)



(b)



Figure 7. AI-generated exploratory reconstruction candidate of the architectural layout of Chichen Itza, produced using Midjourney.

- c. The intricate architectural features of Chichen Itza, emphasizing the design and construction of the ancient Maya city's iconic buildings and structures. Highlight the specific elements of architecture, such as the step-pyramids, temples, ball courts and observatory and provide insights into their unique architectural styles, ornamentation and any notable construction techniques or materials used. Capture the essence of Chichen Itza's architectural marvels in their prime, showcasing the cultural and artistic achievements of the ancient Maya civilization Figure 8.

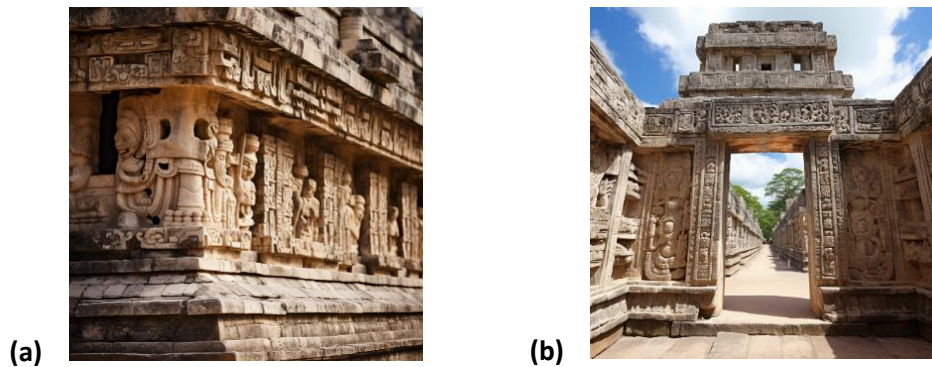


Figure 8. AI-generated exploratory reconstruction candidates showing architectural features of Chichen Itza, produced using Midjourney.

3.3.3 Pompeii (Italy)

Prompt: Create a hyper-realistic image reconstructing the specific section of the ancient Roman building in Pompeii, Italy. The reconstruction should focus only on the visible section in the original image, maintaining the same perspective and architectural elements. The mosaic patterns, column details and architectural style should match the original as closely as possible, restored to how they might have appeared before any damage. The setting remains the same with the surrounding ruins still visible, but the specific section should appear fully restored with accurate colours and details Figure 9 and Figure 10.



Figure 9. Reference photographs of the ruins of Pompeii. Source: J. Frias Velatti and Limes Media/Tim Schnarr.

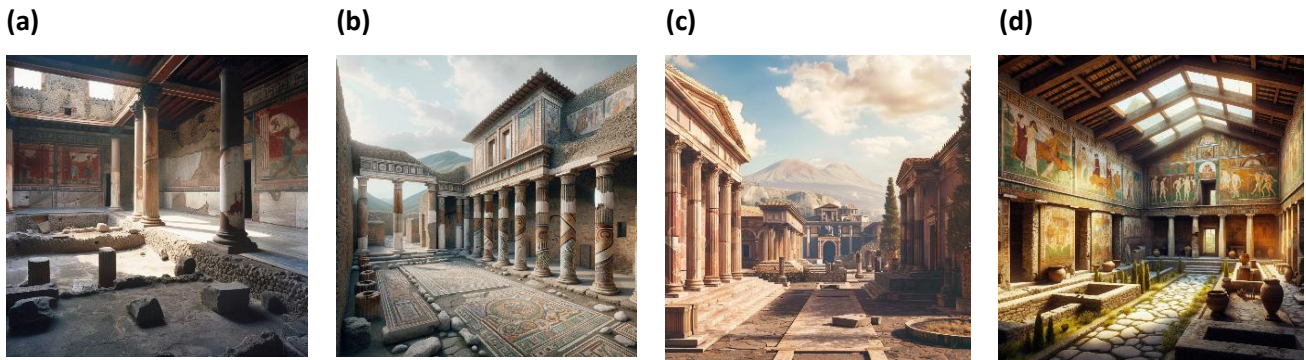


Figure 10. AI-generated exploratory reconstruction candidates showing Pompeii architecture and interiors, produced using Midjourney and DALL-E.

3.3.4 Buddhas of Bamiyan (Afghanistan)

Prompt: Reconstruct the standing Buddha of Bamiyan, standing 180 feet tall and carved into a sandstone cliff. The Buddha is depicted in a standing posture with intricate, layered robes that showcase a blend of Hellenistic and Indian artistic influences. The statue's face reflects a serene, meditative expression typical of Indian Buddhist art. The materials used are traditional mud, straw and lime plaster, giving it an aged yet restored look. The rugged mountainous terrain of the Bamiyan Valley surrounds the Buddha, emphasizing its monumental scale and cultural significance. Side architectures include detailed carvings and Kharosthi scripts to enhance historical accuracy Figure 11.

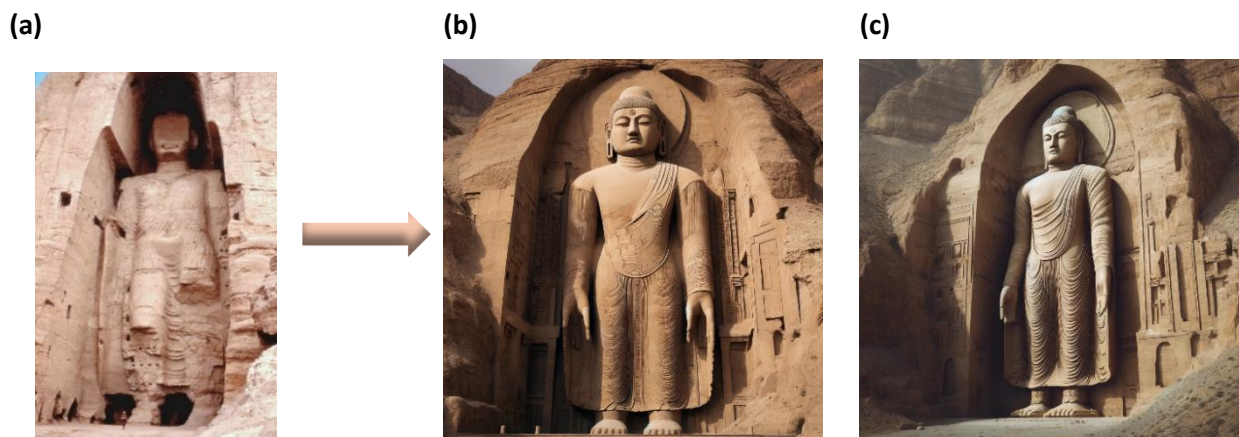


Figure 11. Western Buddha of Bamiyan: (a) reference photograph of the Western Buddha in its niche. Source: James Gordon/Wikimedia Commons; (b–c) AI-generated reconstruction candidates produced using Midjourney/DALL-E.

A set of image-quality metrics – the Structural Similarity Index Measure (SSIM), Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR) and Mean Absolute Error (MAE) – was applied to ten Western Buddha candidate images, labelled Images 0–9. Each candidate was compared with the available reference photograph of the Western Buddha in its niche. The purpose of this comparison

was to record image-level correspondence in structure, pixel difference and signal distortion, rather than to determine historical or cultural adequacy.

The results show a narrow and low range of structural similarity across the candidate set. SSIM values ranged from 0.083 to 0.125, with Image 9 recording the highest score and Image 6 the lowest. Although Image 9 was the strongest candidate on this measure, the absolute SSIM value remains low. The MSE and PSNR results point in the same direction. MSE ranged from 7,568.117 to 12,674.829, with Image 9 producing the lowest squared pixel-level difference and Image 4 the highest. PSNR ranged from 7.101 to 9.341 dB, again with Image 9 at the upper end of the range and Image 4 at the lower end. These values indicate limited pixel-level correspondence with the reference photograph. MAE produced a different ordering. Values ranged from 135.359 to 148.749 on the 0–255 intensity scale. Image 4 recorded the lowest MAE value, while Image 9 recorded the highest. Image 9 is therefore strongest on SSIM, MSE and PSNR, but weakest on MAE. The metric set does not produce a single uniform ranking across all four measures.

Figure 12 summarises the scores for all ten Western Buddha candidates. Figure 13 presents the detailed panel for Image 3 and Figure 14 presents the detailed panel for Image 9. Taken together, the results show that the strongest candidate on three of the four measures remains within a low range of structural and pixel-level correspondence. The metric comparison therefore supports a cautious interpretation: the generated images may be visually coherent as candidate images, but their measured similarity to the reference photograph is limited.



Figure 12. Comparison of image-quality metrics for ten Western Buddha candidate images, labelled Images 0–9.

Comparison Metrics for Image 3

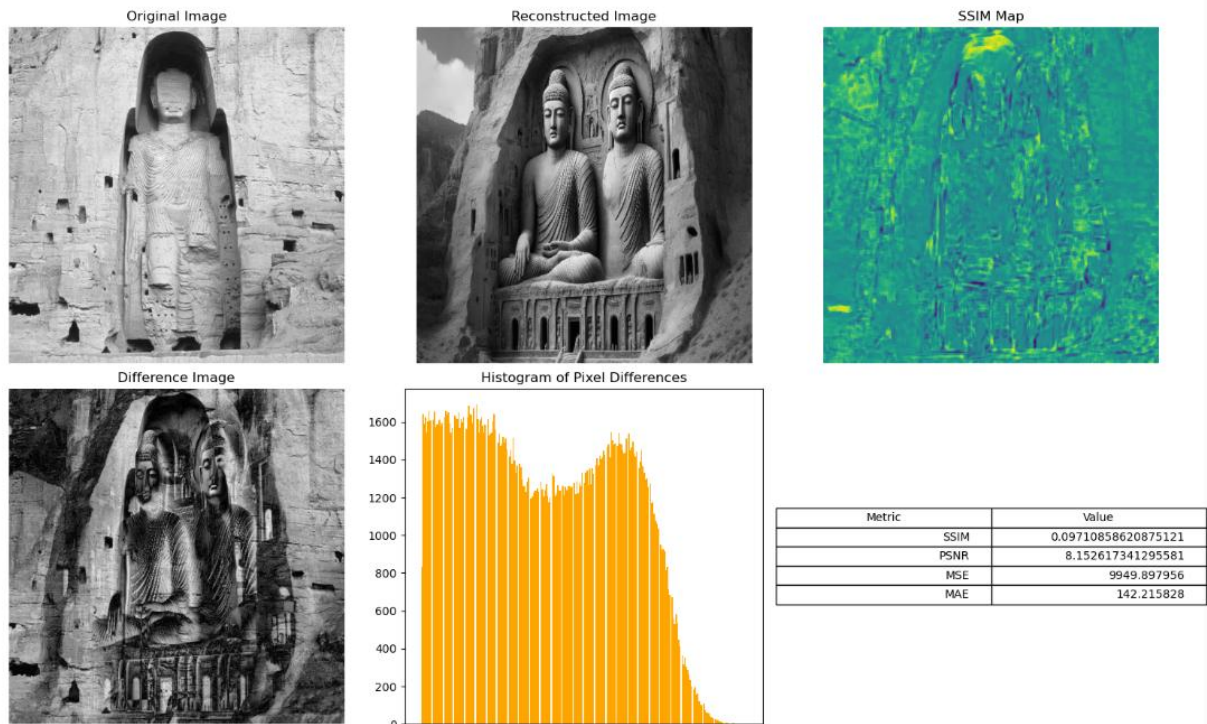


Figure 13. Per-image metric panel for Image 3, showing the reference image, AI-generated candidate image, SSIM map, difference image, pixel-difference histogram and metric values.

Comparison Metrics for Image 9

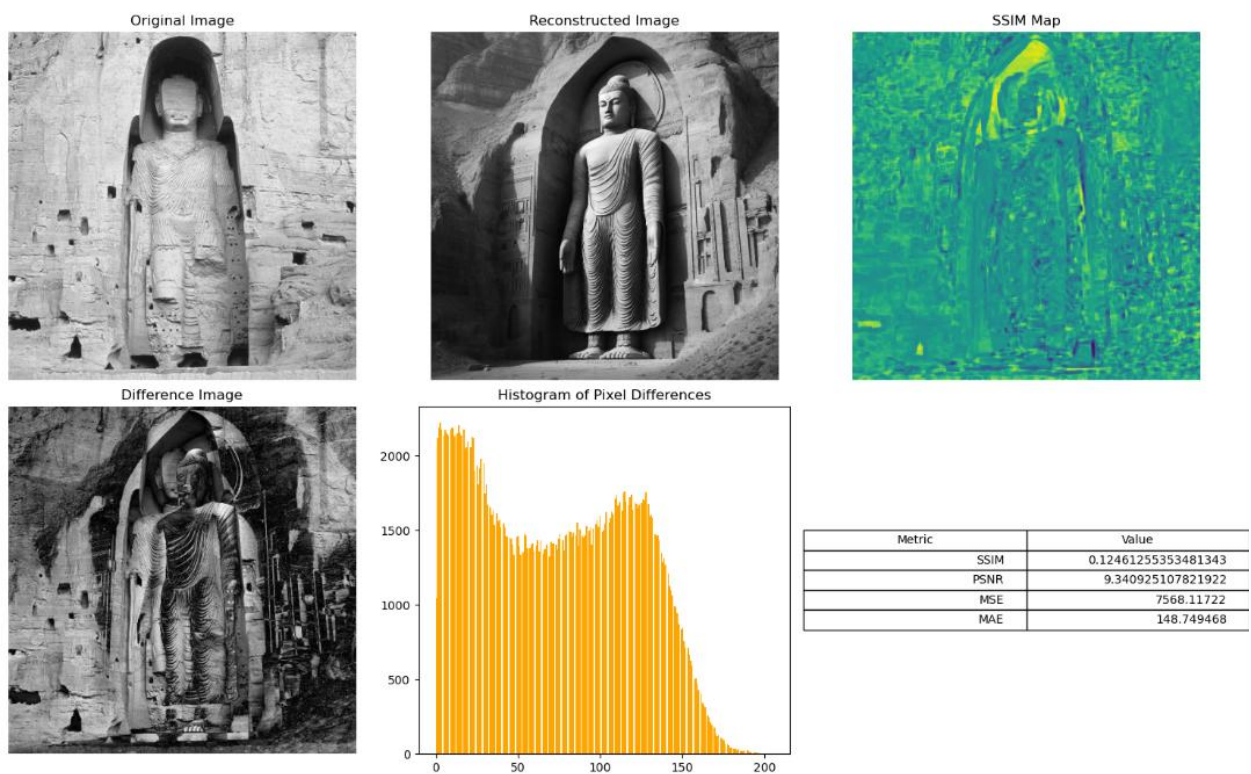


Figure 14. Per-image metric panel for Image 9, showing the reference image, AI-generated candidate image, SSIM map, difference image, pixel-difference histogram and metric values.

3.4 Limitations and Future Directions

This study identifies several limitations that affect the extent to which AI-generated outputs can be treated as heritage reconstruction candidates. These limitations concern the availability of reference material, the manual nature of prompt construction, output variability, limited reproducibility and the absence of structured expert evaluation within this early workflow.

Lack of original visual references: One major limitation is the uneven availability of pre-destruction photographs and visual references. Where reference material is incomplete, direct metric comparison becomes limited or impossible. In such cases, evaluation must rely more heavily on documentary sources, archaeological interpretation and expert judgement. This limitation shows why generated candidates cannot be assessed through image metrics alone.

Automated information extraction and prompt engineering: The workflow relied on manually curated prompts derived from historical documents and academic sources. This limited both coverage and consistency. Future work should examine whether automated or semi-automated extraction of site features can support more systematic prompt construction. Such work would need to preserve source traceability so that generated outputs remain connected to the evidence from which prompts were derived.

Consistency and reproducibility in AI outputs: Text-to-image systems can produce variable outputs from similar prompts, making it difficult to reproduce candidates across iterations or viewpoints. This limits their use as heritage reconstruction evidence. Future work should record platform, model version, prompt wording, generation settings and refinement history so that outputs can be audited and, where possible, reproduced.

Detail-oriented reconstruction candidates: Generated images depend heavily on the specificity of the input prompt and the quality of the supporting evidence. Dimensions, materials, spatial relationships and iconographic details are often incomplete in textual sources. More detailed prompts may improve visual coherence, but they do not remove uncertainty. Future work should therefore connect prompt detail to explicit source documentation and expert assessment.

Methodological constraints: The generated outputs remain interpretive candidates produced from incomplete evidence and model-mediated visual priors. They may include plausible but unsupported details. Collaboration with historians, archaeologists, architects and relevant heritage communities is therefore necessary before any generated candidate can be treated as an accountable reconstruction. This limitation directly motivates the structured evaluation procedure developed in Chapter 4.

These limitations position the Chapter 3 workflow as an exploratory stage. They also define the requirements addressed in the subsequent HARF framework: source traceability, prompt transparency, computational screening, expert review and clearer limits on reconstruction claims.

3.5 Conclusion

This chapter examined an exploratory workflow for generating two-dimensional heritage reconstruction candidates from textual, photographic and architectural evidence. The workflow was tested across Bamiyan, Palmyra, Pompeii and Maya examples. It showed that text-to-image systems can produce visually coherent candidates from historically informed prompts, while also showing that visual coherence is only one part of the reconstruction problem.

The quantitative comparison was limited to ten Western Buddha candidates assessed against the available reference photograph of the Western Buddha in its niche. SSIM ranged from 0.083 to 0.125 and PSNR from 7.101 to 9.341 dB, with Image 9 highest on both measures. MAE ranged from 135.359 to 148.749, with 148.749 representing the maximum of the reported range. These scores place the candidates in a low-correspondence range, even where the generated images remain visually coherent.

The results separate visual coherence from measured similarity. Image-level metrics can indicate structural and pixel-level correspondence with a reference image, but historical plausibility, evidential adequacy, iconographic accuracy and cultural meaning require forms of evaluation beyond pixel-based comparison. This is the methodological gap addressed in the next stage of the thesis.

The workflow examined in this chapter demonstrates that text-to-image generation can support the visualisation of heritage reconstruction candidates from documentary, photographic and architectural evidence. It also exposes the limit of generation-led reconstruction. The workflow provides an initial form of evidential constraint, but it does not yet define a formal prompt schema, prompt sufficiency measure, paradata protocol, computational pre-evaluation procedure or structured expert-review instrument. Chapter 4 addresses that limitation by formalising HARF as the evaluation framework through which GenAI reconstruction candidates are assessed as heritage claims rather than as visually plausible images.

CHAPTER 4

HARF Development and Operationalisation

Chapter 3 established the practical value and methodological limit of evidence-informed text-to-image reconstruction. It showed that dispersed documentary, photographic and architectural evidence can be translated into visual reconstruction candidates, but that the resulting images remain difficult to evaluate unless source use, prompt construction, refinement decisions and expert judgement are made explicit.

This chapter turns that limitation into the methodological framework of the thesis. It formalises the Heritage-Aligned Reconstruction Framework as a structured procedure for evaluating GenAI heritage reconstruction. HARF specifies the Dynamic Prompt Blueprint, Prompt Sufficiency Index, paradata record, computational pre-evaluation and structured expert review. These components standardise the internal evaluation workflow used in this thesis; they do not establish a field-wide standard or claim that generated outputs are historically accurate. The chapter reports the framework design and its initial expert-based operationalisation; the fuller candidate-level Bamiyan evaluation is reserved for Chapter 5.

4.1 Introduction

Cultural heritage across the world is increasingly threatened by armed conflict, ideological destruction and environmental degradation, leaving many monuments accessible only through fragmentary archives and unstable historical memory (Brosché et al., 2017; Kila, 2019; Sloggett & Scott, 2022; Veysel Apaydin, 2020). Among the most emblematic cases is the destruction of the sixth- to seventh-century CE Buddhas of Bamiyan in Afghanistan in March 2001, monumental sculptures widely recognised as masterpieces of Gandhāran Buddhist art. Their absence and the limited photographic documentation that survives, continues to pose a methodological dilemma for digital restitution: how can vanished monuments be reconstructed in ways that respect their historical specificity without drifting into aesthetic speculation or modern reinterpretation (Flood, 2002)?

The rapid advance of generative artificial intelligence has been positioned as a potential response to this dilemma. Diffusion-based systems such as DALL·E, Midjourney and Stable Diffusion can produce visually compelling imagery from textual prompts, enabling new forms of cultural engagement, exhibition and digital preservation (Ajuzieogu, 2021; Z. He et al., 2024; Saharia et al., 2022). Pilot studies have demonstrated the capacity of these models to approximate lost architectural forms or narrate intangible heritage, underscoring their expressive range. Yet this technical promise is accompanied by well-documented risks that are particularly acute in non-Western contexts. Because these models are trained on large, uncurated web corpora dominated by Western aesthetics, they tend to reproduce three recurring errors: stylistic drift, where generated images adopt Westernised lighting or surface treatments; historical anachronism, where modern materials or geometries intrude; and cultural misrepresentation, where region-specific iconographic conventions are flattened, generalised or replaced by model-derived visual defaults (Foka & Griffin, 2024a; Muldoon & Wu, 2023). Together, these issues create a credibility gap: a divergence between visual plausibility and historical authenticity that undermines scholarly trust and curatorial responsibility.

Existing research highlights these challenges. Feasibility studies show that multimodal models can approximate lost architectural forms (Arzomand et al., 2024; Shih, 2025), yet they rarely document the evidentiary basis for their outputs or disclose the interpretive decisions embedded in prompt construction. Benchmarking initiatives such as the Time Travel Benchmark (Ghaboura et al., 2025) offer quantitative metrics but remain disconnected from cultural or ethical considerations. Meanwhile, participatory and community-centred approaches (Foka et al., 2025; Nikolakopoulou & Koutsabasis, 2025) promote inclusivity but are seldom integrated within computational workflows. Across this literature, three methodological deficiencies remain unresolved: prompt design and evaluation are inconsistent and often opaque; domain knowledge is loosely coupled with algorithmic scoring; and reconstruction assumptions are insufficiently documented. These limitations are most visible in non-Western heritage contexts, where training data frequently misrepresent local materials, proportions and iconographic canons (Chang et al., 2023; Nyaaba et al., 2024; S. Tiribelli et al., 2024).

To address these methodological and ethical challenges, this study develops and validates a reproducible, expert-in-the-loop approach for culturally grounded AI-mediated reconstruction. We refer to this integrated approach as the Heritage-Aligned Reconstruction Framework (HARF). HARF structures the reconstruction-evaluation workflow through five connected components. First, the Dynamic Prompt Blueprint (DPB) translates verified architectural dimensions, iconographic evidence and contextual knowledge into a layered prompt structure. Second, prompt sufficiency is controlled through a construction-stage PSI completeness gate, while Phase I expert responses provide a separate assessment of whether the prompt structure captured the historical, architectural, cultural

and artistic information needed for generation. Third, paradata records the evidential sources, interpretive decisions and refinement history of each generated candidate. Fourth, computational pre-evaluation uses two-dimensional structural screening against reference imagery to identify proportional and compositional drift before expert review. Fifth, structured expert evaluation assesses historical, iconographic and cultural plausibility beyond what image metrics can capture. Together, these components establish the evidential, computational and interpretive foundations for the Bamiyan evaluation developed in Chapters 4 and 5.

The Bamiyan implementation formed part of a single staged expert study. Phase I involved thirty-three interdisciplinary experts who assessed prompt sufficiency, representative reconstruction candidates and areas requiring refinement. Phase II involved five Bamiyan-focused specialists who later assessed the refined candidate set through the site-specific specialist evaluation reported in Chapter 5. Chapter 4 reports the framework design and its initial operationalisation; Chapter 5 reports the candidate-level computational and specialist evaluation.

This study makes four methodological contributions. First, it formalises DPB as an evidence-based prompt engineering model that embeds verified cultural and contextual knowledge. Second, it introduces PSI as a construction-stage measure of prompt completeness. Third, it establishes a two-dimensional computational pre-evaluation protocol for screening generated outputs before expert review. The protocol detects proportional inconsistencies and large-scale geometric drift, but remains less sensitive to fine anatomical, iconographic and stylistic deviations, which require specialist judgement. Fourth, it demonstrates how expert-in-the-loop evaluation reconciles algorithmic efficiency with interpretive depth, operationalising the ethical principles of the London Charter (Denard, 2009) and the Seville Principles (López-Menchero Bendicho & Grande, 2011) within a transparent reconstruction workflow.

4.2 Methodology

The HARF was implemented to the Western Buddha of Bamiyan through a sequential mixed-methods workflow that combines historical documentation, architectural analysis, structured prompt engineering, quantitative image comparison and expert review. Within this framework, the documentary record defines the evidentiary limits; architectural and iconographic analysis constraints prompt formulation, quantitative metrics assess geometric and contextual fidelity and expert evaluation addresses inconsistencies not captured computationally. These components form the operational structure of HARF (Figure 15).

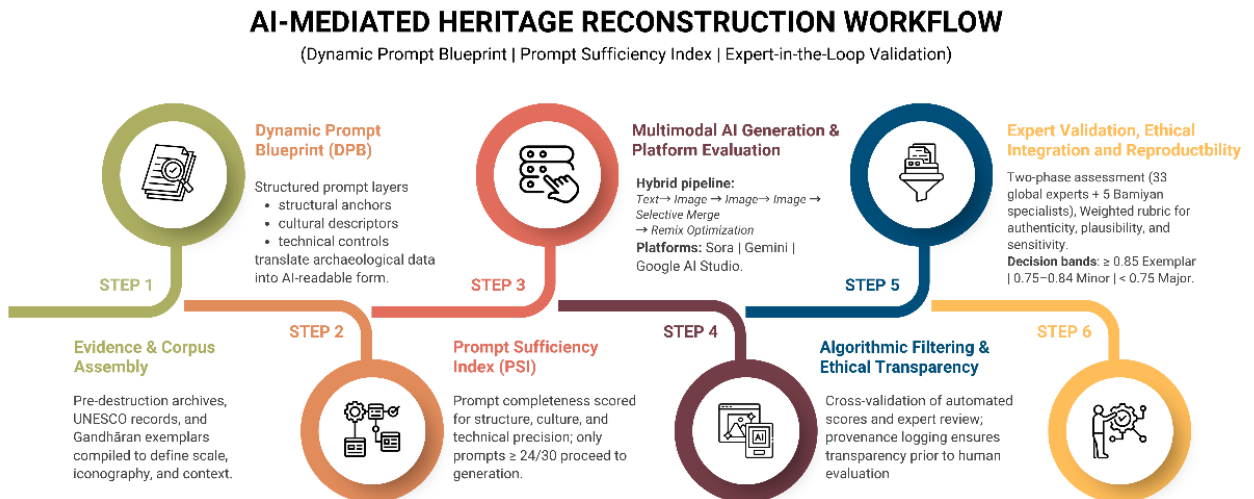


Figure 15. Author-generated overview diagram of the Heritage-Aligned Reconstruction Framework (HARF) and its six interconnected stages for the reconstruction of the Western Buddha of Bamiyan.

4.2.1 Case Study and Evidentiary Corpus

The Western Buddha of Bamiyan serves as the primary case study through which HARF was developed and operationalised. The site was selected for three reasons: the monument is no longer accessible for direct documentation after its destruction in 2001; the surviving documentary record is substantial but incomplete, allowing reconstruction to be anchored in verifiable dimensional and iconographic evidence; and the monument represents one of the most consequential cases of cultural loss in South and Central Asia, requiring a reconstruction approach that balances historical specificity with methodological transparency. The statue occupied the western niche of a monastic cave complex carved into the Bamiyan Valley's sandstone cliffs. It is treated in this thesis as a sixth- to seventh-century CE monument associated with the Gandhāran–Central Asian sculptural tradition (Figure 16).

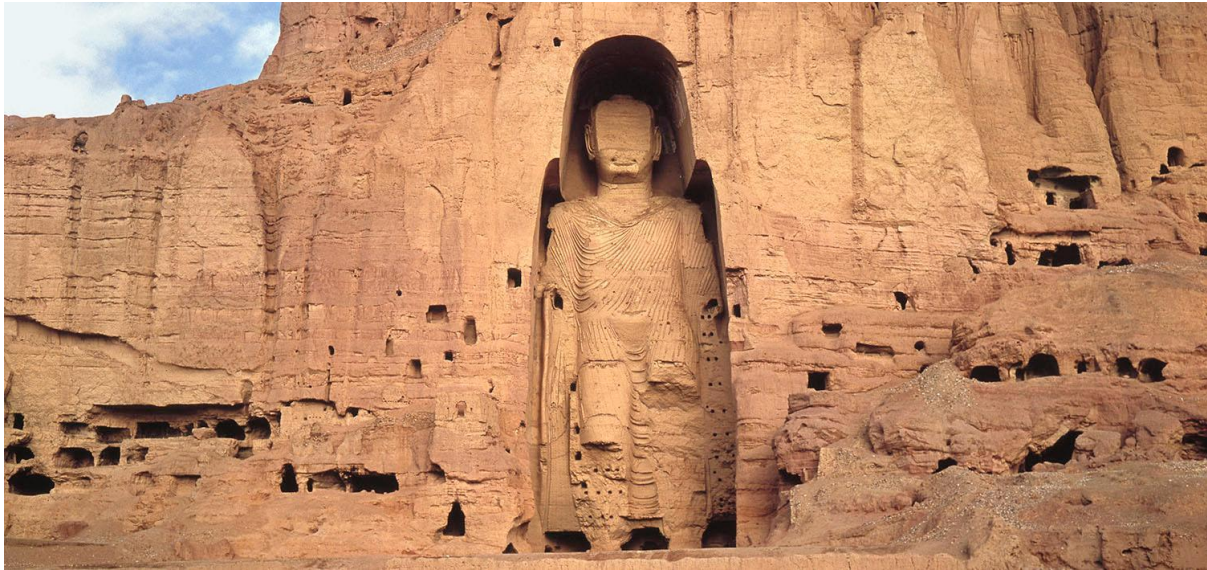


Figure 16. Pre-2001 view of the Western Buddha of Bamiyan showing the niche architecture, drapery contours and surrounding monastic caves. Source: Khan Academy, UNESCO.

A structured evidentiary corpus was assembled to support dimensional grounding, environmental context and iconographic analysis. The corpus integrates archaeological, architectural and conservation documentation and draws on institutional and scholarly sources including UNESCO materials, Afghan archaeological records, SPACH-related documentations and academic repositories. Early photographic campaigns by DAFA (Petzet, 2009; Toubekis et al., 2017) and the German Archaeological Institute record proportional relationships, drapery rhythms and niche morphology (Rienjang & Stewart, 2023). These visual records were supplemented by the Japanese cultural missions of the 1970s, whose measured drawings and architectural surveys remain among the most precise records of the niche's geometry, including vaulted curvature, pedestal organisation and wall contours. Art-historical scholarship clarifies robe structuring, layering conventions and stylistic lineage within the region's sculptural idiom (Klimburg-Salter, 1989). Together, these sources define the dimensional and iconographic evidence base on which historically grounded reconstruction depends (Bevan & Lake, 2016; UNESCO World Heritage Centre, 2011).

Post-destruction photogrammetric and laser-scanning campaigns coordinated through UNESCO documented the surviving niche, pedestal traces, wall contours and vault outline. To enable quantitative comparison with generated imagery, these records were aligned with pre-2001 photographs and measured drawings to establish a consistent spatial reference for the two-dimensional evaluation workflow. Archival measurements for niche height, shoulder width, pedestal proportions and drapery fall lines were derived from measured drawings, architectural surveys, pre-2001 photography and post-destruction photogrammetric and laser-scanning records. These measurements were cross-checked across independent documentary sources and standardised within the HARF workflow as dimensional and proportional anchors. Reference outlines traced from

the clearest pre-2001 imagery then served as geometric anchors for later two-dimensional computational screening.

Where records diverged, the most consistently attested or technically precise values were retained, and remaining uncertainties were documented as paradata in line with the London Charter (Denard, 2009) and subsequent work on interpretive accountability in digital heritage (Bentkowska-Kafel & Denard, 2012; Richards-Rissetto & von Schwerin, 2017). Features lacking sufficient evidence, such as facial attributes eroded long before 2001, were preserved as indeterminate, establishing a clear evidentiary boundary between reconstruction and speculation.

To ensure consistency across HARF's analytical stages, corpus materials were digitised, standardised and catalogued within a structured project repository. Archival photographs were harmonised in resolution where necessary and annotated with viewpoint, source and orientation information to support spatial alignment. Architectural drawings were registered against a shared reference frame where sufficient control information was available, enabling comparison across documentary sources without implying complete survey-grade georeferencing. Textual documentation, including UNESCO mission reports, Japanese survey notes and archaeological monographs, was catalogued using structured metadata fields informed by linked-data and FAIR documentation principles (W3C, 2020), (Wilkinson et al., 2016). Each record specified provenance, date, source type, viewpoint where relevant, technical quality and relevance to architectural or iconographic features. Descriptive metadata and interpretive paradata were kept separate so that source attributes and reconstruction decisions could be audited independently.

From this corpus, dimensional anchors were formalised to guide the reconstruction workflow. The validated measurements and reference outlines were converted into geometric benchmarks used for computational pre-evaluation within HARF. Iconographic and contextual anchors, including robe layering conventions, fold treatment, sandstone texture and the configuration of adjacent monastic caves, were derived from photographic, survey and textual sources and encoded within the DPB. In accordance with current norms emphasising transparency and reproducibility in digital heritage (Huvila, 2022b; Marwick, 2022), only elements supported by verifiable documentation were incorporated into these operational layers.

This evidentiary foundation operationalises the DPB by converting verified documentation into generative text layers encoding factual, iconographic and contextual parameters. The same evidentiary logic can be re-instantiated for other heritage contexts where comparable archival, dimensional and iconographic documentation exists. The corpus therefore provides the evidentiary parameters within which reconstruction proceeds and ensures that subsequent methodological stages remain accountable to the historical record.

4.2.2 Dynamic Prompt Blueprint (DPB) Architecture

Prompt engineering in this study was guided by the Dynamic Prompt Blueprint (DPB), developed within this thesis to structure the translation of archaeological, stylistic and technical evidence into generation instructions. The DPB reduces reliance on generic prompt writing by separating evidential content into structural, cultural-iconographic and technical-control layers. It treats the prompt as a documented reconstruction instrument rather than a creative text description. The Bamiyan prompt record generated through this procedure is reproduced in Appendix C.9.

The first layer, the structural dimension, defines proportional and spatial anchors derived from architectural documentation, including relationships between the niche depths, statue heights and the geometry of the cliff face. The second layer encodes cultural descriptors such as Gandharan stylistic conventions, stone carving and mud-straw plaster techniques and weathering traces consistent with the material ageing of more than fifteen centuries. The third layer governs technical controls, specifying lighting parameters, composition logic and targeted negative prompts that suppress anachronistic elements such as modern materials or pristine surfaces.

Recent research has emphasised the importance of culturally informed prompting in generative heritage applications, where universalised models often fail to capture regionally specific aesthetics. The DPB addresses this limitation by structuring the prompt as a formal representation of archaeological and stylistic knowledge, rather than a creative instruction (S. Wu et al., 2023).

4.2.3 Prompt Sufficiency Assessment

Prompt sufficiency was assessed at two distinct stages of the HARF workflow. First, a construction-stage completeness check was applied before image generation. Each DPB prompt was reviewed against three domains: structural specification, cultural-iconographic specification and technical-control specification. Each domain was scored out of ten, producing a maximum construction-stage score of 30:

$PSI_{30} = \text{Structural specification} + \text{Cultural-iconographic specification} + \text{Technical-control specification}$

A threshold of 24/30 was adopted as the internal readiness gate for generation. This corresponds to 80 per cent completeness, or 4.0/5 when normalised onto a five-point scale:

$PSI_5 = (PSI_{30} / 30) \times 5$

For example, a prompt scoring 8/10 across each of the three domains receives 24/30 and therefore meets the readiness threshold. Prompts below this threshold required refinement before entering the generation workflow.

Second, prompt sufficiency was assessed by experts in the Phase I questionnaire. This expert-rated sufficiency score measured whether the Bamiyan keyword and prompt structure captured the historical, architectural, cultural and artistic information needed for image generation. It is reported separately in Section 4.5. The construction-stage PSI and the expert-rated sufficiency score therefore operate at different stages: the former controls prompt readiness before generation; the latter evaluates prompt adequacy from a heritage-specialist perspective.

4.2.4 Multimodal Generation Workflow and Platform Evaluation

The generation workflow used DPB-structured prompts to produce still-image candidates for subsequent HARF evaluation. Generative systems were treated as constrained image-production tools within an evidentiary workflow. Prompt construction, paradata recording, computational screening and expert review governed how outputs were produced, selected and refined (Brooks & others, 2023; L. Zhang & Agrawala, 2023).

Initial generation and refinement used the retained tri-platform configuration: Sora v1.2, Gemini Pro v1.5 and Google AI Studio API v2.1. Sora v1.2 was used for high-fidelity full-figure outputs requiring proportional coherence and complete monument framing. Gemini Pro v1.5 and Google AI Studio API v2.1 supported rapid prompt testing, localised refinement and reference-conditioned generation. Each platform therefore had a defined methodological role, reducing dependence on a single model and allowing model-specific distortions to be identified through comparison.

Localised refinement, masking and reference-conditioned generation were retained only where traceability to the prompt record and source constraints could be preserved. These operations were logged as paradata, including platform, model version, prompt structure, available generation settings and refinement rationale (Avrahami & others, 2023; Lugmayr & others, 2022). Quality control prioritised architectural integrity, iconographic consistency and traceability. Outputs were checked against DPB-encoded parameters, construction-stage prompt sufficiency criteria and broad proportional anchors before expert review.

The workflow positioned GenAI as a controlled visualisation instrument within an evidentiary reconstruction process. Visual plausibility was not treated as historical accuracy. Archaeological and art-historical judgement remained necessary for assessing cultural plausibility, iconographic coherence and the limits of reconstruction evidence (Bau & others, 2023; Gal & others, 2023).

4.2.4.1 AI Model Evaluation and Comparative Analysis

Seven text-to-image and multimodal systems were compared before the main generation phase: Midjourney v6 (Don-Yehiya et al., 2023), Stable Diffusion XL (Rombach et al., 2022) DALL-E 3 (Nichol et al., 2022; OpenAI, 2024a; Ramesh et al., 2022b), Gemini Pro v1.5, Google AI Studio API v2.1 (OpenAI, 2024a; Saharia et al., 2022), Meta Imagine (Sun et al., 2023) and OpenAI Sora v1.2 (OpenAI, 2024b). The comparison was informed by wider work on text-to-image generation, diffusion-based image synthesis and multimodal generative systems (Birhane et al., 2023; C. Zhang et al., 2023), but its purpose was narrower: to select a working configuration for the Bamiyan workflow, not to benchmark the platforms. The comparison assessed platform behaviour during DPB prompt development, drawing on the Bamiyan prompt record reproduced in Appendix C.9. That record is an iterative archive of full-figure, regional, viewpoint, reference-conditioned and localised-refinement prompts, not a fixed standardised prompt set.[76], [210][211]

Four qualities guided platform selection: the ability to retain proportional and contextual detail specified in the prompt; stability across iterative refinement; handling of iconographically dense compositions; and generation time, which affected the feasibility of repeated refinement cycles. On this basis, the systems separated into a retained tri-platform configuration and a set of comparative baselines (Table 6). Sora v1.2 was retained for primary full-figure generation because it produced the most usable outputs in relation to proportional coherence, monument framing and niche relation, although its generation time was longer. Gemini Pro v1.5 and Google AI Studio API v2.1 were retained for rapid prompt testing, localised refinement and reference-conditioned generation. DALL-E 3, Midjourney v6, Stable Diffusion XL and Meta Imagine remained comparative baselines, as their outputs showed recurrent issues including stylistic drift, inconsistent geometric stability or weaker site-specific control. Cross-platform checks were then used to identify model-specific distortions and omissions that could affect historical or iconographic plausibility [207], [212](Hou & others, 2023; Münster & others, 2024).

Table 6. Platform-selection matrix for the retained HARF generation workflow. The table summarises platform roles during DPB prompt development and the basis for retaining or excluding each system from the main Bamiyan workflow. It reports workflow-selection judgements, not benchmark scores.

Platform	Gen. time (s)	Main use	Limitation	Workflow decision
OpenAI Sora v1.2	45-60 s	Full-figure coherence	Slow iteration	Primary generation
Gemini Pro v1.5	8-15 s	Prompt responsiveness	Weaker full-figure stability	Refinement

Google AI Studio API v2.1	7-11 s	Localised/API testing	Variable consistency	API refinement
DALL-E 3	15-25 s	Visual coherence	Generic site control	Baseline
Midjourney v6	20-35 s	Visual polish	Aesthetic stylisation	Baseline
Stable Diffusion XL	12-20 s	Open diffusion baseline	Configuration burden	Baseline
Meta Imagine	10-18 s	Visual plausibility	Geometric instability	Baseline

The retained configuration is summarised in Table 7. Platform versions, API configurations and available generation parameters were documented to reduce prompt drift and version instability within the Bamiyan workflow. In the final configuration, Sora v1.2, Gemini Pro v1.5 and Google AI Studio API v2.1 were employed. Prompt construction followed the DPB throughout, using consistent terminology, structured tiers and systematic negative prompting. The full site-agnostic DPB schema is specified in Appendix A, and the Bamiyan prompt record used during platform comparison, generation and refinement is provided in Appendix C.9. Output verification combined broad proportional checks, including head-to-height, shoulder-to-height and base-to-height relations, with expert review of cultural and contextual coherence. Each generated artefact was archived with metadata and paradata recording platform, model version, prompt record, available generation settings and refinement rationale.

Table 7. Summary of model versions and generation parameters used in the HARF workflow. Settings apply to all experiments unless otherwise indicated.

Component	Specification
Platforms	OpenAI Sora v1.2; Gemini Pro v1.5; Google AI Studio API v2.1
Roles	Sora v1.2: full-figure generation; Gemini Pro v1.5 / Google AI Studio API v2.1: rapid testing, refinement and localised edits
Prompt window	Platform maximum
Aspect ratio	16:9
Resolution	Platform maximum HD resolution
Guidance settings	Platform defaults unless exposed controls were available
Random seeds	Fixed where exposed
Prompt standard	DPB: structured tiers; consistent terminology; negative prompting
Documentation	Metadata and paradata: platform, model version, prompt record, settings, refinement rationale

Platform selection, the DPB prompt record and the final Bamiyan image corpus formed successive stages of one workflow. After platform selection, the retained configuration was used to generate the Bamiyan corpus analysed in Chapter 5: 245 still-image outputs, comprising 187 generated in Sora v1.2 and 58 generated through Gemini Pro v1.5 and Google AI Studio API v2.1 workflows. The image-level dataset flow in Chapter 5 provides the authoritative corpus map; Appendix C.9 documents the prompt history behind platform comparison, generation and refinement.

4.3 Generation, Pre-selection and Transparency

The retained HARF configuration generated 245 Bamiyan still-image outputs: 187 in Sora v1.2 and 58 through Gemini Pro v1.5 and Google AI Studio API v2.1 workflows. Outputs were generated at 16:9 aspect ratio and maximum available platform resolution, with random seeds fixed where the platform exposed seed control. Image selection followed a two-stage process. First, computational screening assessed proportional alignment and compositional stability. Second, expert review assessed anatomical plausibility, stylistic coherence and the presence of anachronistic or culturally inconsistent features. Several candidates that performed adequately under automated screening were rejected during expert review, confirming that computational similarity could support filtering but could not replace specialist judgement (Conte & others, 2023).

Post-processing within the AI-generated corpus was limited to framing, resolution standardisation and tonal normalisation where required for comparison. Speculative or missing heritage features were not manually completed within the generated corpus; unresolved gaps triggered prompt revision, localised masking or regeneration, consistent with principles of evidentiary restraint and transparency in digital heritage (Huvila, 2022b; López-Menchero Bendicho & Grande, 2017). Where a manually corrected comparator was later used in the candidate set, it is identified separately in Chapter 5 and Appendix C.8. This distinction preserves traceability between generated outputs, localised refinement steps and manually corrected comparison material.

4.4 Expert Validation, Ethical Integration and Reproducibility

The expert evaluation reported across Chapters 4 and 5 forms one two-phase study, not two separate studies. Phase I was an interdisciplinary expert assessment completed by thirty-three respondents. Phase II was a site-specific specialist evaluation completed by five Bamiyan-focused experts. This chapter reports the Phase I assessment of the DPB, prompt sufficiency and evaluation design. The candidate-level Bamiyan assessment, including image screening, specialist scoring and computational comparison, is reported in Chapter 5.

The evaluation framework followed a sequential two-phase design that combined broad methodological assessment with site-specific expertise. This structure addressed a central challenge in digital heritage evaluation: balancing general technical judgement with culturally and historically grounded specialist knowledge (Agapiou, 2021; Morgan & Eve, 2012). Fifty experts were contacted through academic networks, professional associations and institutional channels; more than 80% received individual briefings outlining research aims, consent procedures and withdrawal rights. Thirty-three experts participated in Phase I, representing geographic, disciplinary and institutional diversity (Figure 17).

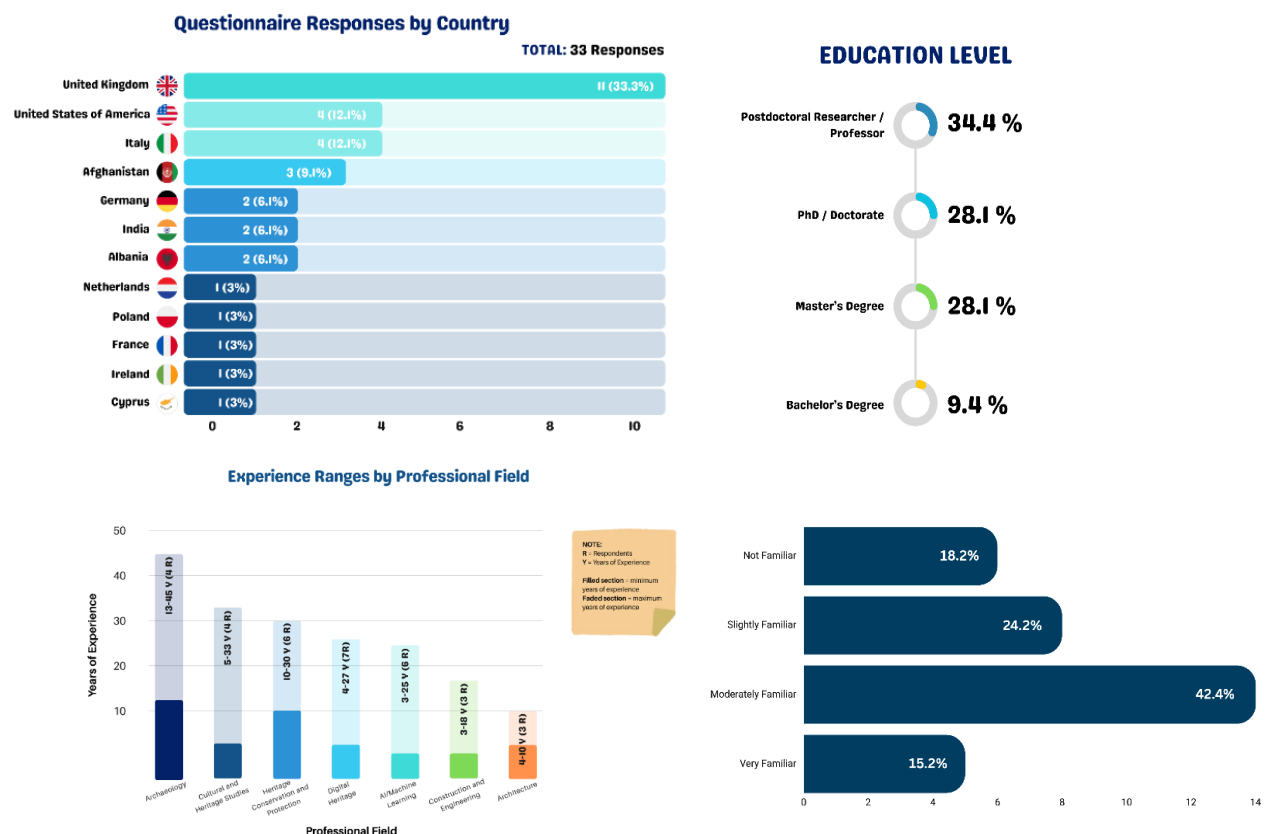


Figure 17. Author-generated demographic profile of the Phase I expert panel ($n = 33$), showing country distribution, educational attainment, years of professional experience and familiarity with the Buddhas of Bamiyan.

Phase I served as the comprehensive assessment stage. Participants completed a structured expert questionnaire comprising 53 items, including Likert-scale ratings, comparative selections and open-ended responses that required approximately 45-60 minutes. The questionnaire operationalises the HARF evaluation domains of structural fidelity, iconographic accuracy, environmental plausibility and prompt sufficiency. The full set of assessment items is provided in Appendix B.

Administrative and consent-related components are not reproduced, as Appendix B reports only the analytical items underpinning the evaluation framework. The questionnaire combined Likert-scale evaluations (photorealism, historical accuracy, proportional fidelity and textural coherence) with

open-ended responses requesting critique, identification of inaccuracies and ethical observations (Dillman et al., 2014; Oppenheim, 2000). Images from three iterative refinement cycles were presented as labelled, side-by-side collages to enable direct comparison of proportional, iconographic and textural evolution. Representative examples of the reconstruction candidates are shown in Figure 18 to illustrate the evaluation material provided to experts. The complete set of evaluated reconstruction outputs and detailed image-level analysis are reported in Chapter 5.[31]

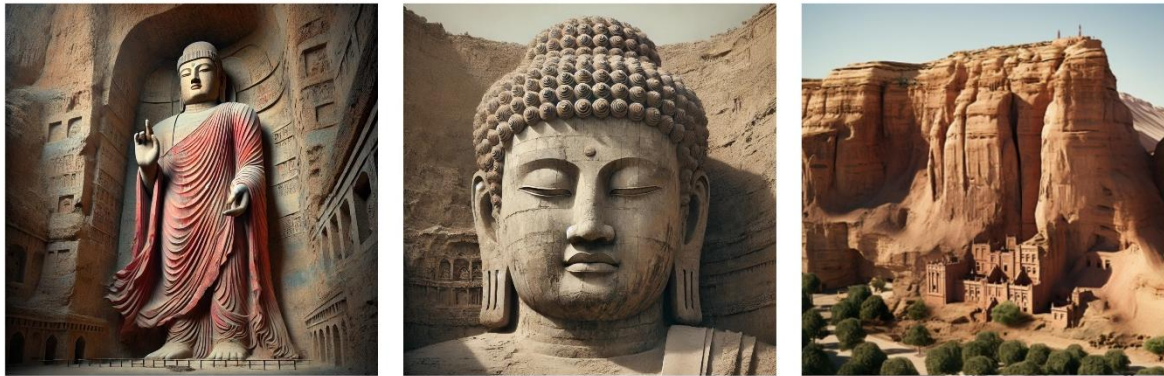


Figure 18. AI-generated reconstruction candidates evaluated in Phase I. (a) structural reconstruction candidate (Image 1 - left), (b) iconographic detail candidate (Image 13 - middle), (c) environmental context candidate (Image 21 - right)

The expert panel represented thirty institutions across Europe, South Asia and North America with disciplinary specialisation distributed across historical (40%), technical (35%) and hybrid (25%) profiles verified through publication and portfolio review. Educational attainment was high: one-third held postdoctoral or professorial appointments, 28% held doctorates and a similar proportion held master's degrees. Professional experience ranged from three to forty-five years. This diversity mitigated the risk of monocultural reasoning and enabled comparison across different interpretive traditions of accuracy and authenticity (Nikolakopoulou & Koutsabasis, 2020).

Phase II provided a focused layer of evaluation by domain specialists with direct involvement in Bamiyan-related conservation or documentation. From the Phase I panel, twelve experts were invited and five completed this stage. Their backgrounds included former surveyors, UNESCO assessors and Buddhist-iconography specialists. Their assessments identified contextual and iconographic misreadings that generalist reviewers and computational metrics were unlikely to detect. This phase prioritised cultural specificity, historical authenticity and evidentiary restraint, reflecting principles widely recognised in heritage assessment (Agapiou, 2021).

Together, the two phases established equilibrium between cross-disciplinary breadth and site-specific authority. Phase I ensured methodological robustness and international representation, while Phase II anchored the evaluation in specialised cultural knowledge essential for assessing accuracy in non-Western heritage contexts. The expert questionnaire is provided in Appendix B to ensure methodological transparency and reproducibility of the evaluation procedure.

4.4.1 Data Collection Tools and Procedures

Data collection employed two complementary instruments aligned with the two phases of expert evaluation. The design prioritised clarity, cross-disciplinary accessibility and methodological transparency, so that quantitative ratings and qualitative comments could inform prompt refinement and candidate assessment.

Phase I used a 53-item questionnaire combining five-point Likert scales with open-ended prompts. Developed with reference to established digital-heritage assessment frameworks (Pietroni, 2021), the instrument was piloted with five colleagues and supervisors to confirm interpretive clarity and estimate completion time. The quantitative component covered six domains: photorealism, historical accuracy, proportional fidelity, material authenticity, environmental context and prompt sufficiency. Participants assessed lighting stability, surface texture, iconographic correctness, dimensional relationships and contextual integration. The qualitative section invited justification for perceived inaccuracies, identification of speculative or missing elements and reflections on ethical constraints. To support interdisciplinary participation, the questionnaire included a concise glossary explaining terminology from Buddhist iconography, archaeology and generative AI. Image sets were presented as labelled, side-by-side collages with consistent framing, resolution and aspect ratio to support comparison across iterative reconstruction stages.

Phase II used a site-specific specialist rubric applied to eighteen refined candidate images. The rubric assessed visible Bamiyan-specific features, including proportional coherence, niche–cliff integration, drapery organisation, drapery colour, hand and foot treatment, head and hair morphology, facial plausibility, surface erosion, landscape context and avoidance of anachronistic features. Criteria were scored as 0, 0.5 or 1, corresponding to absent or incorrect, partially present or ambiguous, and clearly satisfied. Criteria marked N/A were excluded from the denominator for that candidate, so candidate scores were calculated across applicable criteria only. This avoided penalising head studies, drapery studies or localised detail images for features outside the visible scope of the image. The full candidate-level scoring procedure is reported in Chapter 5.

4.4.2 Analytical Workflow and Approach

The analytical workflow integrated quantitative ratings and qualitative commentary from both evaluation phases. Quantitative data were processed in Python (pandas, numpy, scipy) and Excel. Summary statistics, including means, medians, dispersion measures and response distributions, described scoring patterns within each domain and identified where reviewers converged or diverged.

Where relevant, responses were compared descriptively by disciplinary orientation and site familiarity.

Qualitative responses were analysed through a reflexive thematic approach. All comments were reviewed iteratively, with recurrent observations consolidated into higher-level categories to clarify the interpretive assumptions underlying expert judgements. Reflexive annotations documented coding decisions and moments of uncertainty, ensuring transparency in a single-coder context.

Phase II agreement analysis is reported in Chapter 5 using criterion-level inter-rater agreement measures, including quadratic-weighted Cohen's κ and Gwet's AC2 with confidence intervals. Integration of quantitative and qualitative findings was supported by a process-tracing matrix that logged each evaluation cycle, targeted revision and re-assessment, preserving paradata continuity and ensuring that methodological decisions remained explicitly linked to their evidentiary basis.

4.4.3 Ethical Considerations and Data Management

The human-participant components of the research were conducted under Brunel University London research ethics approval, reference 50187-LR-Jan/2025-53624-2. Participants received information about the study purpose, voluntary participation, anonymity, data handling and withdrawal conditions before completing the questionnaire. Responses were processed in anonymised or pseudonymised form, stored securely and handled in accordance with Brunel University research data-management requirements and UK GDPR.

The Bamiyan case required additional ethical restraint because the study concerns a destroyed sacred monument and a site associated with cultural trauma, contested memory and international conservation debate. The reconstruction outputs were therefore treated as research stimuli and evaluation candidates, not as definitive reconstructions. Undocumented iconographic, ceremonial or decorative features were excluded from the prompt structure unless supported by evidence, and speculative elements were either avoided or recorded as interpretive uncertainty. This approach aligned the study with wider ethical expectations in AI-mediated cultural heritage, particularly transparency, cultural sensitivity and accountability (Foka & Griffin, 2024a; OpenAI, 2025).

Data governance covered questionnaire responses, generated images, prompt records, paradata and analytical files. Access was restricted to the research team, and records were organised so that participant data, image-generation provenance and interpretive decisions could be audited separately. Third-party archival and heritage materials were used only within the research and evaluation context, with source attribution and licensing restrictions observed where applicable. The workflow also remained attentive to cultural-property obligations under the 1970 UNESCO

Convention and the 1972 World Heritage Convention by treating outputs as non-commercial research interpretations rather than as authorised reconstructions. Ethical compliance, cultural accountability and procedural transparency were therefore treated as part of the evaluation workflow.

4.4.4 Methodological Limitations, Risk Mitigation and Reproducibility

AI-assisted heritage reconstruction involves unavoidable interpretive and technical uncertainty. Within HARF, these risks were not treated as external limitations after the analysis, but as methodological conditions to be documented during the workflow. Risks were identified through pilot testing, discrepancy logs generated during computational pre-selection and recurring patterns in expert feedback from both evaluation phases. Each risk was mapped to its observable effect within the workflow, such as changes in geometric ranking, regeneration triggers or repeated concerns raised by specialists.

Residual severity was assessed according to three criteria: frequency of occurrence, impact on reconstruction outcomes and cultural or ethical significance. A risk was designated *Critical* where it could compromise cultural or ethical integrity even after mitigation; *High* where it could affect rankings, thresholds or interpretation; *Moderate* where its effects were local and did not alter the main conclusions; and *Low* where its impact was negligible.

Table 8. Summary of methodological risks, their relevance to reproducibility and cultural fidelity, the mitigation strategies applied and the residual severity after mitigation.

Risk	Why it matters	Mitigation implemented	Severity
Platform evolution	Model and version drift can undermine reproducibility.	Version recording; timestamped prompts; saved seeds; metadata and paradata for each generation.	High
Expert fatigue / forced response	A 53-item instrument may affect response quality for less-familiar participants.	Familiarity-based segmentation; “uncertain” response option; monitor completion time monitoring; qualitative cross-checking	Moderate
Computational pre-selection conflicts	Automated scores can mis-rank culturally or iconographically stronger candidates.	Discrepancy logs; expert primacy in final selection; conflicts triggered regeneration or review, not manual fixes.	High
Cultural representation bias	Model outputs may reproduce Westernised or generic visual conventions.	Diverse expert panel; inclusion of Central Asian & Buddhist expertise; negative prompting; explicit review of cultural and iconographic plausibility.	Critical
Temporal-context instability	Sixth- to seventh-century CE visual features may be distorted	Temporal framing; suppression of anachronism; conservative handling	High

	by models trained on contemporary imagery.	of speculative features; documentation of uncertainty.	
--	--	--	--

These mitigations do not remove uncertainty. They make its location and handling visible within the workflow. Version recording, discrepancy logging, expert primacy and paradata capture therefore functioned as safeguards against treating visually coherent outputs as automatically valid reconstructions. Reproducibility in this context means procedural traceability rather than guaranteed regeneration of identical images, since commercial generative platforms remain partly opaque and version-dependent.

4.4.5 Paradata and Metadata Capture

To support traceability within HARF, the study distinguished between metadata recording the technical provenance of each generated output and paradata documenting the interpretive decisions that shaped reconstruction. The structure followed the London Charter (Denard, 2009) and aligned with later work emphasising accountability and transparency in digital-heritage workflows (Bentkowska-Kafel & Denard, 2012; Richards-Rissetto & von Schwerin, 2017).

Metadata recorded the technical parameters available for each generation, including platform and model versions, API settings, seed values where exposed, generation timestamps, prompt record and construction-stage sufficiency status. Each generated output was logged with its computational pre-evaluation measures, including ORB–RANSAC alignment, SSIM, proportional checks and region-of-interest masks. This provided a technical record against which candidate selection and refinement decisions could be audited.

Paradata documented the interpretive pathway that guided iterative refinement. This included expert rationales, explanations for accepting or rejecting images, prompt-revision notes, iteration-change logs and records of cases where computational indicators conflicted with expert assessment. These records clarified how evidentiary constraints, cultural expertise and computational indicators were weighed when resolving uncertainty.

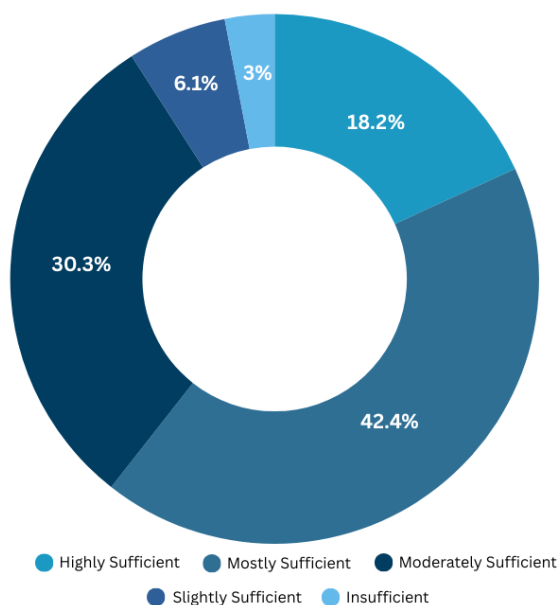
4.5 Results

The evaluation produced three principal findings on the operational performance of HARF. First, expert feedback showed that prompt structure affected proportional, iconographic and textural behaviour across reconstruction cycles. Second, the construction-stage completeness check

distinguished prompts that required further specification before generation, while the expert-rated sufficiency score identified where the prompt and keyword structure still needed refinement. Third, written feedback identified recurrent issues in phrasing, negative constraints and the use of structural anchors. The findings are presented through the numerical patterns evident in the ratings and the explanatory comments provided by participants.

4.5.1 Expert-Evaluation of the DPB and Prompt Structure

Expert evaluation provided the first indication of how effectively the baseline prompt schema captured the descriptive information needed for historically grounded reconstruction. The structured questionnaire combined Likert-scale sufficiency ratings with open-ended comments, following established principles for questionnaire design and expert-response elicitation (Dillman et al., 2014; Oppenheim, 2000). Twenty of the thirty-three respondents rated the prompt structure as Highly or Mostly sufficient, while ten rated it Moderate, two Slightly sufficient and one Insufficient. On a five-point scale, this produced an expert-rated prompt sufficiency mean of 3.67/5 (Figure 19). This finding suggests that the prompt structure was adequate as a working schema, but not yet sufficiently detailed for final generation without refinement. Experts identified the need for stronger environmental description, material treatment and symbolic/contextual detail [220]



Qualitative comments clarified where these refinements were most necessary. Participants prioritised strengthening environmental and geographical context, proportional geometry, architectural dimensions, material descriptions and stylistic features (Figure 20). These were the area's most vulnerable to omission during prompt construction. Experts also favoured grouping descriptive elements into structured clusters to minimise generative drift across iterative outputs.

Figure 19. Author-generated chart showing expert ratings of baseline prompt sufficiency (n = 33).

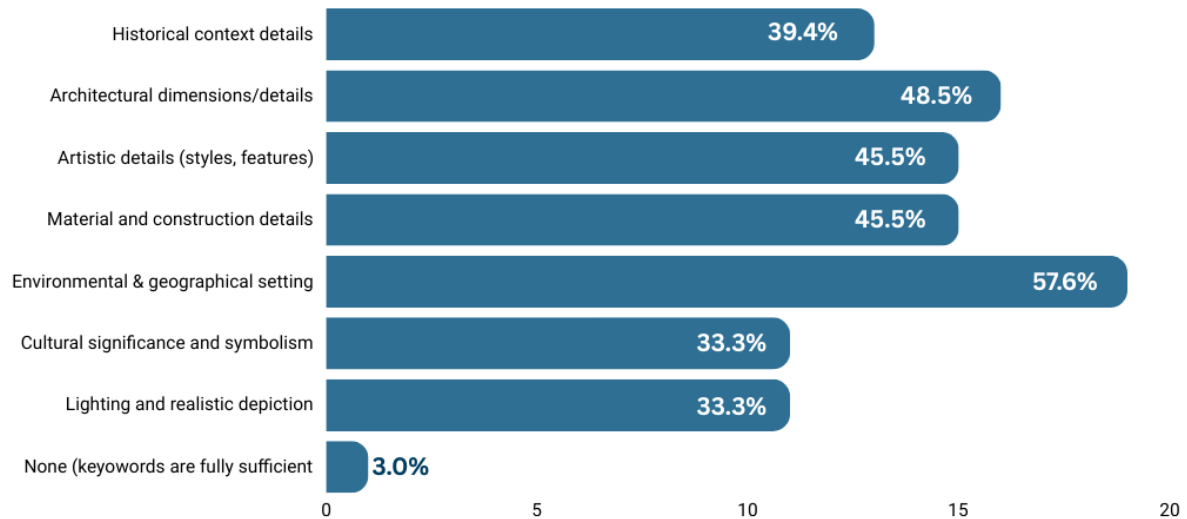


Figure 20. Author-generated chart showing domains prioritised by experts for prompt enrichment ($n = 33$). The distribution visualises which categories were most frequently prioritised for refinement in AI-guided heritage visualisation.

The six refinement themes identified in open-ended responses centred on anatomical and stylistic specificity, environmental anchoring, temporal clarity, terminological precision, spatial depth cues and clearer organisation of prompt information (Table 9). These themes indicate where ambiguity or omission was most likely to occur in the baseline prompt design.

Table 9. Thematic clusters of expert-suggested prompt refinements, with representative quotations illustrating specific areas for improvement. n = number of experts out of 33 who identified that theme during the qualitative analysis.

Theme	n	Example quotation
Anatomical and stylistic specificity	12	"Include reference to anatomy, posture and positioning to maintain iconographic accuracy."
Environmental context and morphology	11	"Prompts should emphasise the tight cave niche and cliff stratigraphy surrounding the Buddha."
Historical and temporal anchoring	9	"Add explicit temporal framing, e.g., 6th–7th century construction or pre-destruction state."
Terminological accuracy	7	"Replace vague terms with correct archaeological vocabulary, e.g., 'polychromy' not 'multicolour.'"
Spatial depth and proportionality	5	"Current measures lack depth; add 3D proportional cues to improve spatial coherence."
Structured prompt refinement	4	"Group keywords into progressive refinement steps to reduce drift and support paradata."

This feedback informed the formalisation of the DPB as a slot-based schema rather than a linear text template. Each slot functions as a semantic tier that can be populated or refined independently, enabling structured control over descriptive density and evidential grounding. The schema comprises three interlinked tier families: core structural anchors, which record non-negotiable evidential

elements such as monument identity, proportional geometry and verified materials; auxiliary contextual descriptors, which capture environmental settings, stylistic lineage and iconographic meaning; and technical control parameters, which govern procedural transparency through negative prompting, dimensional tolerances and metadata for documenting interpretive rationale. Reviewers indicated that this tiered structure improved procedural consistency, since descriptive and ethical decisions were logged in defined fields rather than dispersed across narrative prompt text.

Experts assessed sufficiency according to the completeness and clarity of slot coverage. A prompt was treated as sufficient when its core anchors were specified, contextual descriptors were grounded in verifiable evidence and technical controls recorded the reasoning behind inclusions and exclusions. The expert-rated sufficiency results and written comments identified which components required refinement, rather than treating sufficiency as proof of historical accuracy.

For the Bamiyan case study, archival and archaeological sources were mapped onto the tier schema: anchors captured dimensional geometry and material traces; auxiliary descriptors encoded environmental and stylistic context; and technical controls ensured that proportional constraints and paradata were explicitly recorded. This mapping preserved a consistent link between source evidence and generative output while keeping procedural choices visible.

Overall, the expert evaluation and the slot-based architecture demonstrate that the DPB's modular tier structure provided a stable foundation for evidential mapping and refinement. The feedback not only assessed prompt sufficiency but also clarified how architectural formalisation supports transparency, transferability and reproducibility across heritage reconstruction scenarios.

4.5.2 Quantitative PSI and Refinement Outcomes

Quantitative prompt-sufficiency analysis was used in two distinct ways. First, the construction-stage PSI checked whether a DPB prompt reached the minimum readiness threshold before generation. As defined in Section 4.2.3, this score was calculated across three equally weighted domains: structural specification, cultural-iconographic specification and technical-control specification. Each domain was scored out of ten, producing a maximum score of 30:

$$PSI_{30} = S + C + T$$

where S = structural specification, C = cultural-iconographic specification and T = technical-control specification.

The five-point equivalent was calculated as:

$$PSI_5 = (PSI_{30} \div 30) \times 5$$

A prompt therefore met the readiness threshold when $PSI_{30} \geq 24$, equivalent to $PSI_5 \geq 4.0$. For example, a prompt scoring 8/10 in structural specification, 8/10 in cultural-iconographic specification and 8/10 in technical-control specification receives 24/30 and therefore meets the readiness threshold. This construction-stage PSI was used to prevent under-specified prompts from entering the generation workflow.

Second, the Phase I questionnaire provided an expert-rated sufficiency judgement on the prompt and keyword structure. This produced a mean score of 3.67/5 across thirty-three respondents. The two values should not be conflated. The construction-stage PSI recorded whether prompts were sufficiently complete to support controlled generation, while the expert-rated sufficiency score identified where those prompts still required refinement from a heritage-specialist perspective.

The refinement process therefore did not treat PSI as proof of historical accuracy. It used PSI to identify construction-stage omissions and used expert feedback to refine the vocabulary, hierarchy and evidential specificity of the prompt structure.

Expert feedback also identified individual terms that required greater precision. Early prompts contained generic descriptors such as flowing robe or rocky cliff, which reviewers considered too broad for reliable interpretation. These expressions were therefore replaced with archaeologically grounded terminology, such as stepped Gandhāran pleats consistent with wet-fold drapery, and stratified sandstone with monastic cells. This refinement strengthened the clarity of key descriptive elements and reduced ambiguity in categories associated with lower expert-rated sufficiency.

4.5.3 Cross-Site Evaluation

The DPB was designed to remain structurally stable while allowing evidential descriptors to be re-specified for different heritage contexts. In the Bamiyan case, the schema was populated with descriptors for a figural, cliff-carved Buddhist monument. For the cross-site transferability test, the same tier structure was re-instantiated for the Temple of Bel at Palmyra, replacing Bamiyan-specific descriptors with architectural, material and decorative evidence appropriate to a monumental temple context. The purpose was not to show that identical visual problems recur across sites, but to examine whether the DPB could organise site-specific evidence while preserving a consistent evaluative logic.

Table 10. Cross-Site Substitution of DPB Structural Tiers

Category	Bamiyan Instantiation	Cross-Site Substitution
Monument	Western Buddha of Bamiyan	Temple of Bel, Palmyra
Typology	Figural cliff-carved sculpture	Monumental temple architecture

Material	Mud-straw plaster, sandstone, lime plaster	Limestone, ashlar masonry, carved relief surfaces
Structural context	Tall niche, pedestal, cliff face	Cella, podium, columns, sanctuary enclosure
Visual grammar	Gandhāran robe folds, hand gesture, cranial morphology	Palmyrene relief, column capitals, wall and roofline articulation
Environment	Bamiyan cliff, monastic caves, valley setting	Palmyra oasis, temple precinct, colonnaded urban context
Instability domains	Hands, drapery, hair, facial morphology, niche depth	Roofline, relief detail, column proportion, inscriptions, surface texture

This re-instantiation preserves the DPB’s evaluative structure while changing the site-specific evidential content. The transfer claim is therefore deliberately bounded. Bamiyan tests the framework against figural and iconographic instability; Palmyra tests whether the same structure can organise architectural and decorative instability. Chapter 7 reports the Palmyra transferability test in full.

4.5.4 Evaluation of Ethical and Transparency Measures

Expert feedback also addressed the framework’s transparency mechanisms, particularly the extent to which evidential sources, interpretive decisions and areas of uncertainty were made explicit. Reviewers indicated that the DPB’s structured recording of evidential anchors, contextual assumptions and negative prompts improved the traceability of reconstruction decisions. This was consistent with the documentation principles of the London Charter (Denard, 2009) and the Seville Principles (López-Mencheró Bendicho & Grande, 2011).

Three aspects of the workflow supported ethical clarity:

1. **Controlled exclusion:** Undocumented features, including unverified decorative elements, haloes, unrecorded wall paintings and uncertain façade symmetries, were excluded from prompt formulation unless supported by evidence. Negative prompting was used to reduce speculative inflation and to limit culturally inconsistent visual additions.
2. **Documentation of inference:** Interpretive steps are explicitly annotated with their rationale, comparative reasoning and confidence levels. This documentation allowed verifiable evidence to be distinguished from scholarly interpretation within the prompt structure, aligning with established expectations for paradata documentation (Bentkowska-Kafel & Denard, 2012)
3. **Transparent output metadata:** Each completed reconstruction candidate was accompanied by metadata outlining its evidential scope, temporal framing and uncertainty status. This

supported scholarly interpretation and helped communicate limitations to non-expert audiences (Richards-Rissetto & von Schwerin, 2017).

Across these areas, expert feedback suggested that the DPB's transparency protocols provided a practical means of maintaining cultural accountability and interpretive clarity. The evaluation did not treat transparency as evidence of historical accuracy; rather, it showed that HARF made the evidential limits, exclusions and interpretive decisions behind generated outputs more visible.

4.5.5 Novelty, Contribution and Scholarly Significance

The DPB represents a methodological contribution to AI-assisted heritage reconstruction by formalising prompt design as an evidence-driven process rather than an intuitive or platform-dependent practice. Its novelty arises from the consolidation of several elements typically treated separately in digital-heritage workflows. The first contribution is the incorporation of verifiable archaeological, architectural and iconographic data directly into the prompt structure. This ensures that descriptive clauses reflect evidential constraints rather than stylistic assumptions. The second is the use of a tiered semantic architecture, which organises material, spatial, environmental and symbolic information into discrete slots. This structure provides a stable grammar for prompt assembly and prevents descriptive drift across iterative generations.

A further contribution is the introduction of a construction-stage prompt-completeness mechanism. The PSI enables descriptive completeness to be assessed systematically before generation, allowing under-specified domains to be identified and refined. It does not validate historical accuracy. Expert review remains necessary to assess archaeological plausibility, art-historical consistency and cultural appropriateness. The final contribution concerns ethical clarity: the DPB embeds procedures for identifying uncertainty, recording interpretive rationale and excluding undocumented or speculative features, consistent with established digital-heritage principles of transparency and paradata (Denard, 2009; López-Menchero Bendicho & Grande, 2011), and with current concerns about bias, accountability and ethical responsibility in AI-mediated cultural heritage (Foka & Griffin, 2024a; S. Ocón et al., 2023).

These mechanisms reposition the prompt as a scholarly instrument linking empirical evidence and computational generation. The DPB provides a means of producing visualisation candidates whose descriptive content, interpretive reasoning and procedural constraints are explicitly documented and therefore open to review. The framework is not universal in an unrestricted sense. Its tier-based structure allows evidential descriptors to be re-specified for other heritage contexts while preserving the underlying evaluation logic as tested in bounded form through the Palmyra case.

By demonstrating how prompts can be transformed from ad hoc cues into reproducible analytical instruments, the DPB advances digital-heritage practice towards reconstruction workflows that are culturally sensitive, archaeologically defensible and procedurally transparent. It provides a model through which generative systems can support, rather than displace, responsible heritage evaluation.

4.6 Limitations and Future Work

The framework presented here demonstrates that prompt design for heritage reconstruction can be made evidential, auditable and procedurally traceable. However, several limitations qualify its current scope and must be made explicit in line with the transparency principles of the London Charter and the Seville Principles (Denard, 2009; López-Mencheró Bendicho & Grande, 2011, 2017).

First, the reproducibility of generative outputs remains partly dependent on the volatility of commercial models. Version updates, opaque changes to training data and non-deterministic sampling mean that prompt-level traceability does not guarantee identical results across time or across platforms, even where versions and available metadata are recorded. HARF mitigates this instability through the DPB, the construction-stage completeness check, expert-rated prompt sufficiency and paradata, but complete temporal reproducibility is not currently achievable within commercial GenAI systems.

Second, the evidentiary record also limits the framework's precision. Where documentation is fragmentary, incomplete, or contested, the DPB cannot eliminate uncertainty. In such cases, prompt sufficiency reflects the completeness of the schema in relation to available evidence, not the completeness of the historical record itself. This is particularly evident for anatomical and facial regions of the Western Buddha that had deteriorated before 2001, where the absence of stable visual anchors constrains what can be responsibly specified.

Third, training-data bias continues to affect representational accuracy. Generative systems trained on large-scale and uneven visual datasets may reproduce stylistic conventions that conflict with regional sculptural grammars, especially in non-Western contexts. These biases can appear as symmetrical drapery, cinematic surface treatment or hybridised morphological features. Constraint-based prompting and expert evaluation reduce these risks, but they cannot fully compensate for structural imbalances in the training data.

Fourth, the expert-evaluation component has limits of scope. Although the validation process included disciplinary breadth and methodological diversity, the expert panel did not by itself represent all local, descendant or religious perspectives connected to Bamiyan. Those perspectives are addressed

separately through the community-grounded authenticity study in Chapter 6, but they remain methodologically distinct from the expert evaluation reported in this chapter.

Finally, this implementation focuses on two-dimensional still imagery. Extending HARF to volumetric or temporal reconstructions, including 3D models, animations or immersive environments, would introduce separate questions of spatial coherence, lighting stability, uncertainty communication and geometric calibration. Those extensions require dedicated validation procedures and are not evaluated in this chapter.

Future work should refine prompt-sufficiency scoring across different heritage typologies and examine how HARF can be linked to photogrammetric, laser-scanning or survey-grade anchors where such data exist, without treating GenAI output as equivalent to measured reconstruction. Collaboration between heritage institutions and AI developers will be necessary for building more balanced training corpora and reducing stylistic drift. FAIR-compliant repositories for prompts, model versions, generated outputs and paradata would strengthen procedural reproducibility as platforms evolve. Further development should also address how uncertainty can be communicated through layered visualisation, alternative reconstruction scenarios and interfaces that distinguish evidence from interpretation.

4.7 Conclusion

This chapter has formalised HARF as the evaluation framework of the thesis. It has defined the DPB as the prompt-construction schema, the PSI as a construction-stage sufficiency gate, the paradata protocol as the record of interpretive and technical decisions, and computational pre-evaluation as a two-dimensional screening step before expert review. The chapter has also clarified the role of platform selection within the workflow. The comparison of Sora v1.2, Gemini Pro v1.5, Google AI Studio API v2.1 and baseline systems informed the retained configuration, but it did not constitute a fixed prompt benchmark.

The Phase I expert evaluation showed that the prompt structure was sufficiently developed to operate as a working schema, but required refinement in environmental description, material specification, proportional anchoring and cultural terminology. These findings supported the conversion of prompt writing into a more controlled DPB procedure and provided the methodological basis for the refined Bamiyan corpus analysed in Chapter 5.

The contribution of this chapter is therefore methodological. It does not claim that the generated images are historically accurate reconstructions. It establishes the procedure through which such images become assessable: evidence is translated into structured prompts, prompt completeness is

checked before generation, platform and refinement decisions are recorded as paradata, computational screening identifies visible structural drift, and expert review evaluates historical and cultural plausibility.

The limitations identified in this chapter also define the scope of HARF. Commercial GenAI systems remain unstable across time and platform versions; the surviving Bamiyan record remains incomplete; training-data bias continues to affect visual outputs; expert evaluation cannot substitute for community-grounded authenticity; and the present implementation remains limited to two-dimensional still imagery. These limits do not undermine the framework. They specify the conditions under which its claims should be read.

Chapter 5 moves from framework development to candidate-level empirical assessment. It applies HARF to the Bamiyan reconstruction corpus and examines how computational similarity, interdisciplinary screening and specialist archaeological judgement evaluate the same set of generated outputs.

Chapter 5

Computational and Expert Evaluation at Bamiyan

Chapter 4 formalised HARF as the evaluation framework for GenAI heritage reconstruction. This chapter tests the framework at the Western Buddha of Bamiyan by applying it to a corpus of generated reconstruction candidates and comparing computational similarity, interdisciplinary screening and specialist archaeological judgement. The purpose is not to identify a definitive reconstruction. It is to examine how different evaluative registers behave when applied to the same evidentially constrained reconstruction problem.

This chapter provides the computational and specialist evaluation of HARF at candidate level. Chapters 4 and 5 report different analyses of a single two-phase expert study: an interdisciplinary screening phase (Phase I, $n = 33$) and a site-specific specialist phase (Phase II, $n = 5$). Chapter 4 establishes the framework; this chapter reports the empirical evaluation of the Bamiyan corpus across both phases. It asks where algorithmic measures of structural correspondence support expert judgement, where they diverge, and which visual domains remain unstable despite prompt refinement. The analysis remains confined to specialist evaluation; community-grounded authenticity is addressed separately in Chapter 6.

5.1 Introduction

Generative models are increasingly used in cultural heritage reconstruction to visualise monuments that are destroyed, inaccessible, or only partially documented. Artificial intelligence techniques are increasingly used in heritage digitisation, restoration and interpretation, enabling new forms of digital preservation and analysis of historical artefacts and monuments (Li et al., 2025; Spennemann, 2024; Westenberger & Farmaki, 2025). Despite these technological advances, however, generative systems remain methodologically difficult to evaluate when applied to historically specific monuments. Reconstructions produced by large-scale AI image models frequently reflect aesthetic plausibility rather than evidentiary fidelity, raising questions about how such outputs should be assessed within archaeological and heritage research (Wei et al., 2026; Xu et al., 2025).

This challenge is particularly visible in cases where surviving documentation is fragmentary and interpretive uncertainty is unavoidable. The destruction of the sixth- to seventh-century CE Buddhas of Bamiyan in 2001 represents one of the most prominent examples. Although the statues were extensively documented in the twentieth century, the surviving record remains incomplete. Reconstruction must therefore navigate the tension between visual plausibility and archaeological accountability (Flood, 2002).

Recent scholarship emphasises the need for transparent methodological frameworks integrating computational techniques with documentary evidence and expert interpretation with particular emphasis on reproducibility and expert oversight in AI-assisted heritage analysis (Colace et al., 2025; Paolanti et al., 2026). These principles extend earlier guidelines such as the London Charter and the Seville Principles, which emphasise the need to distinguish clearly between evidence-based reconstruction and interpretive conjecture in digital heritage visualisation (Denard, 2009; López-Mencheró Bendicho & Grande, 2011).

Within this context, the Heritage-Aligned Reconstruction Framework (HARF) was introduced as a methodology integrating structured prompting, quantitative filtering and expert-in-the-loop validation within AI-mediated reconstruction workflows (Arzomand et al., 2026). While this framework establishes a methodological architecture for reconstruction, its behaviour in complex heritage contexts remains to be examined. Generative models may exhibit unpredictable behaviour when operating within culturally specific visual grammars, particularly when training data insufficiently represent non-Western monuments.

The study does not seek to reproduce the monument with historical certainty, but to test whether a reproducible, multi-stage evaluation process can distinguish between formally plausible and archaeologically unsupported reconstructions and to define the role of expert validation in this process.

5.2 Methodology

5.2.1 Corpus Generation and Iterative Refinement

The empirical evaluation proceeded through four stages: corpus formation, algorithmic pre-evaluation, interdisciplinary screening and specialist evaluation. Three datasets were distinguished within the reconstruction workflow: the full generation corpus, the interdisciplinary screening dataset and the specialist evaluation dataset. Computational pre-evaluation was applied only to the eleven

full-figure candidates, as head-only and drapery studies lacked sufficient statue–niche context for reliable geometric comparison.

5.2.2 Implementation Parameters and Iterative Cycles

The reconstruction corpus comprised 245 still-image outputs, of which 187 were generated in OpenAI Sora v1.2 and 58 through Gemini Pro v1.5 via the Google AI Studio API v2.1. All generations followed the Dynamic Prompt Blueprint (DPB) schema and Prompt Sufficiency Index (PSI) criteria defined within HARF. In this study, “DPB-compliant” refers to prompts satisfying DPB slot-completion and PSI sufficiency requirements, including explicit dimensional anchors. Platform defaults were retained where applicable, fixed seeds were used when available, aspect ratio was standardised to 16:9 and all outputs were archived with metadata and paradata.

The corpus comprised two iteration categories: exploratory and controlled generations. Exploratory outputs tested prompt stability across platforms and revealed recurring distortion. Pre-screening was applied to the controlled subset to remove outputs exhibiting explicit anachronism, geometric instability, or recurrent artefacts identified during exploratory generation. This stage combined quantitative proportion checks (head:height, shoulder:height, base:height ratios) with visual inspection. The resulting shortlist formed the interdisciplinary screening set.

From this shortlist, 23 images (Image 1–23) were selected for Phase I screening, enabling broader assessment of architectural plausibility, stylistic coherence and environmental context. Following refinement cycles informed by Phase I feedback and the exclusion of recurrent artefacts, 18 final-generation images were retained for Phase II specialist evaluation. The staged evaluation workflow is summarised in Figure 21.

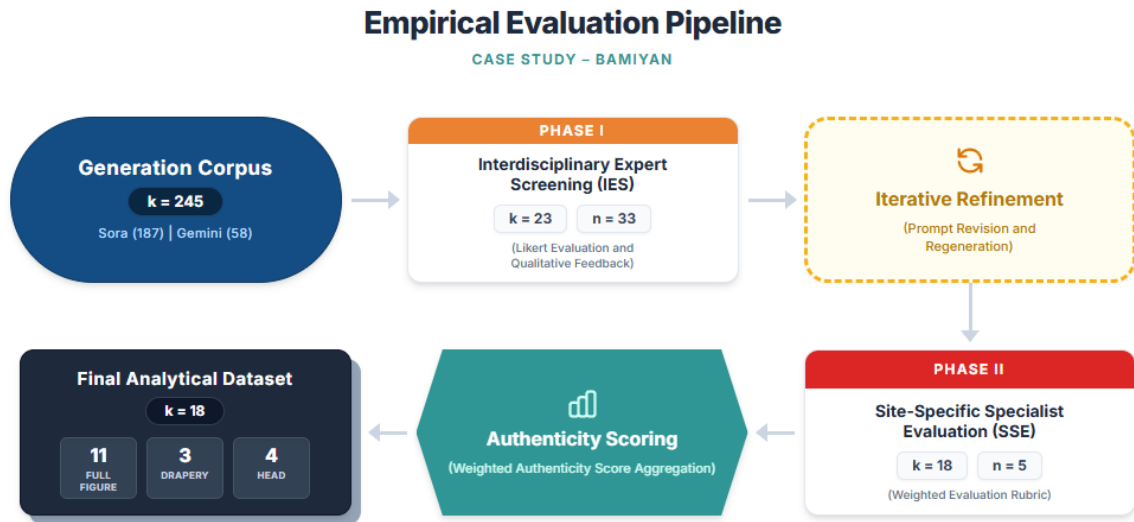


Figure 21. Author-generated workflow diagram showing the empirical evaluation pipeline for the Bamiyan reconstruction case study. The workflow summarises the staged evaluation process applied to the Western Buddha reconstruction corpus. An initial generation corpus ($k = 245$ images; Sora = 187, Gemini = 58) underwent interdisciplinary expert screening ($k = 23$ images; $n = 33$ experts) followed by iterative prompt refinement and regeneration. The refined candidate set ($k = 18$ images) was subsequently evaluated in Phase II specialist assessment ($n = 5$) using a structured applicable-criteria rubric, producing the final analytical dataset comprising 11 full-figure reconstructions, 3 drapery studies and 4 head studies. Note: k denotes the number of images; n denotes the number of experts.

Figure 21 shows the staged evaluation pipeline. Table 11 clarifies the analytical status of each dataset used in the Bamiyan evaluation and distinguishes corpus formation, expert screening, specialist scoring and computational comparison.

Table 11. Formation and use of the Bamiyan evaluation datasets

Stage	Dataset	Size	Function	Evaluation use
Generation corpus	AI outputs	245	Platform behaviour and iterative refinement	Corpus formation
Phase I	IES images	23	Interdisciplinary screening across structural, iconographic and environmental domains	Expert screening
Phase II	SSE candidates	18	Specialist archaeological assessment	Specialist scoring
Computational subset	Full-figure candidates	11	Structural similarity and proportional comparison	Computational evaluation
Visual reference	Manual composite	1	Visual comparison only	Excluded from scoring

Image numbering in this chapter is stage-specific. Phase I screening images (Images 1–23) refer to the interdisciplinary questionnaire set and are organised into structural (Images 1–7), iconographic

(Images 8–17) and environmental (Images 18–23) subsets. The Phase II specialist-evaluation candidates were derived from this screening set and are numbered and reported separately.

The final 18 images comprised 11 full-figure reconstructions, 3 drapery studies and 4 head studies, enabling separate assessment of proportional integrity, drapery articulation and cranial morphology. Iterative refinement proceeded through prompt revision and controlled regeneration, but the workflow differed across platforms. Gemini Pro refinement occurred recursively within a single conversational session, allowing constraints to accumulate across iterations. In contrast, Sora generation used independent prompt instances requiring re-specification of the DPB structure. Failure criteria included anatomical impossibilities, iconographic inconsistencies, non-Bamiyan architectural insertions, geometric proportion drift and stylistic contamination. A representative expert-refined prompt is provided in the Appendix C.

Distortions such as asymmetrical drapery, exaggerated cranial curvature and speculative architectural insertions were corrected by modifying DPB slots, refining negative prompts, or adjusting dimensional anchors prior to regeneration. Selected outputs were composited using platform-native image-to-image workflows to merge proportionally stable segments; no manual pixel-level alterations were introduced into the scored AI-generated candidates. Image 12 was a separately labelled Photoshop-corrected comparator and was excluded from computational evaluation and expert scoring. Region-based masking constrained regeneration to specific anatomical zones (e.g. hands, hair morphology, drapery segments) without altering global composition. All such merges were logged in paradata and treated as new iterations rather than post-production edits.

Across iterations, recurrent instability patterns included halo-like insertions, over-symmetrised drapery, excessive surface smoothing, cranial height exaggeration, bun-silhouette distortion and speculative cave insertions in background stratigraphy.

Platform behaviour also differed across iterations. Gemini enabled rapid prompt refinement but exhibited greater proportional variance, whereas Sora tended to produce stronger geometric stability alongside occasional surface-level stylisation drift.

Documented architectural updates and safety-driven deployment safeguards indicate that model behaviour is subject to version-specific constraints and alignment adjustments (Gemini Team et al., 2024; OpenAI, 2024a, 2025). Platform outputs were therefore treated as temporally bounded results and prompts, timestamps, model identifiers and API configuration states were systematically recorded as paradata to preserve interpretability and conditional replicability. Exact pixel-level reproducibility across subsequent model updates cannot be guaranteed.

5.2.3 Computational Structural Evaluation

A computational evaluation pipeline was implemented to measure geometric correspondence between generated reconstructions and the archival record of the Western Buddha of Bamiyan. This stage was intended to provide a reproducible structural baseline for subsequent specialist evaluation rather than to replace expert judgement.

(i) Dataset and Reference Image

Computational structural evaluation was applied to the 11 full-figure reconstruction candidates assessed during specialist evaluation. Head-only and drapery-focused studies were excluded because they lacked sufficient statue–niche context for reliable geometric comparison. Each candidate was compared against a single archival photograph of the Western Buddha of Bamiyan prior to destruction (Figure 24), which served as the geometric reference for all computational comparisons.

(ii) Region of Interest (ROI) Definition

Two rectangular regions of interest (ROIs) were defined for spatially constrained comparison between generated images and the archival reference. The first captured the statue from head to feet and the second the surrounding niche cavity. This enabled separate measurement of statue morphology and niche geometry relative to the archival reference (Figure 22).

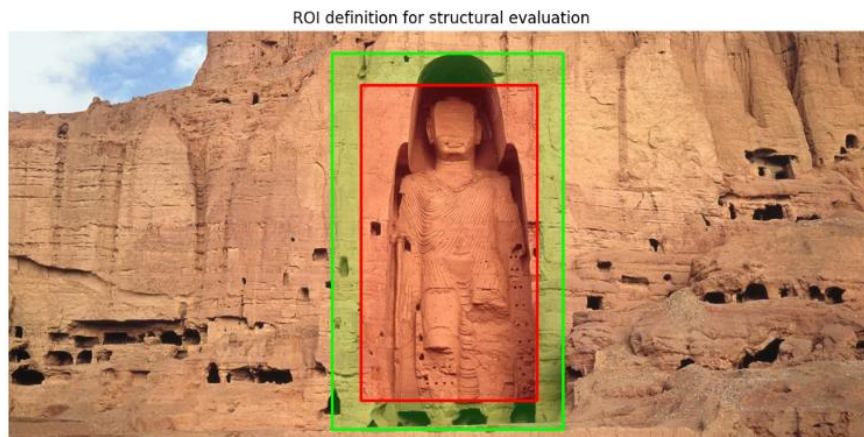


Figure 22. Author-defined ROI annotation used for computational structural evaluation. The statue ROI (red) and niche ROI (green) enable independent measurement of morphological similarity and architectural alignment relative to the archival reference.

(iii) Geometric Registration

Generated images were aligned to the archival reference using Oriented FAST and Rotated BRIEF (ORB) keypoint detection with Random Sample Consensus (RANSAC) homography estimation (Fischler & Bolles, 1981; Rublee et al., 2011). This reduced variation due to scale and viewpoint and enabled pixel-level comparison within defined statue and niche ROIs.

ORB–RANSAC identified gross geometric misalignments and proportion drift, enabling consistent spatial alignment prior to descriptor extraction. Registration was performed in two dimensions: planar homography aligned generated candidates to the archival reference image to support structural comparison and screening before expert review. It did not reconstruct three-dimensional geometry, recover depth or volumetric form, or assess cultural, stylistic or iconographic plausibility; these remained matters for specialist evaluation.

Figure 23 illustrates a representative alignment result, showing the archival reference image, a generated reconstruction prior to registration and the overlay produced after geometric alignment.

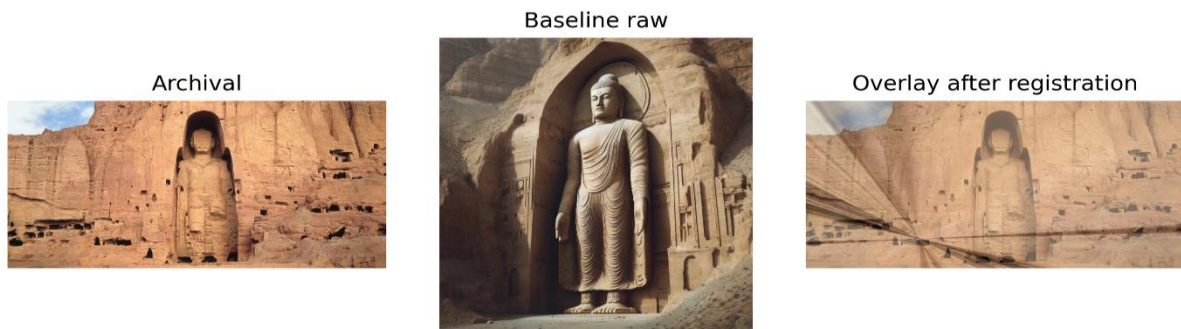


Figure 23. Representative geometric registration to the archival reference. The figure combines the archival photograph (left), an AI-generated reconstruction candidate prior to registration (centre), and an author-generated registration overlay (right). The overlay illustrates the two-dimensional alignment procedure applied before structural similarity measurement.

(iv) Descriptor Extraction

Two descriptors were selected to capture complementary, reproducible dimensions of two-dimensional structural fidelity. SSIM measured luminance-based structural similarity within bounded statue and niche regions, while proportional similarity measured gross statue–niche geometry. These descriptors were used as transparent, region-constrained measures for pre-expert screening, not as measures of historical, cultural or stylistic accuracy.

- **Structural Similarity Index (SSIM):** computed separately for the statue region and the niche region using the luminance channel, providing a measure of structural similarity between generated images and the archival reference (Zhou Wang et al., 2004).
- **Proportional Similarity:** measured the statue height-to-width ratio relative to the niche cavity dimensions and compared it with the archival reference proportions.

(v) Composite Scoring

ROI-level SSIM values were computed after geometric registration and aggregated with proportional similarity to produce a composite structural similarity score. Equal weighting was adopted as a neutral aggregation rule in the absence of a validated domain-specific weighting scheme.

$$CompositeScore = \frac{SSIM_{statue} + SSIM_{niche} + PropSimilarity}{3}$$

CompositeScore $\in [0,1]$, with higher values indicating stronger structural correspondence with the archival reference. $SSIM_{statue}$ and $SSIM_{niche}$ represent structural similarity measured within the statue and niche regions and $PropSimilarity$ denotes proportional agreement between the generated statue and the archival reference.

5.2.4 Expert Evaluation Design

Thirty-three experts completed Phase I — Interdisciplinary Expert Screening (IES), while five Bamiyan specialists completed Phase II — Site-Specific Specialist Evaluation (SSE). IES provided breadth of interdisciplinary judgement, whereas SSE offered depth of site-specific specialist assessment.

The following notation is used throughout the evaluation sections: **n** denotes the number of participants (experts), **k** the number of reconstruction images evaluated and **c** the number of evaluation criteria. Phase I — Interdisciplinary Expert Screening (IES) involved experts from heritage studies, digital heritage research, civil and architectural engineering and AI-related computational fields. During IES ($n = 33$), cross-disciplinary experts evaluated the screening set through Likert ratings and two structured preference instruments:

- (i) a multi-select endorsement task applied to the structural subset ($k = 7$; Images 1–7) and
- (ii) a forced-choice comparison between the initial baseline reconstruction, the refined reconstruction and the option “neither accurately represents historical reality.”

Feedback obtained during IES informed subsequent prompt refinement and image regeneration prior to the specialist evaluation stage.

Phase II — Site-Specific Specialist Evaluation (SSE) involved experts with direct knowledge of the Bamiyan archaeological context. During SSE ($n = 5$), a structured scoring rubric was applied to $k = 18$ refined reconstructions presented in domain-specific subsets (architecture/structure, iconographic detail and environmental context). Each reconstruction was evaluated across $c = 10$ criteria covering architectural configuration, iconographic detail and environmental plausibility.

SSE employed a constrained ordinal rubric anchored to explicit archaeological criteria. Specialists were instructed to score each applicable criterion using a three-point scale: 0 = absent or incorrect, 0.5 = partially correct or ambiguous and 1 = correct or clearly satisfied.

Mean criterion scores were calculated as:

$$\bar{s}_{i,c} = \frac{1}{n} \sum_{r=1}^n s_{i,c,r}$$

where $s_{i,c,r}$ is the score assigned by rater r to candidate i on criterion c and $n = 5$. Mean scores therefore lie within the interval $[0, 1]$.

The SSE dataset comprised three candidate types: 11 full reconstructions, 3 drapery studies and 4 head studies. Criteria were applied only where relevant and non-applicable criteria were excluded.

5.3 Results

5.3.1 Baseline Generative Behaviour

Baseline outputs conveyed plausible scale and surface realism but diverged from the archaeological and iconographic record (Figure 24). Common deviations included non-Gandhāran drapery grammars, anatomically implausible hands and stylised hair inconsistent with documented Gandhāran forms (Klimburg-Salter, 1989; Rienjang & Stewart, 2023; Yamauchi & Bunka, 2012). Structural and contextual distortions were also frequent, including exaggerated niche depth, Petra-like façades and unrealistic vegetation (Figure 25). These deviations were identified by experts in both evaluation rounds (Phase 1: cross-disciplinary panel; Phase 2: Bamiyan specialist), stemmed from generic prompt clauses such as “flowing robe,” “rocky cliffs,” and “curly hair,” as well as unconstrained lighting descriptors that introduced cinematic stylisation. Figure 26 presents Images 1–7, representative baseline AI-only reconstructions, illustrating structural and iconographic instability prior to iterative refinement.

Progress across iterations is illustrated in Figure 27, where successive prompt adjustments improved head proportions, hand articulation and robe drapery. Despite these improvements, residual symmetry and stylisation persisted. These residual inconsistencies motivated the introduction of a structured computational pre-evaluation stage, detailed in Section 5.2.3, prior to expert scoring.

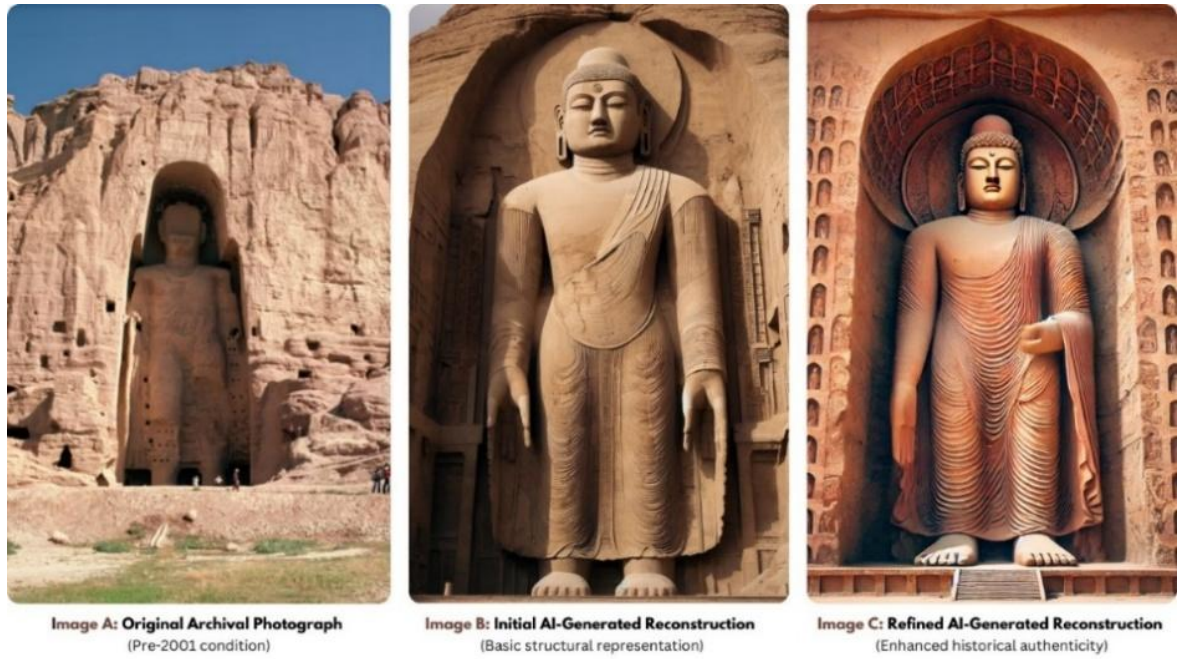


Figure 24. Archival photograph of the Giant Buddha at Bamiyan © UNESCO (left), with AI-generated baseline reconstruction (centre) and AI-generated refined reconstruction (right).

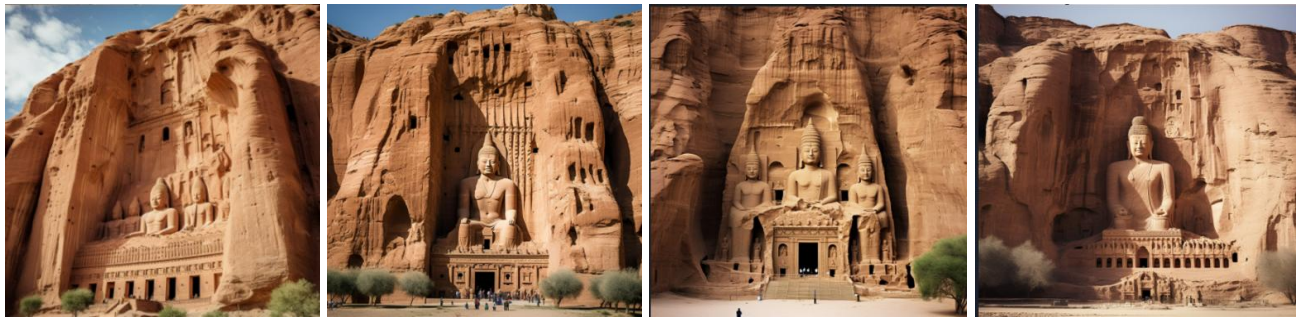


Figure 25. Structurally inconsistent AI-generated reconstructions presenting architectural and contextual deviations. All images were generated without DPB slot constraints.

- (a) Incorporation of Petra-like carved façade geometry within the niche (far left);
- (b) Addition of a temple-style podium and secondary monumental figures (second from left);
- (c) Introduction of dense foreground vegetation not evidenced in archival documentation (third from left);
- (d) Symmetrical niche redesign and halo-like sculptural additions (far right).

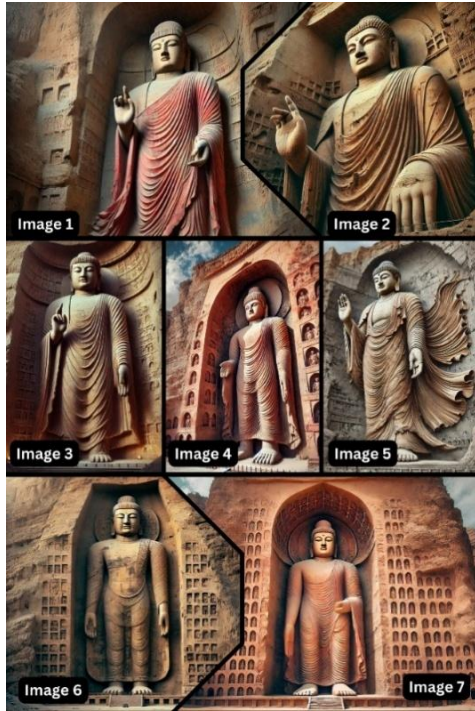


Figure 26. Phase I baseline AI-generated reconstruction candidates (Images 1–7) generated using OpenAI DALL-E. Each image was produced from a single-shot prompt without DPB slot constraints and iterative refinement. No post-generation editing was applied. These images formed the first part of the Phase I questionnaire screening set and were used to assess baseline generative behaviour before DPB-guided refinement. The outputs demonstrate structural and iconographic inconsistencies characteristic of unconstrained generative prompting.

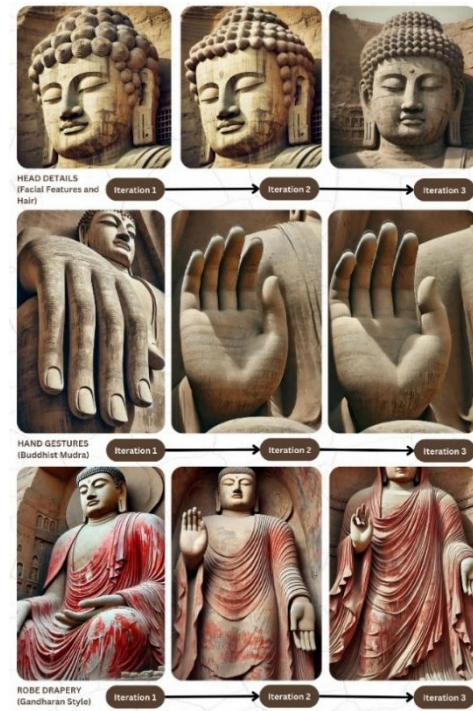


Figure 27. AI-generated iterative DPB-guided refinement of head morphology, hand articulation and drapery across three successive generations. Iteration 1 was generated using an initial DPB-structured prompt. Iterations 2 and 3 were produced within the same OpenAI ChatGPT image-generation session through progressive prompt modification, introducing increasingly specific morphological and iconographic constraints for facial features, hand gestures and drapery articulation. The underlying prompt architecture remained constant, while constraint granularity was incrementally increased. All outputs are direct model generations without manual pixel-level editing.

5.3.2 Corpus Refinement and Platform Behaviour

[30]The retained corpus described in Section 5.2.2 was iteratively refined through controlled prompt revision, targeted API-based edits and selected masked-generation workflows, following the DPB-guided HARF procedure (Arzomand et al., 2026). It comprised full-figure reconstructions, drapery-focused variants and head-only studies. Earlier exploratory baseline outputs, including DALL-E examples used to illustrate pre-HARF behaviour, were retained only as comparative examples and were not counted within the retained corpus. Early generations recovered some dimensional anchors but frequently contained artefactual insertions (Appendix C.5 and Appendix C.6). A subset of unconstrained Gemini exploratory outputs introduced speculative architectural elements or distorted proportions and was therefore excluded from the evaluated shortlist, while remaining documented in the paradata.

Other candidates exhibited canonical drift, particularly in robe-fold logic, mudrā configuration and Petra-like architectural insertions inconsistent with the Bamiyan archaeological record (Appendix C. 1 and C. 2). Selected outputs were then refined through targeted API-based editing to stabilise features that remained archaeologically important but visually unstable. These edits addressed robe-fold articulation, hand positioning, cranial height, bun silhouette, wave morphology and polychromy cues while retaining the evidential constraints of the reconstruction workflow (Figure 28-Figure 29; Appendix C. 3-C. 4). Through this process of generation, exclusion and targeted refinement, the corpus was reduced to the candidates retained for structured expert evaluation.



Figure 28. AI-assisted reference-conditioned refinement of Buddha head morphology using Sora. A reference image of a comparable Gandhāran Buddha head was used to guide generation through successive prompt refinements describing cranial height, bun form and hair wave structure. The sequence shows the progressive adjustments leading to the final generated head.

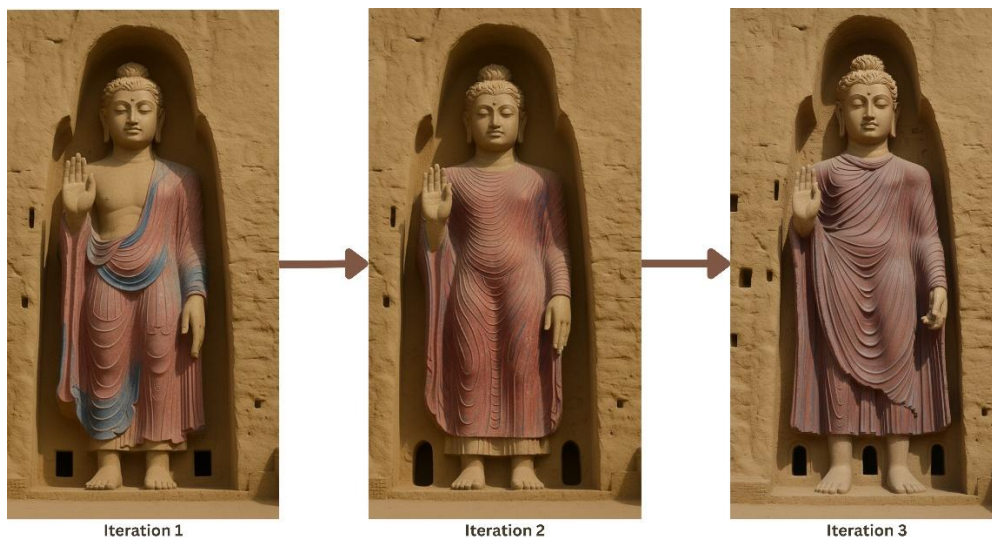


Figure 29. AI-assisted targeted regional refinement of drapery articulation using the Gemini API. Targeted regional edits were applied through the Gemini API to correct robe folds and garment layering. Successive iterations progressively stabilised the wet-fold structure, garment layering and under-robe visibility.

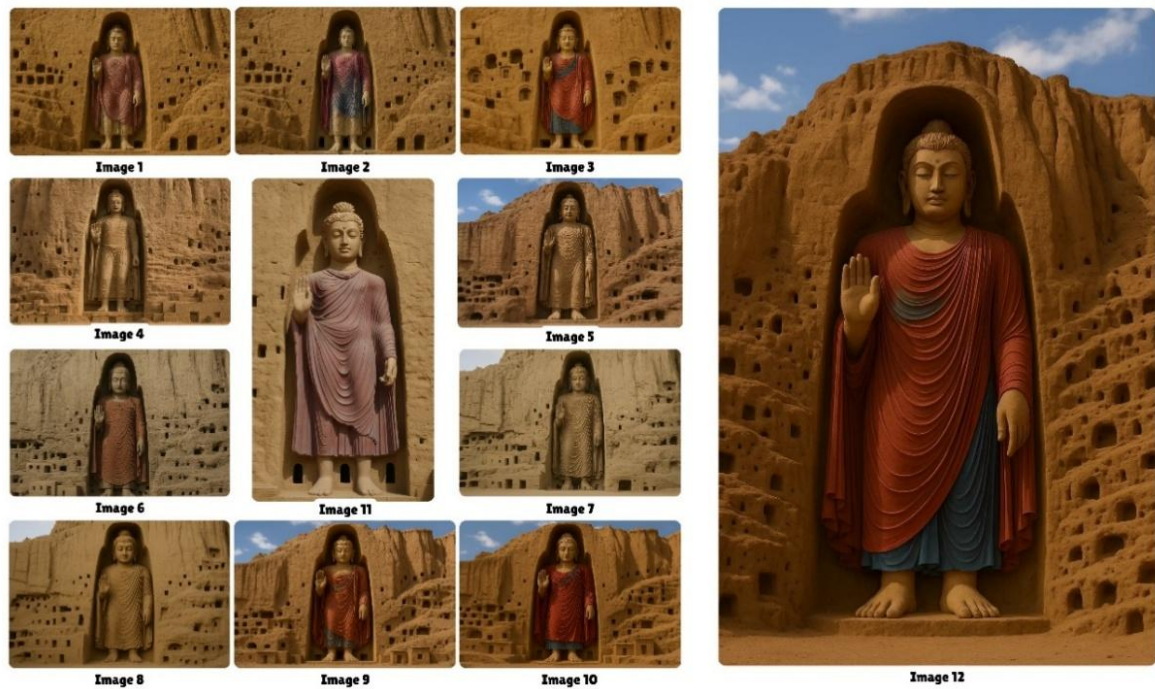


Figure 30. Final set of reconstruction candidates retained for expert evaluation. Images 1–11 were generated through AI workflows, while Image 12 (right) represents a manually corrected reconstruction produced in Adobe Photoshop.

Through generation, exclusion and targeted refinement, the corpus was narrowed to 23 images for Phase I — IES organised into three evaluation domains: structural (Images 1-7), iconographic (Images 8-17) and environmental (Images 18-23). Feedback from IES informed further prompt refinement, producing a final set of 18 refined images for Phase II — SSE.

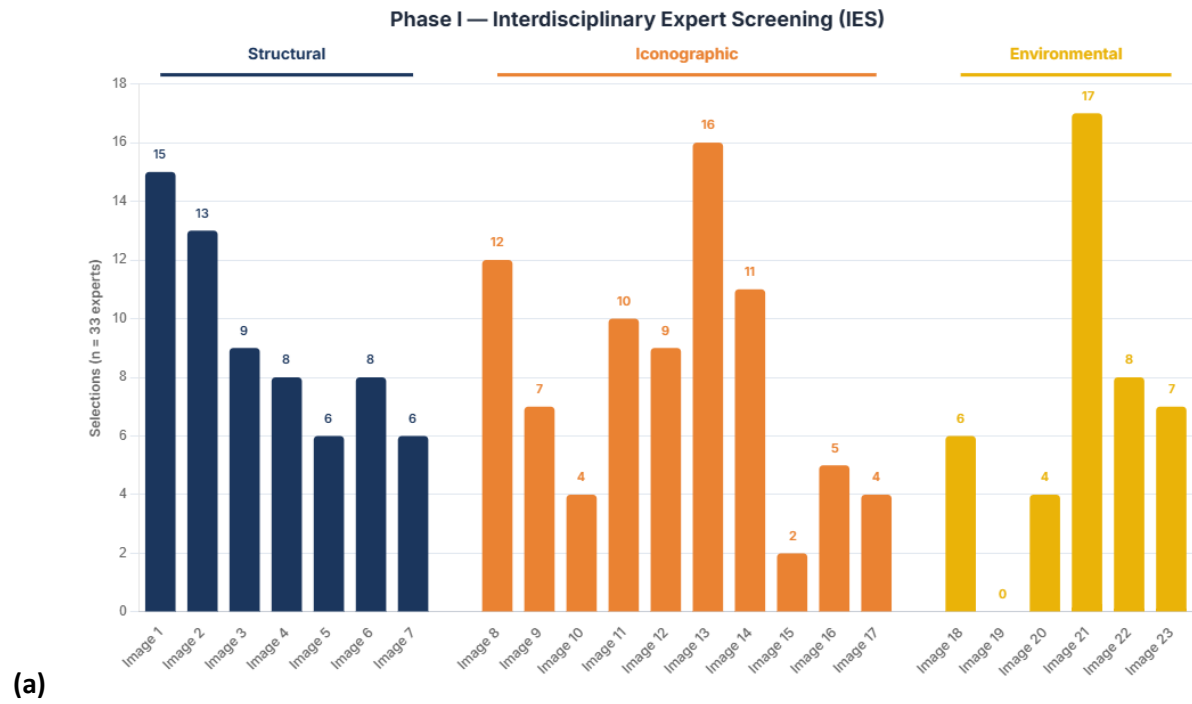
The SSE dataset comprised 11 full-figure reconstructions (Figure 30), 3 drapery studies and 4 head studies (Figure 28 and Figure 29). A composite visual reference (Image 12; Figure 30) was additionally produced in Adobe Photoshop. It was excluded from computational evaluation and expert scoring and is presented solely for visual comparison. Interactive masks were used to isolate local regions during refinement and constrain regional regeneration (Figure 31).



Figure 31. Author-produced masking illustration from the AI-assisted refinement workflow. Highlighted regions indicate areas selected for targeted prompt refinement, including hair morphology, hand articulation, niche openings behind the feet and foreground architectural elements. These regions guided the iterative correction process during AI image generation. These masks were used during generation refinement and were not part of the computational evaluation pipeline.

5.3.3 Candidate Selection and Structural Evaluation

Experts were permitted to select more than one reconstruction as archaeologically plausible; endorsement totals therefore exceed the number of respondents ($n = 33$). As shown in Figure 32(a), selection frequencies varied across the structural, iconographic and environmental screening domains. Within the structural subset (Images 1–7; Figure 26), Images 1 and 2 received the strongest endorsement. In the iconographic subset (Images 8–17; Figure 33), Image 13 was selected most frequently, while Image 21 dominated the environmental subset (Images 18–23; Figure 34). These patterns indicate domain-specific variation in plausibility assessment.



(a)

	Image	SSIM_statue	SSIM_niche	PropSimilarity	CompositeScore
0	Image 2	0.198000	0.210000	0.944000	0.451000
1	Image 7	0.168000	0.170000	0.944000	0.427000
2	Image 3	0.144000	0.141000	0.944000	0.410000
3	Image 9	0.108000	0.122000	0.944000	0.391000
4	Image 10	0.021000	0.029000	0.944000	0.331000
5	Image 11	0.229000	0.243000	0.518000	0.330000
6	Image 4	0.012000	0.010000	0.944000	0.322000
7	Image 6	0.091000	0.097000	0.776000	0.321000
8	Image 1	0.003000	0.011000	0.944000	0.320000
9	Image 5	0.128000	0.132000	0.554000	0.272000
10	Image 8	0.156000	0.153000	0.434000	0.247000

(b)

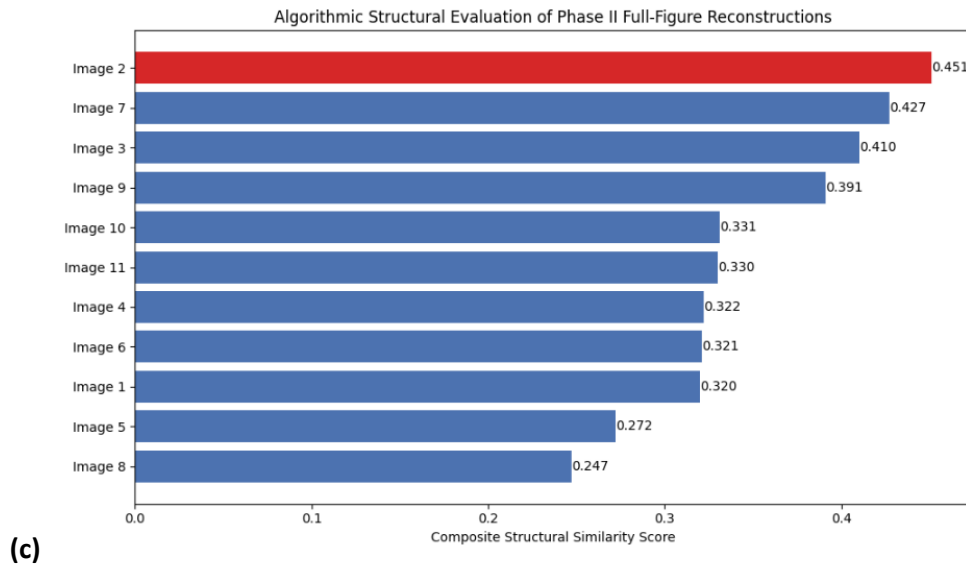


Figure 32. Author-generated charts. Candidate selection and structural evaluation across analysis phases. (a) Phase I – Interdisciplinary Expert Screening (IES) selections across evaluation domains. Bars show the frequency with which reconstruction candidates were selected by respondents ($n = 33$) across the three screening domains: structural plausibility ($k=7$; Images 1–7), iconographic authenticity ($k=10$; Images 8–17) and environmental context ($k=6$; Images 18–23). Multiple selections were permitted within each domain. (b) Structural similarity metrics for Phase II full-figure reconstruction candidates. The figure reports statue-region SSIM, niche-region SSIM, proportional similarity and the resulting composite structural similarity score used to rank the eleven reconstruction candidates relative to the archival reference image. (c) Algorithmic structural ranking of Phase II full-figure reconstructions. Higher values indicate stronger geometric correspondence with the archival reference photograph.



Figure 33. AI-generated iconographic reconstruction candidates used during Phase I — Interdisciplinary Expert Screening (IES), including head morphology, hand gesture and drapery studies (Images 8–17).

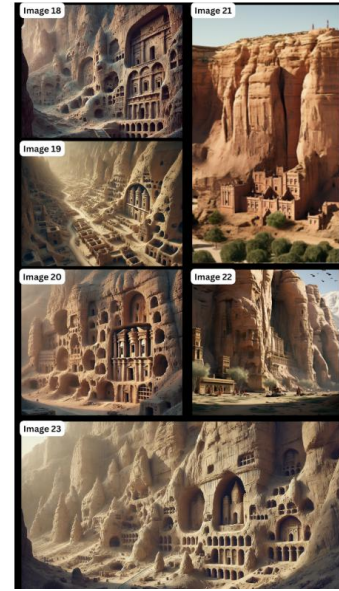


Figure 34. AI-generated environmental context reconstruction candidates used during Phase I — Interdisciplinary Expert Screening (IES), illustrating alternative interpretations of the Bamiyan cliff architecture and surrounding cave landscape (Images 18–23).

In addition to numerical voting, experts provided qualitative justifications for their selections (Table 12), which consistently emphasised robe realism, proportional balance and niche integration in Images 1 and 2, whereas Images 4 and 7 were more often selected for visual resemblance or pigmentation despite weaker archaeological correspondence.

Per-candidate structural similarity scores for the eleven full-figure reconstructions are shown in Figure 32(b-c). Image 2 ranked highest (CompositeScore = 0.451), followed by Images 7 (0.427) and 3 (0.410). Lower-ranked candidates demonstrated weaker structural correspondence and greater deviation from archival proportions.

Table 12. Expert rationales for structural-domain preferences during IES (Images 1–7; Figure 26). Key reasons and common feedback cited by respondents (n = 33).

Image	Key Reasons Experts Selected It	Details from Feedback
Image 1 (14 votes)	<i>Most selected overall</i> Red pigmentation, realistic niche, proportions, environment and overall authenticity	Accurate drapery, red robes, balanced proportions and plausible cliff context; frequently described as the most realistic overall. Minor critique concerned finger count and niche-wall treatment.
Image 2 (12 votes)	Photographic resemblance, natural sculpture feel, aging and realistic imperfections	Strong visual similarity to historical photos and memory- Correct pose, colour and textural detail- Recognized by experts referencing oral history or elders' recollections- Sculptural surface is imperfect, which adds authenticity
Image 3 (8 votes)	Realism with full-figure composition	Viewed as most photorealistic by multiple experts- Includes top-to-bottom view of the statue- Strong proportions and pose- Slightly exaggerated niche width, but otherwise highly praised
Image 4 (7 votes)	Memory resemblance and partial niche/pigment accuracy	Recalled based on visual memory or imagination- Cited for red pigmentation and somewhat accurate cave/niche shape- Considered closer to the original than others by some, even though structurally less complete
Image 5 (5 votes)	Robe flow, arm position and gesture accuracy	Robes noted for realistic draping and flow- Correct hand positioning (based on analogical examples)- Several respondents criticised excessive symmetry and surface regularity, which they associated with a digitally generated aesthetic
Image 6 (7 votes)	Background, niche shape, leaner anatomy and facial detail	Background and niche context closest to reality, though niche still too wide and shallow- Statue has a leaner body and better facial sculpting than others- Critiques included broad shoulders, wrong mudra and robe inaccuracy
Image 7 (5 votes)	Red pigment, facial impression and familiarity	Selected due to red colour, facial impression and subjective resemblance to memory- Sometimes grouped with Image 4 for similar visual appeal- Critiqued for lack of historical evidence or accuracy to 20th-century documentation

Note: Participants could endorse multiple images; endorsement counts therefore exceed n = 33.

5.3.4 Forced-Choice Comparison Results

In addition to the multi-select screening task, respondents completed a forced-choice comparison between the initial baseline reconstruction, the refined AI reconstruction (Figure 24) and the option “neither accurately represents historical reality”.

Most respondents did not consider either reconstruction historically plausible. A majority of respondents (51.5%) selected “neither”, while (27.3%) preferred the refined reconstruction and 21.2% the initial baseline. Those preferring the initial reconstruction emphasised restraint and the avoidance of speculative additions, whereas those selecting the refined version cited improvements in drapery articulation and proportional balance. Respondents choosing “neither” referred to fabricated architectural elements, anatomical instability and stylistic features inconsistent with documented Bamiyan and Gandhāran conventions (Figure 27).

Although the refined reconstruction appeared more visually polished, several experts criticised it for stylistic amplification and the introduction of non-local motifs (Bentkowska-Kafel & Denard, 2012). The predominance of the “neither” response indicates that visual refinement alone did not resolve concerns regarding historical authenticity. Expert preferences in IES also diverged from the computational ranking, converging on Images 1 and 2 rather than following the full algorithmic ordering.

5.3.5 Full-Figure Reconstruction Scores (Images 1–11; Figure 30)

Table 13 reports SSE scores for the 11 full-figure candidates across ten criteria: proportion, niche–cliff integration, drapery structure, drapery colour, hands, feet, head/hair morphology, facial plausibility, surface erosion and landscape context.

To compare the full-figure candidates, for which all ten criteria were applicable, an overall candidate-level mean score was calculated across the ten criteria:

$$S_i = (1 / 10) \sum c = 1^{10} s_{i,c}$$

where $s_{i,c}$ denotes the mean SSE score for candidate i on criterion c .

Across the full-figure subset, candidate means ranged from 0.39 to 0.71. Image 2 scored highest (≈ 0.71), followed by Image 1 (≈ 0.68) and Image 11 (≈ 0.65). Lower-scoring candidates, including Images 5, 3 and 8, tended to combine weaker performance across proportion, facial plausibility and contextual integration rather than failing on a single criterion alone.

Table 13. Mean SSE scores for full-figure reconstruction candidates (Images 1–11; Figure 30). Values represent the arithmetic mean of ratings assigned independently by five specialists using the SSE rubric (0 = absent/incorrect, 0.5 = partially correct/ambiguous, 1 = correct/clearly satisfied). Candidate-level means were calculated as the arithmetic mean across the ten applicable criteria. The three highest-ranking candidates (Images 2, 1 and 11) are highlighted for visual clarity. Note: N/A entries were excluded from criterion-specific means.

Image	Proportion	Niche Cliff	Drapery	Drapery Colour	Hands	Feet	Head Hair	Facial	Surface Erosion	Landscape
Image 1	0.70	0.60	0.80	0.70	0.70	0.90	0.60	0.60	0.50	0.70
Image 2	0.80	0.70	0.70	0.50	0.80	0.90	0.70	0.60	0.60	0.80
Image 3	0.20	0.30	0.80	0.40	0.80	0.70	0.40	0.50	0.30	0.20
Image 4	1.00	0.90	0.50	0.10	0.70	0.50	0.30	0.50	0.50	0.70
Image 5	0.40	0.50	0.40	0.10	0.60	0.50	0.20	0.20	0.50	0.50
Image 6	0.30	0.80	0.70	0.50	0.60	0.50	0.30	0.40	0.50	0.60
Image 7	0.70	0.70	0.60	0.20	0.40	0.25	0.30	0.70	0.30	0.60
Image 8	0.20	0.50	0.50	0.20	0.50	0.50	0.60	0.60	0.30	0.60
Image 9	0.30	0.70	0.80	0.60	0.70	0.40	0.40	0.60	0.50	0.70
Image 10	0.40	0.60	0.80	0.50	0.70	0.50	0.50	0.70	0.40	0.70
Image 11	0.90	0.50	0.80	0.50	0.60	0.80	0.80	0.70	0.50	0.38

5.3.5.1 Comparison with Algorithmic Structural Similarity

SSE results for the 11 full-figure candidates were compared with the corresponding composite algorithmic similarity scores (Figure 35(a)). The association was positive but weak and statistically non-significant. The relative magnitude of expert and algorithmic scores for each candidate is shown in Figure 35(b). Spearman's rank correlation analysis produced $p = 0.369$ ($p = 0.264$) across the eleven full-figure candidates. Given the small subset, the coefficient is used only to characterise the observed ranking pattern. It shows that computational structural similarity and specialist archaeological assessment aligned only partially, rather than operating as interchangeable measures. Correlations between SSE mean scores and individual computational metrics are reported in Table 14.

Table 14. Spearman correlations between computational similarity metrics and expert SSE scores.

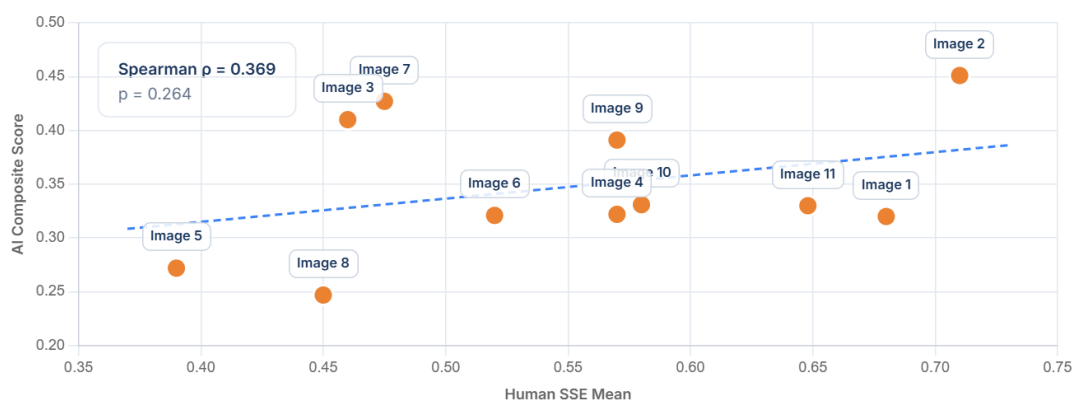
Metric	Spearman ρ	p
SSIM_statue	-0.077	0.821
SSIM_niche	-0.046	0.894
PropSimilarity	0.412	0.209
CompositeScore	0.369	0.264

Structural similarity measures (SSIM_statue and SSIM_niche) showed negligible correlation with expert evaluations, whereas proportion similarity showed the strongest association with SSE scores ($\rho = 0.412$). The composite algorithmic score ($\rho = 0.369$) therefore indicates only partial alignment between computational filtering and specialist evaluation.

Alignment between the two ranking systems was uneven (Table 15). Image 2 ranked first in both, but Images 7 and 3 received relatively high algorithmic scores alongside lower expert evaluations, whereas Image 1 received a comparatively high specialist score despite modest algorithmic similarity. Image 1 showed the largest discrepancy between the two systems and Image 7 the smallest.

Table 15. Comparison of expert and algorithmic rankings for full reconstruction candidates (Images 1–11; Figure 30). Human rank is based on the candidate-level mean SSE score across ten criteria. AI rank is based on the composite structural similarity score reported in Figure 32. Rank difference is calculated as the absolute difference between human and algorithmic positions. Agreement categories are defined as: Match (difference = 0), Close (difference = 1), Moderate divergence (difference = 2–3) and Large divergence (difference ≥ 4).

Image	Human Rank	AI Rank	Rank Difference	Agreement
Image 2	1	1	0	Match
Image 1	2	9	7	Large divergence
Image 11	3	6	3	Moderate
Image 9	4	4	0	Match
Image 10	5	5	0	Match
Image 4	6	7	1	Close
Image 6	7	8	1	Close
Image 7	8	2	6	Large divergence
Image 5	9	10	1	Close
Image 8	10	11	1	Close
Image 3	11	3	8	Large divergence



(a)



Figure 35. Author-generated chart showing the relationship between expert SSE mean scores and AI composite similarity scores for the eleven full-figure reconstruction candidates. Panel (a) reports the relationship between expert and computational scores; panel (b) compares expert SSE mean scores and AI composite similarity scores across the eleven candidates.

5.3.5.2 Mean Applicable-Criteria Scores Across All Candidates (k=18)

Eighteen shortlisted candidates were evaluated by five Bamiyan specialists using the SSE rubric. Candidate provenance, including subset category, platform of origin and targeted merges or corrective edits, is reported in Appendix C. 8.

Specialist evaluations were aggregated as unweighted mean scores across applicable criteria, with non-applicable (N/A) criteria excluded. No criterion weighting was applied; all applicable criteria contributed equally. For each candidate, each specialist's mean score was first calculated across the criteria applicable to that candidate. The candidate-level score was then calculated as the mean of the five specialist means:

$$S_i = (1/n) \sum_{r=1}^n [(1/|C_{i,r}|) \sum_{c \in C_{i,r}} s_{i,r,c}]$$

where $s_{i,r,c}$ is the score assigned by specialist r to candidate i on applicable criterion c , $C_{i,r}$ is the set of criteria applicable to candidate i for specialist r , and $n = 5$.

Figure 36 summarises mean specialist scores across the three candidate subsets. Head-focused reconstructions achieved the highest scores, with head3 and head1 both reaching 0.83. Drapery-focused studies occupied an intermediate range (0.63–0.69), whereas full-figure reconstructions showed lower scores and greater variability. Among the full-figure candidates, Image 2 scored highest (0.71), followed by Image 1 (0.68) and Image 11 (0.65).

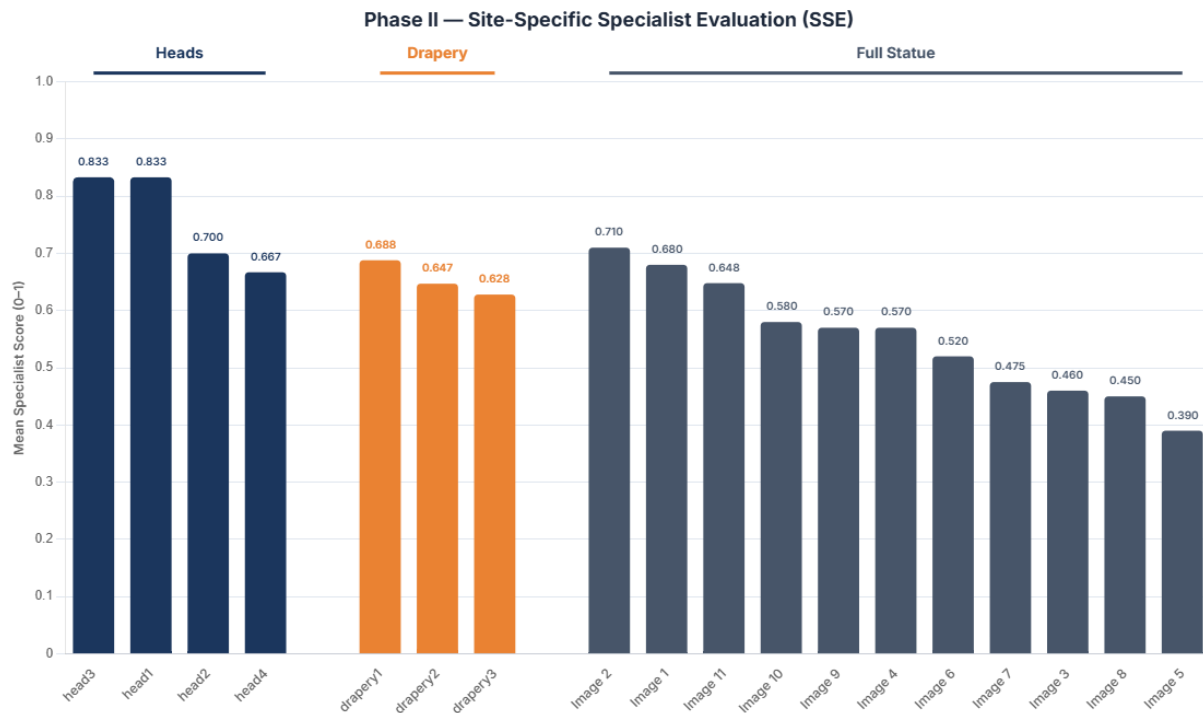


Figure 36. Author-generated chart based on Phase II — Site-Specific Specialist Evaluation (SSE) mean applicable-criteria scores for the 18 candidates. Mean specialist scores are reported on a 0–1 scale for head studies, drapery studies and full-figure reconstructions.

Specialist evaluations showed a clear structural pattern: reconstructions targeting morphologically constrained zones, particularly cranial structure and facial morphology, were judged more plausible than full-figure reconstructions requiring simultaneous resolution of pose, drapery and environmental integration. Full-figure candidates frequently accumulated smaller inaccuracies across multiple domains, especially robe-fold grammar and hand articulation. These patterns are consistent with the candidate provenance reported in Appendix C. 8.

5.3.5.3 Criterion-Level Means Across Candidates

To examine how specialist assessments varied across individual domains of the SSE rubric, criterion-level mean scores were calculated across candidates and are reported in Figure 37. Performance was strongest for structural proportionality and cranial morphology and weakest for articulated stylistic domains, particularly Drapery articulation, Drapery colour coherence and Hand morphology.

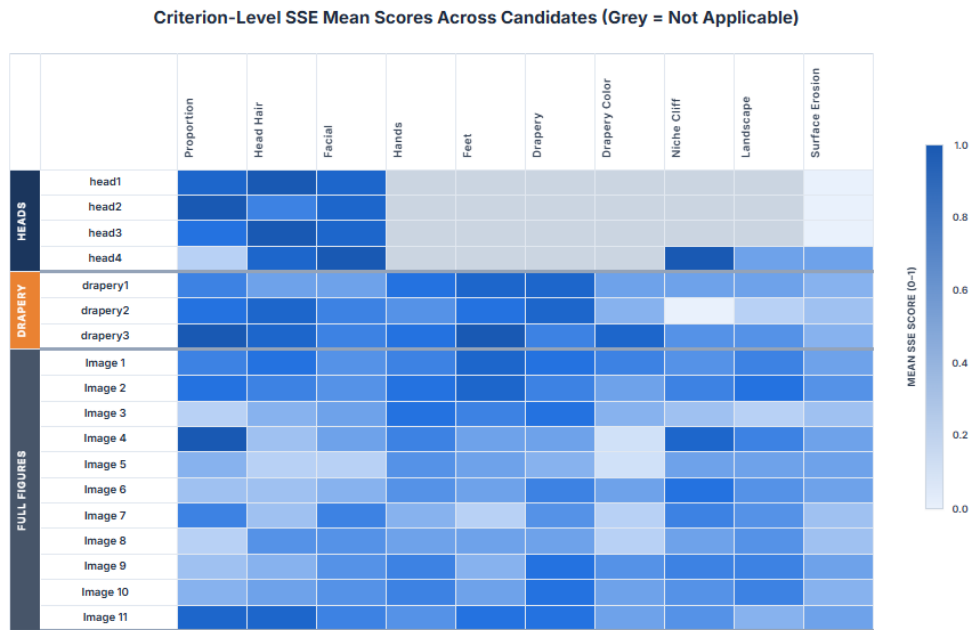


Figure 37. Author-generated heat map based on Phase II SSE criterion-level scores across candidates. Darker cells represent higher mean specialist scores (0–1). Grey cells indicate criteria not applicable to a given candidate subset.

5.3.5.4 Inter-Rater Agreement by Criterion

Inter-rater agreement across evaluation criteria is reported in Figure 38 as quadratic-weighted Cohen’s κ and Gwet’s AC2 coefficients with 95% confidence intervals (K. L. Gwet, 2008; Landis & Koch, 1977). Agreement was strongest for Proportion ($\kappa = 0.55$; AC2 = 0.66), while the weakest agreement appeared in criteria requiring fine-grained anatomical, stylistic and surface-detail judgement, including Hand morphology and Surface Erosion. The difference between κ and AC2 is consistent with the known sensitivity of κ to uneven rating distributions, with AC2 providing a more stable estimate in such cases (K. L. Gwet, 2008). The criterion-level pattern indicates that the five-rater SSE design was most stable for broad structural assessment and more variable where the judgement required specialist interpretation of damaged, partially inferred or visually unstable heritage features.[87]

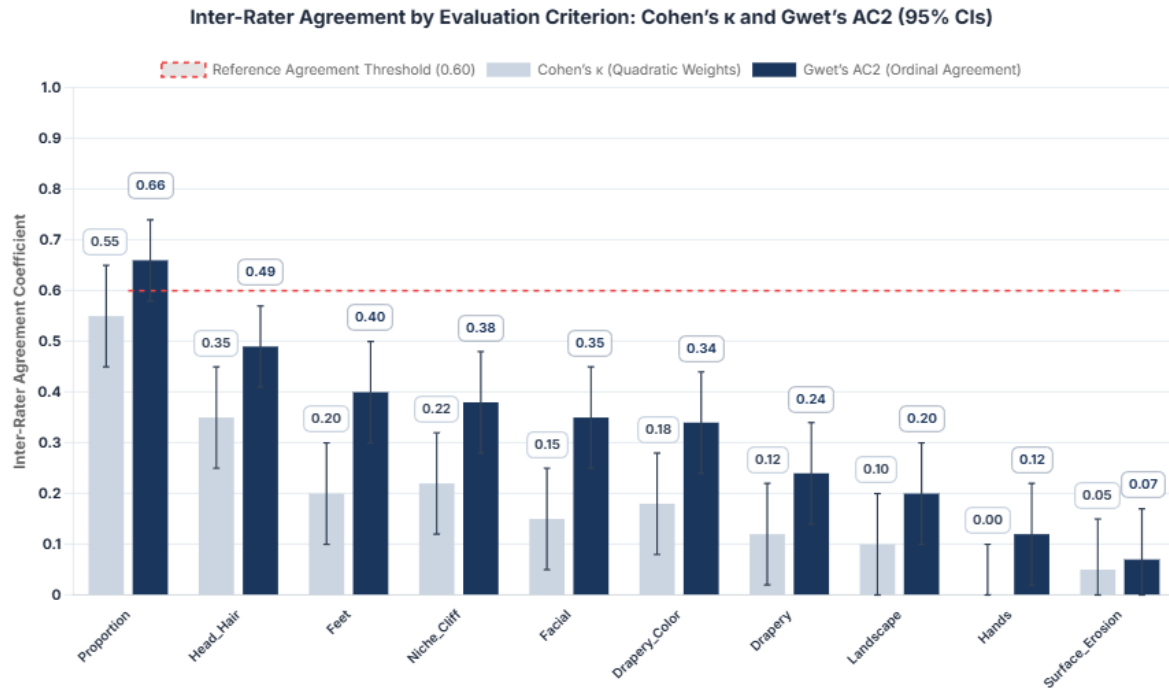


Figure 38. Author-generated inter-rater agreement chart across SSE evaluation criteria. Criterion-wise agreement was calculated from the five Phase II specialist ratings using quadratic-weighted Cohen's κ and Gwet's AC2, with 95% confidence intervals. N/A entries were excluded from each criterion-specific calculation. The dashed line at 0.60 is used as a visual reference threshold for interpretive comparison.

5.4 Discussion

The results indicate that structured prompting and algorithmic filtering improve proportional stability within the reconstruction workflow, but do not eliminate the need for expert-guided correction.

5.4.1 Divergence Between Algorithmic Similarity and Expert Evaluation

These findings indicate a structural asymmetry between computational screening and archaeological assessment. Computational filtering provides a reproducible measure of two-dimensional geometric correspondence to the archival reference, but stylistic credibility and iconographic plausibility remain dependent on domain expertise. Comparable divergences between visual plausibility and historically grounded authenticity have been noted in recent work on generative heritage reconstruction (Xu et al., 2025).

The Bamiyan results show that computational metrics remain less sensitive to archaeologically significant stylistic deviations. Several candidates with high structural similarity scores retained non-canonical drapery articulation or stylised facial features that were later identified during expert review (Appendix C. 1-C. 2; C. 5-C. 6). The divergence should therefore be read as a difference in evaluative

function: computational screening identifies two-dimensional structural correspondence, while specialist assessment judges archaeological and stylistic plausibility.

Provenance mapping (Appendix C. 8) and SSE mean applicable-criteria scores (Figure 36 and Figure 37) indicate that candidates derived from hybrid cross-platform refinement generally achieved higher specialist evaluations among non-head reconstructions. These candidates combined geometrically stable base generations with targeted regional corrections applied through API-guided edits or masked regeneration.

The head and drapery subsets further demonstrate that such targeted interventions can partially correct recurrent artefacts, particularly in cranial morphology and fold articulation, by isolating morphologically constrained regions. However, improvements remained localised and did not consistently extend to full-figure coherence, where interactions between pose, drapery and environmental integration introduced additional instability.

These deviations instead identify structural domains in which generative modelling exhibits systematic instability during cultural heritage reconstruction. Specialist evaluation remains disciplinary and archaeological in scope; community-grounded cultural acceptability is addressed separately in Chapter 6. Because Phase I (IES) evaluated a broader image set than Phase II (SSE), the instability analysis below draws primarily on Phase I observations.

5.4.2 Recurrent Anatomical and Drapery Instabilities

Expert evaluation consistently identified three principal instability domains: hand articulation, drapery structure and cranial treatment. Hand gestures remained unstable, often appearing mirrored, incomplete, or iconographically speculative. Drapery was often rendered in overly symmetrical or stylised folds that diverged from the gravity-driven “wet-cloth” articulation characteristic of Gandhāran sculpture (British Museum & Zwalf, 1996; Rienjang & Stewart, 2023).

Cranial and facial treatments were similarly affected by excessive surface smoothing. Many outputs displayed excessively uniform facial surfaces lacking erosion, asymmetry and material irregularities characteristic of ancient stone sculpture. Additional deviations included stylised hair textures resembling “snail curls” but not the elongated wavy strand morphology typical of Gandhāran statuary.

These instabilities concentrated in morphologically complex regions, indicating difficulty in reproducing interactions between structural geometry and culturally specific stylistic conventions. Such domains require the simultaneous resolution of anatomical structure, sculptural grammar and culturally specific stylistic conventions, which proved more difficult to reproduce consistently within

the generated reconstructions. By contrast, contextual variables such as niche–cliff integration and landscape coherence remained relatively stable. The recurrent substitution of Gandhāran wavy strand morphology with stylised “snail curls” is consistent with documented bias propagation in cultural-heritage AI pipelines, in which dominant visual patterns present in training data are reproduced or amplified during deployment (Foka et al., 2025).

Despite these instabilities, several outputs demonstrated partial strengths, including plausible robe silhouettes, proportional balance and coherent statue–niche relationships. Generative systems can therefore approximate elements of archaeological plausibility when structural constraints are encoded, but full-figure reconstructions remain limited by persistent drapery and hand-articulation errors. HARF appears to stabilise proportional structure more effectively than culturally embedded stylistic grammar (Appendix C. 7).

Environmental reconstructions frequently introduced speculative or culturally inconsistent elements, including Petra-like façades and vegetation inconsistent with the Bamiyan Valley (ICOMOS, 2003). Although some outputs achieved plausible cliff stratification, most exhibited over-generation of contextual detail (Appendix C. 7), suggesting that generative models prioritise visual completeness over archaeological constraint.

5.4.3 Iterative Refinement Limits

Later reconstruction iterations (Figure 27) achieved noticeable improvements in proportional balance, robe clarity and niche integration. Across Phase I, 16 expert responses identified drapery articulation as the most unstable reconstruction domain, with robe folds described as overly symmetrical and inconsistent with Gandhāran “wet-cloth” grammar (British Museum & Zwalf, 1996).

Approximately 73% of responses criticised cranial treatments as overly smooth, culturally inconsistent, or insufficiently weathered. Hand gestures also remained unstable across iterations, frequently appearing mirrored, incomplete, or iconographically speculative. These observations indicate that visual refinement often progressed faster than historical fidelity, as reflected in specialist feedback, where improvements in surface realism did not consistently translate into historically grounded iconographic accuracy. Later generations appeared increasingly polished and persuasive to non-specialist yet were still judged by experts to be “visually compelling yet archaeologically misleading” (Bentkowska-Kafel & Denard, 2012).

Figure 39 summarises expert observations across the iterative reconstruction stages shown in Figure 27. Although cranial detail improved between Iterations 1 and 3, instability in drapery articulation and hand gestures persisted throughout the refinement process. Reconstructions combining proportional

balance, credible drapery and coherent niche integration were evaluated most favourably, whereas outputs with speculative or unsupported elements were rejected.

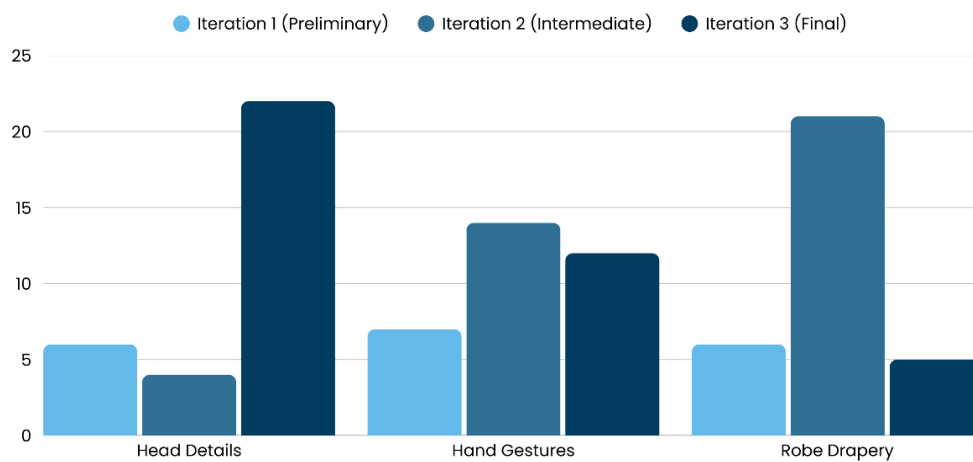


Figure 39. Author-generated chart based on expert identification of reconstruction instabilities across iterative refinement stages based on the reconstruction outputs illustrated in Figure 27. The chart compares the frequency with which experts identified three diagnostic features, head details, hand gestures and robe drapery, as problematic across Iterations 1 (Preliminary), 2 (Intermediate) and 3 (Final). The vertical axis represents the number of expert observations.

This divergence highlights a fundamental credibility gap between algorithmic optimisation and archaeological interpretation. Generative models prioritise visual coherence and aesthetic realism, whereas expert evaluation emphasises evidentiary grounding and historical plausibility. These findings therefore reinforce the necessity of expert review within the reconstruction workflow.

Paradata documenting expert rationales and subsequent prompt adjustments recorded the basis for candidate refinement and helped interpret recurring failure patterns. The results indicate a divergence between geometric stabilisation and iconographic fidelity: while proportional anchors converge under structured prompting, stylistic micro-grammar remains resistant to automated generation. Expert mediation therefore remains essential when reconstructing iconographically complex elements.

5.4.4 Expert Recommendations and Methodological Implications

Building on the identified instability patterns, specialists proposed methodological recommendations to improve the reliability of AI-assisted reconstructions of the Buddhas of Bamiyan. These recommendations primarily concerned three areas: material realism, contextual grounding and interpretive transparency.

Experts emphasised the need to improve material and morphological fidelity. Cliff textures were often rendered with digitally uniform surfaces that did not reflect Bamiyan geology, while drapery remained overly symmetrical or stylised and inconsistent with Gandhāran “wet-cloth” grammar (British Museum & Zwalf, 1996; Rienjang & Stewart, 2023). They therefore recommended tighter prompt constraints,

including stronger pose control for hand gestures and the explicit exclusion of speculative architectural additions.

Experts also highlighted the need for clearer contextual and temporal framing. In several outputs it remained unclear whether the images represented the monument at original construction, during later weathering phases, or immediately prior to its destruction in 2001. Without such clarification, generated images risk conflating restorative restitution with hypothetical visualisation. To reduce this ambiguity, experts recommended anchoring prompt design more explicitly in Afghan and Gandhāran sculptural exemplars and in established conservation documentation (ICOMOS, 2003; Toubekis et al., 2017).

Several experts further cautioned that these outputs should be treated as interpretive visualisations rather than definitive reconstructions, given the evidentiary gaps surrounding the original appearance of the statues (Lowenthal, 1998). In digital heritage contexts, highly photorealistic renderings can imply a degree of historical certainty not supported by the evidence (Bentkowska-Kafel & Denard, 2012; Denard, 2009; Smith, 2006). These observations reinforce HARF's premise: reconstructions must remain evidentially constrained, interpretively transparent and supported by expert-in-the-loop evaluation.

5.5 Limitations

Several limitations condition the interpretation of these findings. Consistent with the transparency principles of the London Charter for the Computer-based Visualisation of Cultural Heritage and the Seville Principles for Virtual Archaeology, these are addressed across technical, methodological and interpretive dimensions (Denard, 2009; López-Menchero Bendicho & Grande, 2011, 2017).

A primary limitation concerns the divergence between computational pre-evaluation and expert archaeological judgement. Structural similarity measures such as SSIM and ORB-RANSAC enable reproducible geometric comparison between generated images and archival references; however, in this chapter they were applied only to the eleven full-figure candidates and remain insensitive to culturally embedded stylistic cues and iconographic nuance. In archaeological reconstruction, where evidence is incomplete and historically mediated, algorithmic scoring therefore operates only as a preliminary filtering mechanism, with final plausibility assessment dependent on expert evaluation.

A further methodological limitation is that computational structural comparison relied on a single archival reference photograph. Because alignment used planar homography to one two-dimensional view, the composite score remained conditioned by that photograph's viewpoint, illumination and

lens geometry. Differences in camera angle, lighting, photographic distortion and occlusion could not be separated fully from structural differences in the generated candidates.

A further limitation arises from biases embedded within the training data of large-scale generative models (E. M. Bender et al., 2021). Trained on heterogeneous internet image corpora rather than domain-specific archaeological datasets, these systems frequently reproduce stylistic artefacts inconsistent with Gandhāran sculptural traditions, including symmetrical drapery folds, stylised “snail-curl” hair forms and Petra-like architectural façades. Although expert-guided prompt refinement reduced these distortions, it did not eliminate them.

Reproducibility also remains constrained. Outputs varied over time as generative platforms underwent undocumented model updates, limiting exact replication despite the use of seed locking and metadata capture. This instability underscores the importance of paradata disclosure (Bentkowska-Kafel & Denard, 2012) and model reporting (Mitchell et al., 2019), including prompt structure, seeds and model identifiers, to support partial replication.

Finally, the composition of the expert panels shapes the interpretation of results. Phase I achieved interdisciplinary breadth through 33 international experts, while Phase II provided depth through five Bamiyan specialists. However, Afghan community representatives and Buddhist practitioners were underrepresented and the length of the evaluation instrument may have introduced fatigue effects influencing response consistency.

5.6 Conclusions

This study evaluated the Heritage-Aligned Reconstruction Framework (HARF) through an empirical case study of the Western Buddha of Bamiyan, examining how structured prompting, computational filtering and expert assessment interact within an evidentially constrained reconstruction workflow.

The results demonstrate that computational pre-evaluation can identify outputs with strong geometric correspondence to archival references; however, a consistent divergence was observed between algorithmic ranking and expert judgement. While computational metrics capture structural coherence, they remain insufficient for assessing culturally embedded stylistic and iconographic validity. Archaeological plausibility therefore continues to depend on expert interpretation grounded in domain-specific knowledge.

The findings further indicate that generative outputs are most unstable in morphologically complex regions, particularly drapery articulation and anatomical detail, where structural geometry intersects

with culturally specific visual grammar. These instabilities reflect systemic limitations in the ability of generative systems to reconcile geometric accuracy with historically grounded representation.

Taken together, the results reposition generative AI within heritage reconstruction workflows not as a tool of definitive restitution, but as an exploratory instrument capable of generating and testing visual hypotheses under evidentiary constraint. When embedded within a structured architecture integrating prompt design, computational filtering and expert review, such methods can support iterative interpretation while maintaining methodological transparency. Although the reconstructions presented here do not constitute definitive archaeological restorations, HARF provides a transferable methodological basis for the critical use of AI in the reconstruction of culturally significant heritage.

The evaluation in this chapter establishes the computational and specialist limits of HARF at Bamiyan. It shows that structural similarity and archaeological plausibility are related but not equivalent and that expert assessment remains necessary where stylistic, iconographic and culturally specific features cannot be resolved by metrics. These findings leave a further question unresolved: whether reconstruction candidates that are computationally and archaeologically plausible are also culturally acceptable to communities connected to the site. Chapter 6 addresses that question through community-grounded evidence from Afghan respondents.

Chapter 6

Community-Grounded Authenticity at Bamiyan

Chapter 5 tested HARF at Bamiyan through computational similarity, interdisciplinary screening and specialist archaeological assessment. It showed that computational correspondence and expert plausibility are necessary but not sufficient for evaluating AI-mediated heritage reconstruction. A reconstruction candidate may be structurally filtered and archaeologically assessed while still leaving unresolved questions of cultural meaning, absence, memory and acceptable representation.

This chapter addresses that unresolved register through Afghan respondent evidence. Its role in the thesis is to extend the Bamiyan validation from computational and specialist assessment to community-grounded authenticity. The chapter does not treat the respondent sample as a representative account of all Afghan, Hazara or Buddhist perspectives. It uses the sample as bounded empirical evidence for how reconstruction, loss and AI-mediated representation are understood by respondents connected to the cultural and historical context of Bamiyan. The chapter therefore examines whether AI-mediated reconstruction at Bamiyan also requires provenance visibility, rupture visibility and governance visibility before generated outputs can be treated as culturally accountable.

6.1 Introduction

The Taliban's destruction of the Buddhas of Bamiyan in March 2001 removed two monumental rock-cut sculptures from the cultural landscape of central Afghanistan and transformed the site into a continuing problem of heritage reconstruction, representation and loss. The act was widely treated as a turning point in debates on cultural-heritage protection in conflict (Flood, 2002; Lostal & Vasconcelos Vilaça, 2015). Since then, Bamiyan has been the subject of photogrammetric documentation, long-term conservation and virtual-reality planning, remote-sensing assessment of the wider archaeological landscape and, more recently, generative-AI reconstruction workflows (Arzomand & Kalganova, 2026b; Grün et al., 2004; Hammer et al., 2018; Nagaoka, 2020; Remondino, 2011; Toubekis et al., 2017). Critical heritage scholarship has accompanied this technical record with parallel concern for memory, erasure and representational authority at the site (Connerton, 2008; Flood, 2002; Lostal

& Vasconcelos Vilaça, 2015; Macdonald, 2008; Risam, 2018; Stobiecka, 2020). What remains comparatively thin is primary empirical evidence on how Afghan respondents themselves understand authenticity, loss and acceptable reconstruction at Bamiyan.

This chapter addresses that empirical gap. Where Afghan perspectives appear in the existing literature, they are more often mediated through secondary scholarship or institutional documents than examined through primary data collected from Afghan respondents themselves (Dupree, 2002; Monsutti, 2005). That asymmetry matters because earlier reconstruction work has already identified the under-representation of Afghan and Buddhist-practitioner perspectives as a limitation. The present study extends that research programme by adding the community layer that the earlier studies identified as missing. Earlier work in this research programme established AI-assisted reconstruction methodology and formalised it through the Heritage-Aligned Reconstruction Framework (HARF), with a subsequent paired computational–expert evaluation showing that similarity-based metrics and specialist judgement can diverge on stylistic and iconographic grounds (Arzomand et al., 2024, 2026; Arzomand & Kalganova, 2026b). Computational plausibility does not on its own resolve questions of historical adequacy, particularly where stylistic judgement and community-grounded perspectives remain underrepresented. The present study responds to that limitation directly.

This study advances a focused argument. A reconstruction workflow may be evidentially disciplined and expert-validated, yet remain incomplete if the terms of authenticity, absence and legitimacy are never examined from the perspective of those most closely connected to the site. Drawing on questionnaire responses from sixty-two Afghan respondents, the study shows that support for reconstruction does not equate to support for seamless visual replacement alone. Respondents attached importance not only to material fidelity, but also to historical context, cultural meaning, associated skills and knowledge, selected sensory dimensions and the visibility of rupture. On that basis, the study proposes a community-grounded extension to AI-mediated reconstruction at Bamiyan centred on provenance, rupture and governance. Three questions structure the analysis: how Afghan respondents conceptualise authenticity, loss and acceptable reconstruction at Bamiyan; whether views on reconstruction, sensory inclusion and governance vary across the demographic sub-groups within the sample; and how these findings should inform an AI-mediated reconstruction workflow previously assessed through computational and expert evaluation.

6.2 Bamiyan in the Literature: Post-2001 Documentation and the Question of Authenticity

The Bamiyan case is well documented technically, but less well examined through primary respondent evidence. The Buddhas were monumental rock-cut sculptures dating to the sixth- to seventh-century CE set within a wider Buddhist monastic complex along the Silk Road and surrounded by hundreds of decorated meditation cave structures and mural remains (Grün et al., 2004; Klimburg-Salter, 1989; Petzet, 2009). The valley's contemporary population is predominantly Hazara, a Shia Muslim ethnic group concentrated in the central highlands of Afghanistan, including Bamiyan, Daikundi, Ghazni and Samangan (Monsutti, 2005). Hazara communities have lived in the valley for centuries, cultivated the land around the cliffs and continued to act as residents and custodians of the wider site through successive political regimes, while also maintaining their own reinterpretation of the statues' identity. Indeed, in Hazara oral traditions, the statues in Bamiyan are not understood solely as Buddhas but are woven into local legends and associated with heroic figures known as Salsal and Shahmama, tied to the region's cultural landscape (Singh, 2022). Hazara have also been subject to repeated persecution, including during the 1990s Taliban advance on the central highlands and again after 2021. Acknowledging the Hazara presence in the context of Bamiyan reconstruction is therefore not identity-labelling but a substantive point about custodianship and about which community bears the local consequences of reconstruction decisions taken elsewhere.

Post-2001 responses to the site have been technically diverse and institutionally distributed. Photogrammetric reconstruction of the Great Buddha was carried out by Grün and colleagues at ETH Zurich on the basis of pre-destruction imagery, establishing the principal dimensional record on which subsequent work has drawn (Grün et al., 2004). Long-term conservation and virtual-reality planning have been led by Toubekis, Jansen and the RWTH Aachen team in cooperation with ICOMOS Germany, supporting evaluation of physical-restoration options for the niches and surrounding fabric (Petzet, 2009; Toubekis et al., 2017). Remote-sensing assessment of the wider Bamiyan cultural landscape has been undertaken by Hammer and collaborators, extending the documentary record beyond the niches themselves (Hammer et al., 2018). Earlier Japanese cultural missions in the 1970s produced measured drawings and architectural surveys that remain the most precise records of niche geometry. These institutional efforts converge on documentation and physical-restoration planning rather than on community consultation. The 2017 UNESCO symposium, whose proceedings were edited by Nagaoka (Nagaoka, 2020), presents positions ranging from anastylosis and partial reconstruction to non-reconstruction, alongside debates about what authenticity means when the original fabric is almost entirely lost (Holtorf, 2020; Jokilehto, 2020; Rössler, 2020). The consensus that emerges is

methodologically instructive and substantively limited: it confirms that authenticity at Bamiyan cannot be reduced to material fidelity alone, but it is authored almost entirely by international experts and institutions rather than by Afghan respondents. Table 16 summarises these post-2001 efforts and locates the contribution of the present study in relation to them.

Table 16. Post-2001 reconstruction approaches to the Bamiyan Buddhas, the institutional teams involved and what each approach addresses and does not address.

Approach	Centre / lead team	Date	Output Type	Addresses	Doesn't Address
Photogrammetric reconstruction	ETH Zurich (Grün et al.)	2004	Dimensional model of Great Buddha	Geometry, proportion, scale	Community meaning; absence; sensory layer
Long-term conservation and VR planning	RWTH Aachen / ICOMOS Germany (Petzet, Toubekis, Jansen)	2009–2017	VR environment; restoration scenarios	Physical-restoration evaluation; niche fabric	Community consultation; oral testimony
Remote-sensing landscape assessment	Hammer et al.	2018	Damage and landscape mapping	Wider archaeological landscape	Site-specific representational ethics
UNESCO symposium synthesis	UNESCO / Nagaoka (Ed.)	2020	Edited volume, multi-author	Authenticity debates, restoration options	Primary Afghan respondent data
Generative AI reconstruction (HARF)	This thesis	2024–2026	AI-generated imagery, expert-validated	Visual hypothesis generation, evidentiary discipline	Community-epistemic layer

The second body of literature concerns authenticity in heritage studies. The Venice Charter (ICOMOS, 1964) defines restoration as preserving and revealing aesthetic and historic value while distinguishing original material from reconstructive intervention. The Nara Document on Authenticity (ICOMOS, 1994) widens authenticity beyond material fabric to include form and design, use and function, traditions and techniques, location and setting and spirit and feeling, while rejecting any single universal standard across cultures. The London Charter (Denard, 2009) and the Seville Principles (López-Menchero Bendicho & Grande, 2011) translate these obligations into computer-based heritage visualisation by requiring evidentiary grounding, paradata and explicit marking of inference and uncertainty. Across the heritage-studies literature that runs alongside these standards, authenticity is increasingly treated not as a property inhering in an object, but as a relation between people, objects

and places that must be negotiated rather than simply inspected (Harrison, 2012; Jones, 2010; Kirshenblatt-gimblett, 2004; Silberman, 2012; Smith, 2006). This echoes M. Forte's elaboration of Gibson's concept of affordances, according to which cultural heritage environments afford multiple possibilities for social perception, interaction and cultural interpretation (Forte, 2024).

Recent scholarship has distinguished several senses of authenticity that operate at this site, including historical, material, experiential, community and procedural-transparent authenticity (Jones, 2010; Silberman, 2012). This study adopts the relational sense established in the Nara Document and developed by Jones (Jones, 2010) and uses the term community-grounded authenticity as shorthand for authenticity constituted through the relations that descendant and custodial communities maintain with a heritage site and that therefore cannot be adjudicated without their participation (ICOMOS, 1994; Jones, 2010). Using this as the master term keeps the analysis on a single relational concept rather than across the neighbouring concepts of legitimacy, epistemic authority and representational accountability, which are treated here as components of community-grounded authenticity rather than as rival organising terms.

The empirical gap is narrow but important. No published study of the Bamiyan case treats Afghan respondents themselves as the epistemic source on what authenticity requires for this site. The preceding empirical work in this thesis examined Bamiyan through computational and expert registers: a thirty-three-expert interdisciplinary panel and a paired evaluation by five Bamiyan specialists (Arzomand et al., 2026; Arzomand & Kalganova, 2026b). Those stages establish evidential discipline and specialist review, but they leave unresolved the community register of authenticity. This chapter therefore adds the respondent layer required for a community-grounded account of authenticity.

6.3 Methodology

This study is based on an online questionnaire titled *Memory, Absence and Digital Reconstruction of Lost Cultural Heritage in the Bamiyan Valley*, approved by the Brunel University of London Research Ethics Committee under reference 50187-LR-Jan/2025-53624-2. The participant information sheet stated the study's purpose, expected completion time, the voluntary nature of participation, the right to withdraw up to the point of submission, data-handling procedures and the researcher's contact details. Consent was indicated by continuing past the consent page. Participation was voluntary and informed. Identifying contact details collected for recruitment, correspondence or consent administration were stored separately from the analytical dataset and were not used in the analysis. The analysed dataset contains demographic categories, country-level location information,

timestamps and questionnaire responses. Findings are reported only in anonymised or aggregated form. The full ethics statement, consent procedure, anonymisation measures and quotation policy are provided in Appendix D, Section S1.

The instrument comprised forty-six items organised in six thematic blocks: demographic context; connection and memory; interpretations of destruction; reconstruction preferences and ethical considerations; intangible and sensory dimensions; and open reflections. Items combined single-select, multi-select and short open-ended formats. The instrument was refined through three pilot rounds ($n = 5$, $n = 4$, $n = 3$; where n denotes the number of respondents in each pilot round) with Afghan pilot respondents drawn from diaspora academic networks. Pilot feedback was used to refine wording and remove double-barrelled items. The questionnaire was provided in English, with a Dari/Farsi and Pashto translation available on request and quoted open-ended responses were either submitted in English or translated by the first author with back-translation checks conducted on a subset. The full instrument is reproduced in Appendix D, Section S2.

Recruitment used purposive and snowball sampling through Afghan diaspora community organisations, academic networks in the UK, Europe and North America and local contacts inside Afghanistan. The resulting sample ($N = 62$, where N denotes the total number of respondents) is a convenience sample of Afghan respondents reachable through those channels, not a probability sample of the Afghan population. A goodness-of-fit comparison of the named-ethnicity portion of the sample against widely cited estimates of Afghanistan's ethnic composition returned $\chi^2(2) = 10.49$, $p = .005$, with Hazara respondents over-represented relative to the wider population. This is consistent with the study's substantive focus on the Bamiyan Valley, but it confirms that the sample should not be treated as nationally representative; the analysis therefore reports within-sample patterns and does not generalise to Afghan respondents more broadly. Full demographic breakdown, current-location detail and sample-composition analysis are provided in Appendix D, Section S3 and Tables S1–S5, while the population-comparison exercise is reported in Appendix D, Section S7 and Table S17.

Closed items are reported as counts and proportions with Wilson 95% confidence intervals. Multi-select items report selection counts that exceed N by design and are interpreted as overlapping endorsements rather than mutually exclusive categories. Broad subgroup contrasts on core reconstruction and governance items are reported with chi-square or Fisher tests and bias-corrected Cramér's V , but the sample size and thinness of several demographic cells do not support regression modelling or fine-grained three-way contrasts. Open-ended responses were analysed thematically (Braun & Clarke, 2006). Coding proceeded in two passes: an initial inductive pass to generate candidate themes, followed by a second pass that cross-referenced those themes against the closed-item findings. Coding detail, including the blinded subset check, is provided in Appendix D, Section S9.

Quotations are presented only in anonymised form and are not accompanied by combined ethnicity-by-gender-by-country identifiers, because cell sizes at those intersections are too small to permit such attribution without re-identification risk. Full descriptive tables, subgroup contrasts, the Hazara subsample breakdown, thematic coding detail and reproducibility materials are provided in Appendix D, Sections S4-S6 and S8-S10, including Tables S6–S19 where relevant.

6.4 Results

Table 17 reports the headline closed-item findings. Support for reconstruction was high, although respondents did not equate acceptable reconstruction with visual replacement. Material fidelity, cultural connection, physical reconstruction, sensory inclusion, hybrid past-present representation of absence and concern about AI-mediated oversimplification of cultural meaning were all selected by majorities or near-majorities. Counts are out of $N = 62$ unless otherwise stated; multi-select totals exceed N by design and full descriptive tables are reproduced in Appendix D, Section S4-S5 and Tables S6–S15.

Table 17. Headline proportions with Wilson 95% confidence intervals ($N = 62$).

Item	n	Proportion	95% CI
Support reconstruction (Yes)	51	82.3%	71.0–89.8%
Material fidelity prioritised (multi-select)	52	83.9%	72.8–91.0%
Cultural connection selected (multi-select)	48	77.4%	65.6–86.0%
Physical reconstruction preferred (multi-select)	49	79.0%	67.4–87.3%
Sensory elements should be included (Yes)	38	61.3%	48.8–72.4%
Hybrid past-present representation of absence	37	59.7%	47.3–71.0%
AI concern: oversimplification of cultural meaning	36	58.1%	45.7–69.5%
Hazara voices perceived as included	19	30.6%	20.6–43.0%
Hazara voices: don't know	31	50.0%	37.9–62.1%

Ethnic self-identification comprised Pashtun (26, 41.9%), Hazara (13, 21.0%), Tajik (12, 19.4%), Arab (2) and Sadaat (1); six respondents indicated that ethnicity did not matter to them and two responses were missing. The age distribution was 21–35 (35), 36–50 (18), over 50 (8) and under 20 (1); gender, where reported, was male (43) and female (18), with one unspecified response. Twenty-five respondents resided in Afghanistan and thirty-six across the diaspora, with country of residence not

reported by one respondent. Pre-destruction familiarity with the statues was almost universal, with 59 of 62 respondents reporting prior knowledge of the site.

Multiple, overlapping modes of connection were reported (Figure 40). Cultural connection through stories, ancestry and traditions was the dominant mode (48, 77.4%; 95% CI 65.6–86.0%), followed by emotional connection (24, 38.7%), personal connection through family history or visits (18, 29.0%), political connection (15, 24.2%) and spiritual connection (10, 16.1%). Three respondents reported no strong connection to the site. Intergenerational transmission was substantial: 33 of 62 respondents (53.2%) stated that their family or community had shared stories about the statues and 24 reported continuing to remember those stories in detail. Pre-destruction familiarity was acquired predominantly through schooling, family narrative and media exposure and several open-ended responses referred to the statues by their local names *Salsal* and *Shahmama*, indicating that intergenerational naming practice survives even among respondents who learned of the site only after its destruction. The statues are accordingly framed within the sample as inheritances of cultural and civilisational identity, with devotional registers occupying a clear minority position.

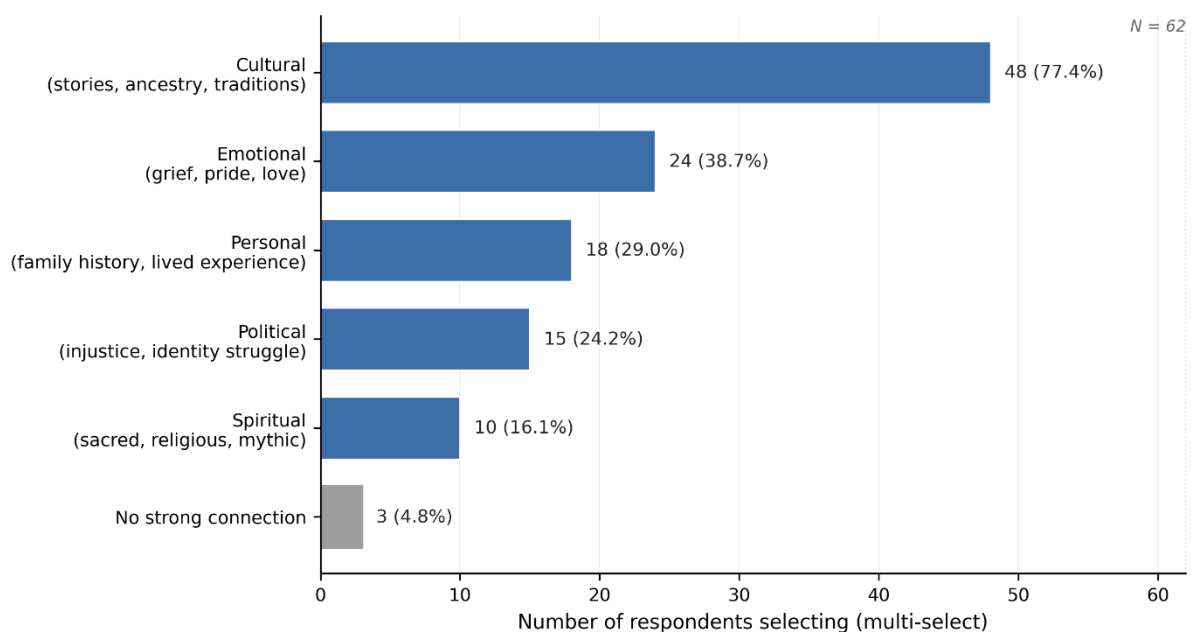


Figure 40. Author-generated chart showing the modes through which respondents reported a connection to the Bamiyan Buddhas (multi-select; *N* = 62). Respondents could select more than one mode, so totals exceed *N* by design. Counts and percentages are shown alongside each bar.

Open-ended responses interpreted the 2001 destruction as politically motivated. The act was repeatedly characterised as a message directed at the international community, as targeted violence against the Hazara population and as the imposition of an exclusionary ideology. A minority of responses described the religious justification advanced at the time as a misinterpretation of Islamic scholarly tradition. The present research does not adjudicate that theological dispute and confines its

claim to the empirical observation that respondents did not treat the religious framing as settled. The interpretive pattern is consistent with academic accounts of the destruction as political signalling, in which the 2001 act is read against the longer history of contested heritage and political erasure (Flood, 2002; Lostal & Vasconcelos Vilaça, 2015). Support for reconstruction was 51 of 62 respondents (82.3%; 95% CI 71.0–89.8%), with 6 uncertain (9.7%) and 5 opposed (8.1%) (Figure 41A). The substantive finding lies in the operational meaning that respondents attached to "reconstruction", which extended well beyond visual completeness (Figure 41B). Multi-select responses on what should be reconstructed prioritised material fidelity to the original (52, 83.9%); selection of non-material aspects was substantial, comprising the skills and knowledge associated with the statues' creation (24, 38.7%), the destruction and its historical context (21, 33.9%) and the meaning and symbolism of the statues (20, 32.3%); four respondents indicated that nothing should be reconstructed. Modal preference was physical reconstruction (49, 79.0%), followed by digital imagery (19, 30.6%), VR/AR (18, 29.0%) and holograms (14, 22.6%); 11 respondents recorded no preference. The sample therefore treats physical reconstruction as the primary modality and digital reconstruction as complementary. Governance preferences were distributed across multiple actors: UNESCO-style organisations (39, 62.9%), Afghan archaeologists and cultural historians (37, 59.7%), heritage scholars (26, 41.9%), the Afghan government (25, 40.3%), the local Hazara community (18, 29.0%), all parties equally (16, 25.8%) and Buddhist monks (13, 21.0%). Guiding principles were headed by historical accuracy (52, 83.9%) and economic considerations including tourism (40, 64.5%); cultural resistance or protest (28, 45.2%), community healing (24, 38.7%) and memorialisation of trauma (18, 29.0%) followed.

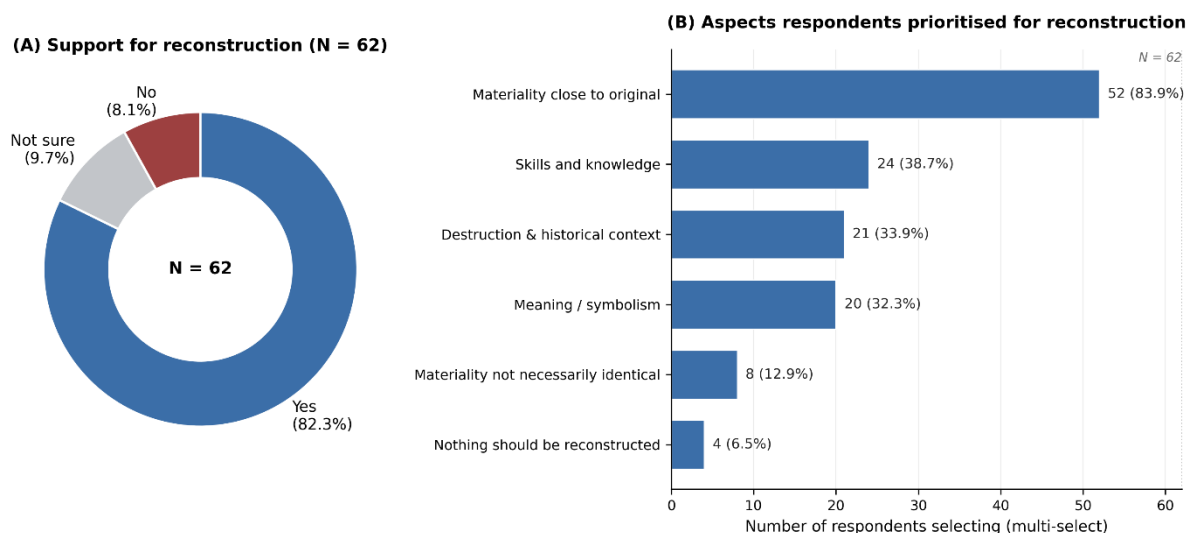


Figure 41. Author-generated chart showing reconstruction support and aspects respondents prioritised for reconstruction (N = 62). Panel A shows responses to the single-select item on whether the Buddhas should be reconstructed. Panel B shows multi-select responses on which aspects should be reconstructed; respondents could select more than one aspect, so totals exceed N by design. Counts and percentages are shown alongside each bar.

On the representation of destruction in future reconstructions, the modal selection was a hybrid past-present representation, with holograms given only as an example in the questionnaire item (37, 59.7%); full digital restoration was the second-ranked option (30, 48.4%); a smaller group endorsed leaving the ruined niches untouched (8, 12.9%); three respondents supported no reconstruction at all. The pattern is congruent with the Venice Charter's requirement that restoration distinguish original material from reconstructive intervention (ICOMOS, 1964) and with critical-heritage scholarship that treats absence and rupture as historically consequential conditions (Connerton, 2008; Macdonald, 2008; Rico, 2016).

Thirty-eight respondents (61.3%; 95% CI 48.8–72.4%) endorsed the inclusion of sensory elements in digital reconstructions; 17 were uncertain and 7 opposed. Asked to prioritise a single sensory modality, respondents ranked light and shadow first (27, 43.5%), followed by scent of earth (13), echo of prayer (12) and wind in the valley (6). On the content of digital reconstructions, accurate visual appearance was selected most frequently (53, 85.5%), followed by the caves and murals (31, 50.0%), Buddhist spiritual symbolism and economic impact (26 each, 41.9%), Hazara oral histories (22, 35.5%), environmental and sensory experiences (15, 24.2%), ruins and trauma of destruction (13, 21.0%), the Taliban's justification and international reactions (11, 17.7%) and soundscapes (8, 12.9%). The dominant AI-related concern was oversimplification or misrepresentation of cultural meaning (36, 58.1%), with secondary concerns about the marginalisation of local voices (19, 30.6%), fabrication of history (18, 29.0%) and the imposition of Western aesthetic defaults (13, 21.0%); 16 respondents reported no concerns. The salience of meaning-based and voice-based concerns over technical-failure concerns is consistent with critiques of digitally mediated heritage representation as a site of epistemic and representational asymmetry (Risam, 2018; Stobiecka, 2020).

On the perception of Hazara voice inclusion in current reconstruction discussions, 19 respondents (30.6%) answered yes, 12 (19.4%) answered no and 31 (50.0%) did not know. Within the Hazara sub-sample ($n = 13$), opinion was evenly divided. The "don't know" share is the most substantive feature of this distribution: the procedural visibility of community involvement is opaque to half the sample, an opacity that bears directly on participation arguments in community-based heritage practice (Acabado & Martin, 2020; Atalay, 2012). A more detailed breakdown of the Hazara sub-sample is reported in Appendix D, Section S8 and Table S18.

Table 18 reports sub-group contrasts on the core items, including ethnicity, gender, location and age band. No contrast reached $\alpha = .05$ and bias-corrected Cramér's V remained small. The non-significant results have limited statistical weight. The tests are unadjusted, the total sample is $N = 62$ and some cells are small, including the Hazara subgroup ($n = 13$). The absence of significant differences is

therefore not used to claim that the groups held the same views. It is reported only as the pattern found in this dataset. The complete contrast set is provided in Appendix D, Section S6 and Table S16.

Table 18. Sub-group contrasts on core items

Comparison	Test	Statistic	p	Cramér's V
Reconstruction support × Ethnicity (Pashtun / Hazara / Tajik)	χ^2	$\chi^2 (4) = 5.42$	.247	0.12
Reconstruction support × Gender	χ^2	$\chi^2 (2) = 0.36$	.835	0.00
Reconstruction support × Location (inside / diaspora)	χ^2	$\chi^2(2) = 2.25$	.325	0.06
Reconstruction support × Age band	χ^2	$\chi^2 (6) = 3.20$	.784	0.00
Sensory inclusion × Ethnicity	χ^2	$\chi^2 (4) = 6.51$	.164	0.16
Hazara-voice perception × Ethnicity	χ^2	$\chi^2 (4) = 4.34$	.361	0.05
Hazara-voice perception (Yes vs No/DK) × Location	Fisher exact	—	.092	—

Note: No contrast reached $\alpha = .05$. χ^2 = Pearson chi-square statistic; p = p-value; V = bias-corrected Cramér's V effect size. Fisher exact tests are reported where expected cell counts were too small for chi-square.

Three negative findings should be stated explicitly. The data do not support the claim that respondents differ substantively by ethnicity on the core reconstruction-ethics items; the convergence reported above is consistent across the contrasts tested. The data do not support a uniform rejection of digital reconstruction by Afghan respondents; what the data support is qualified endorsement, conditional on transparency and community inclusion. And the data do not support a preference for digital over physical modalities: the reverse holds in this sample, with physical reconstruction the modal preference (79.0%) and digital imagery a complementary rather than substitutive option (30.6%).

6.5 A design orientation for the AI-mediated workflow

The findings do not warrant a new heritage-reconstruction framework. They support three design principles that extend the AI-mediated workflow. These principles are offered as orientation for practice, not as audited criteria and remain confined to what the survey directly supports. They sit within obligations already articulated in the Venice Charter, the Nara Document, the London Charter and the Seville Principles.

The first concerns provenance visibility. Material fidelity and historical accuracy were each prioritised by 83.9% of the sample, while concerns about AI-mediated oversimplification of meaning (58.1%) and fabrication of history (29.0%) were also substantial. Reconstructions should therefore make their evidentiary basis visible at the level of the output, indicating which source underlies which element and whether that element is documented, inferred, or speculative. This extends the Venice Charter's requirement that reconstruction remain distinguishable from original material (ICOMOS, 1964) translated into digital visualisation through the London Charter's paradata requirement (Denard, 2009) and the Seville Principles' documentation logic (López-Menchero Bendicho & Grande, 2011). Within HARF, the principle is implementable through the Dynamic Prompt Blueprint's evidential slot structure, with those anchors surfaced in the published reconstruction record (Arzomand et al., 2026).

The second concerns rupture visibility. The modal preference for representing absence was a hybrid past-present output (59.7%); 12.9% preferred the ruined niches to remain untouched and a small minority supported no reconstruction at all. Reconstructions that present only the pre-2001 state do not align with this preference and the implication is that reconstruction outputs should be published alongside the documented post-2001 condition rather than in place of it. This is consistent with heritage scholarship that treats absence and rupture as historically consequential conditions rather than defects to be erased (Connerton, 2008; Harrison, 2012; Rico, 2016). Within the existing workflow, this can be implemented as a dual-view publication or delivery specification.

The third concerns governance visibility. Only 30.6% of respondents perceived Hazara voices as included in reconstruction discussions, while 50.0% did not know. Governance preferences were distributed across international organisations, Afghan experts and local-community actors rather than concentrated in any single authority. AI-mediated reconstruction should therefore declare, in its published record, which parties were involved in evidential, interpretive and curatorial decisions and which were not. This transparency obligation extends the London Charter's paradata requirement (Denard, 2009) and the Seville Principles' documentation conventions (López-Menchero Bendicho & Grande, 2011) and aligns the disclosure of reconstruction governance with community-based archaeology's concern for whose knowledge counts as evidence and whose absence remains unmarked (Acabado & Martin, 2020; Atalay, 2012).

Three further observations follow from the findings but are not elevated to principles. First, the sample gave substantial weight to economic and livelihood dimensions of reconstruction, with 64.5% endorsing tourism and economy as a guiding principle. Reconstruction outputs may therefore be judged partly through utility criteria that remain underdeveloped in heritage-ethics scholarship. Second, support for sensory inclusion was present but not overwhelming and light and shadow were prioritised above acoustic and olfactory elements. Any expansion into multisensory reconstruction

should follow that ordering rather than assume that sound or atmosphere are the primary sensory additions (Bao & Bowen, 2025; A. Bender et al., 2023; Kirshenblatt-gimblett, 2004). Third, these principles should not be read as community validation of AI-mediated reconstruction in general. A majority of respondents expressed at least one substantive concern about AI mediation, while roughly a quarter reported no concerns. The principles indicate the conditions under which AI-mediated reconstruction may become more acceptable to respondents who already hold reservations; they do not establish that such acceptance has been achieved.

6.6 Discussion, Limitations and Future Work

Three findings structure the discussion. First, support for reconstruction in some form was strong and widely distributed within the sample. Second, respondents used the term reconstruction in a broader sense than visual replacement alone, attaching importance to skills, knowledge, historical context, cultural meaning, sensory elements and the visible marking of destruction. Third, the community layer around reconstruction was perceived as unevenly included, with half the sample uncertain whether Hazara voices were represented in current discussions. The sub-group comparisons did not identify significant variation across the demographic categories tested. The principal patterns reported here are therefore presented as descriptive patterns within the respondent group, rather than as findings confined to one ethnic, gendered, age-based, or locational subgroup.

Across the preceding empirical chapters, exploratory text-to-image generation established the feasibility of historically informed AI reconstruction; HARF formalised the evidentiary discipline required for that process within a structured expert-guided workflow; and the paired computational–expert evaluation on the Western Buddha showed that computational similarity and specialist judgement can diverge on stylistic and iconographic grounds, with community under-representation identified as an outstanding limitation (Arzomand et al., 2024, 2026; Arzomand & Kalganova, 2026b). The present study addresses that limitation by identifying, in empirical terms, what the community layer would need to accommodate. The design orientation proposed in the previous section therefore extends the earlier workflow by specifying how its published outputs should register provenance, rupture and participation.

The limits of the study arise from the case and from the access conditions under which research on it must now be conducted. The sample is a convenience sample and it should not be read as statistically representative of Afghan opinion as a whole. Participation from within Afghanistan was also constrained by the present political and security environment. Some potential respondents, particularly those residing under the current regime, were unwilling to participate or to answer openly

because of perceived personal risk. That restriction is methodologically important: it likely affected both response rates and the degree of candour available from respondents inside Afghanistan. The questionnaire was administered in English, with translation available on request, which also limited reach even with piloting and translation support. It also shaped the form in which the headline findings were produced: concepts such as authenticity, absence, reconstruction and digital reconstruction were presented through an English-language instrument, although respondents could answer open-ended questions in their own language and one Dari response was translated for analysis.

The sample size supports descriptive analysis and cautious subgroup contrasts, but not finer-grained modelling or stable interaction analysis. The measure on Hazara inclusion captures respondent perception, not an external audit of actual participation in reconstruction projects. The study also reports respondents' characterisation of the religious framing of the destruction without adjudicating the theological dispute itself. These limits define the scope of the findings, but they do not remove the value of the empirical layer the study provides.

The most important methodological implication is that the survey establishes distribution and pattern, but not the depth of explanation that follow-up interviews could provide. The next step is therefore not a broader theoretical claim, but a second method. Semi-structured interviews with Afghan heritage practitioners, Bamiyan residents and diaspora community representatives would allow the preferences reported here to be tested against more detailed accounts of memory, legitimacy and acceptable reconstruction practice. The methodological basis for such follow-up work is well established in community-based heritage scholarship (Acabado & Martin, 2020; Atalay, 2012; Cusack-McVeigh, 2014).

Two further directions follow. A language-stratified replication of the questionnaire, administered in Dari, Pashto and Hazaragi with more deliberately stratified recruitment, would improve coverage and reduce dependence on English-mediated participation. A second line of work would test the design orientation against a concrete HARF reconstruction run and document what changes in the published reconstruction record when provenance visibility, rupture visibility and governance visibility are explicitly applied. That would provide a stronger methodological test than a survey-based design study can offer on its own. Reconstruction at Bamiyan does not occur against a neutral background. It intervenes in a field already shaped by deliberate destruction, contested memory and uneven participation.

6.7 Conclusion

This chapter examined how Afghan respondents sampled across Afghanistan and the diaspora understand reconstruction, authenticity, absence and participation in relation to the Buddhas of Bamiyan. The findings show that support for reconstruction is strong, but that respondents do not treat reconstruction as a question of visual replacement alone. Historical accuracy, cultural meaning, associated skills and knowledge, sensory dimensions and the visible marking of destruction all form part of the operative meaning of acceptable reconstruction. The study also shows that uncertainty around whose voices are included in reconstruction discussions remains substantial.

The contribution of this study is bounded. It does not propose a new reconstruction framework or claim to resolve the ethics of reconstruction at Bamiyan. It identifies a missing empirical layer in the existing literature and turns the findings into design recommendations for an existing AI-mediated workflow. In doing so, it argues that provenance visibility, rupture visibility and governance visibility should be treated as minimum conditions for publishing AI-mediated reconstruction outputs at a contested site.

This completes the Bamiyan validation across computational, specialist and community-grounded registers. Chapter 5 showed that computational similarity and expert archaeological judgement do not measure the same thing. This chapter has shown that expert-assessed plausibility also does not settle cultural acceptability. Chapter 7 therefore moves beyond Bamiyan to test whether HARF can be re-instantiated for a different heritage typology at Palmyra, without treating Palmyra as a second full validation case.

Chapter 7

Bounded Cross-Site Generalisation of the Temple of Bel, Palmyra

Chapters 5 and 6 provided the deepest validation of HARF at Bamiyan, first through computational and specialist evaluation and then through community-grounded authenticity. The remaining question is whether the framework can be re-instantiated beyond the case through which it was developed. This chapter tests that question through the Temple of Bel at Palmyra.

Palmyra is used because it changes the typological conditions of evaluation. Bamiyan concerns a figural Gandhāran monument; Palmyra concerns monumental architecture with different evidentiary sources, expert criteria and likely failure domains. The chapter therefore tests transferability at the level of evaluative schema, not historical accuracy of the generated outputs. It keeps the top-level HARF structure stable while replacing Bamiyan-specific evidence, prompt content and evaluative criteria with Palmyra-specific material. It does not treat Palmyra as a second full validation case.

7.1 Introduction

The deliberate destruction of cultural heritage during armed conflict has placed sustained pressure on digital reconstruction to develop methods that are both technically reliable and ethically accountable (Brosché et al., 2017; Cunliffe et al., 2016). Among the most consequential losses of the twenty-first century are the sixth- to seventh-century CE Buddhas of Bamiyan, destroyed in March 2001 (Flood, 2002) and the Temple of Bel at Palmyra, intentionally destroyed with explosives in August 2015 (Schmidt-Colinet, 2019). Digital reconstruction methods have been applied to Palmyra through visual reconstruction, photogrammetry, crowdsourced imagery and immersive 3D environments (Rihani, 2023; Silver et al., 2018). Generative-AI reconstruction raises a different evaluation problem: its outputs may appear visually coherent while remaining archaeologically uncertain. The methodological issue is therefore not only how to generate plausible images, but how to distinguish visual plausibility from evidential plausibility (Arzomand et al., 2024, 2026; Arzomand & Kalganova, 2026b).

The Heritage-Aligned Reconstruction Framework (HARF) was introduced to address that gap (Arzomand et al., 2026). HARF integrates a Dynamic Prompt Blueprint, a Prompt Sufficiency Index,

computational pre-evaluation and expert-in-the-loop assessment within an evidentially constrained workflow. Chapter 5 evaluated HARF on the Western Buddha of Bamiyan and reported a structural divergence between algorithmic similarity scoring and expert archaeological judgement, with the most unstable reconstruction domains concentrated in drapery articulation, cranial morphology and hand gesture (Arzomand & Kalganova, 2026b). The Bamiyan validation, however, leaves one substantive question open. A framework calibrated on a figural Gandhāran sculpture may behave differently on a heritage site of an entirely different typological order and this has not yet been examined empirically.

This chapter reports on a bounded cross-site evaluation of HARF on the Temple of Bel at Palmyra. The contrast with Bamiyan is deliberate. Bamiyan is a figural sculpture in mud-straw plaster on a sandstone cliff, governed by Gandhāran iconographic conventions; Palmyra is a Graeco-Roman temple in locally quarried limestone, governed by classical orders and an architrave with relief frieze (Arzomand et al., 2024; Seyrig et al., 1975). Each typology stresses different aspects of a reconstruction framework. The exercise reported here is questionnaire-based: it tests whether the framework’s schema and evaluative logic transfer, not whether AI outputs constitute defensible reconstructions of the temple itself. Specialist Palmyrene archaeology and debates on post-destruction preservation and reconstruction are addressed elsewhere (Abdulkarim & Seigne, 2024; Munawar, 2017; Schmidt-Colinet, 2019). Syrian and Palmyrene community perspectives are not examined empirically in this study and are treated as a limitation.

This study makes three contributions. First, it provides framework-level generalisation evidence by showing that HARF’s top-level schema and evaluative logic can be instantiated for a heritage typology fundamentally different from its original Bamiyan validation case without redesigning the framework architecture. Secondly, it identifies a typology-dependent instability pattern in generative reconstruction: when the framework is held constant and the heritage typology changes, the dominant failure domains shift from figural micro-grammar to architectural and decorative micro-grammar. Thirdly, it clarifies what HARF adds beyond existing evaluation approaches. Photogrammetric validation, visual plausibility assessment and computational similarity pipelines can assess geometry, appearance, or structural correspondence, but they do not by themselves explain whether an evidence-constrained generative-AI evaluation framework remains stable across heritage typologies, or where expert review is likely to expose evidential instability. What is newly shown here is that HARF transfers at the level of schema and evaluative logic, while historical plausibility remains site-dependent. The contribution is therefore not a new reconstruction of the Temple of Bel, but a set of principles for evaluating when, how and why generative-AI reconstruction frameworks transfer across heritage typologies and where their outputs require expert intervention.

7.2 Methodology

7.2.1 Adapting the Dynamic Prompt Blueprint to Palmyra

The Dynamic Prompt Blueprint is structured as a tiered slot schema with three levels: core structural anchors, auxiliary contextual descriptors and technical control parameters (Arzomand et al., 2026). The schema is intended to remain stable across heritage contexts; only its slot contents are expected to change with the case. To test that property, the slot contents were re-populated for the Temple of Bel using UNESCO records, the principal Temple of Bel architectural study and relevant digital-reconstruction literature (Schmidt-Colinet & Seigne, 2022; Seyrig et al., 1975; UNESCO, 2026; Wahbeh et al., 2016). Bamiyan-specific descriptors such as mud-straw plaster, Gandhāran wet-fold drapery and Abhaya *mudrā* were replaced by limestone construction, fluted Corinthian columns of recorded height, a pseudoperipteral colonnade, a *cella* of recorded dimensions, an architrave frieze depicting Palmyrene deities, two interior *thalamoi* with zodiac ceiling reliefs and Palmyrene–Aramaic inscription content. Negative-prompt content was reformulated to exclude Petra-like façades and undocumented decorative elements and to mark uncertain roofline treatment as interpretive in line with the recent specialist reconsideration of the temple’s crenellations (Schmidt-Colinet & Seigne, 2022). The transferability test is therefore not whether a fixed prompt can be reused, but whether a fixed evaluative schema can be re-instantiated when the evidential content changes. The DPB supports transferability because it separates the stable structure of evaluation from the variable content of the case. Core anchors, contextual descriptors, negative constraints and paradata remain constant as categories, while their contents change according to the monument or artefact under examination (Arzomand et al., 2026; Denard, 2009). In the Palmyra case, transferability is assessed by whether those categories remain usable when the object changes from a figural sculpture to an architectural monument. Table 19 reports the resulting cross-site instantiation. The schema, the tier structure and the paradata-logging conventions were not altered.

Table 19. Cross-site instantiation of the Dynamic Prompt Blueprint for the Temple of Bel, with Bamiyan referents shown for comparability. The schema is preserved; the slot contents are replaced with site-specific descriptors drawn from documentary sources.

Tier	Slot	Bamiyan referent	Temple of Bel instantiation
Core anchor	Monument identity	Western Buddha (Salsal)	Temple of Bel, Palmyra
Core anchor	Period	6th-7th century AD	1st century AD; consecrated AD 32
Core anchor	Geometry	Statue ~55 m, niche-aligned	Cella 39.45 × 13.86 × 14 m; 41 Corinthian columns at 15.81 m, 1.33 m base

Core anchor	Material	Mud-straw plaster on sandstone	Locally quarried limestone; fluted Corinthian columns with gilded capital treatment where documented
Auxiliary	Figural / architectural context	Standing Buddha, Abhaya mudrā, wet-fold drapery	Pseudoperipteral colonnade, fluted shafts, architrave frieze, monumental staircase, elevated podium
Auxiliary	Environmental setting	Sandstone cliff, monastic caves, Bamiyan Valley	Palmyrene urban sanctuary in oasis landscape; Great Colonnade and civic context
Auxiliary	Symbolic content	Indo-Hellenistic Buddhist iconography, halo, lotus	Palmyrene–Aramaic inscriptions; relief friezes of Palmyrene deities; ceiling reliefs of seven planets and twelve zodiac signs
Technical	Negative prompts	Exclude Petra-like façades, modern materials, anachronistic vegetation	Render roofline only where documented; mark uncertain roofline treatment as interpretive; exclude Roman-imperial conflation
Technical	Paradata	Seed, model, prompt version, refinement cycle	Seed, model, prompt version, refinement cycle

7.2.2 Evaluation Instrument

The evaluation instrument was designed in response to a gap in existing digital-heritage evaluation practice. Photogrammetric and image-based reconstruction workflows can assess spatial correspondence and geometric reconstruction quality, but they do not address the specific problem posed by generative AI: visually coherent outputs may remain archaeologically or culturally unsupported (Rihani, 2023; Wahbeh et al., 2016). Digital-heritage charters emphasise transparency, paradata and the distinction between documented evidence and interpretive reconstruction, but they do not provide a site-transfer test for generative-AI outputs (Denard, 2009; López-Menchero Bendicho & Grande, 2011). Recent benchmark work on historical and cultural artefacts likewise confirms the need for structured, expert-verified evaluation rather than reliance on output plausibility alone (Ghaboura et al., 2025). The Palmyra instrument therefore assesses framework behaviour rather than final reconstruction validity: whether HARF’s prompt architecture, prompt-sufficiency procedure, paradata record and expert-review process can identify the domains in which generative outputs remain evidentially unstable.

A fourteen-item Palmyra block was embedded within the Phase I expert questionnaire used for the Bamiyan study (Arzomand et al., 2026; Arzomand & Kalganova, 2026b). The block was preceded by a structured monument briefing reproducing the temple’s principal dimensions, its architectural and artistic style and its cultural and religious significance, so that respondents could rate against a defined evidentiary baseline rather than against personal knowledge alone (UNESCO, 2026).

Six items elicited five-point Likert ratings (5 = Highly Accurate/Highly Sufficient; 4 = Mostly; 3 = Moderately; 2 = Slightly; 1 = Not Accurate/Insufficient) of structural fidelity, proportional accuracy of podium, staircase and colonnade, Corinthian-column representation, fidelity of remaining columns, walls and roofline against the photographic record, realism of restored relief, inscription and texture detail and overall keyword sufficiency.

Four multi-select items elicited the architectural features judged missing or inaccurate, the aspects perceived to improve across iterations, the elements appearing most divergent from the documented ruins and the descriptive domains where additional keyword coverage was required. Three single-choice items recorded the most accurate iteration for relief iconography, the most convincing iteration for the symbolic objects held by figures within the relief and the preferred candidate among the ten reconstructions assessed. One open-ended item invited free-text suggestions for prompt refinement. Respondents evaluated the candidates against the image plates reproduced in Figure 42–Figure 44.

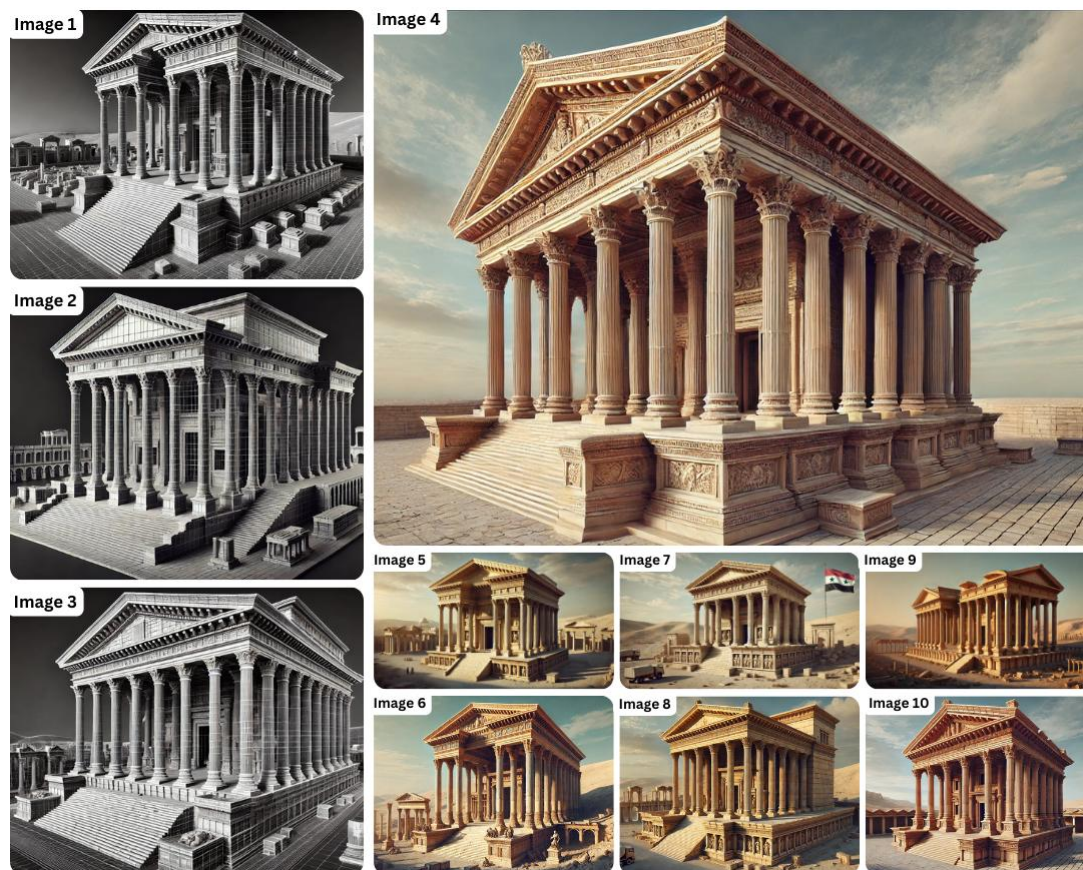


Figure 42. Architectural and structural-accuracy stimuli: the ten AI-generated Temple of Bel reconstruction candidates evaluated by the panel. AI-generated evaluation stimuli, not presented as final reconstructions; platform metadata, model identifiers, prompt version, seed values and refinement-cycle records are recorded in the project paradata in line with HARF and the London Charter (Arzomand et al., 2026; Denard, 2009).

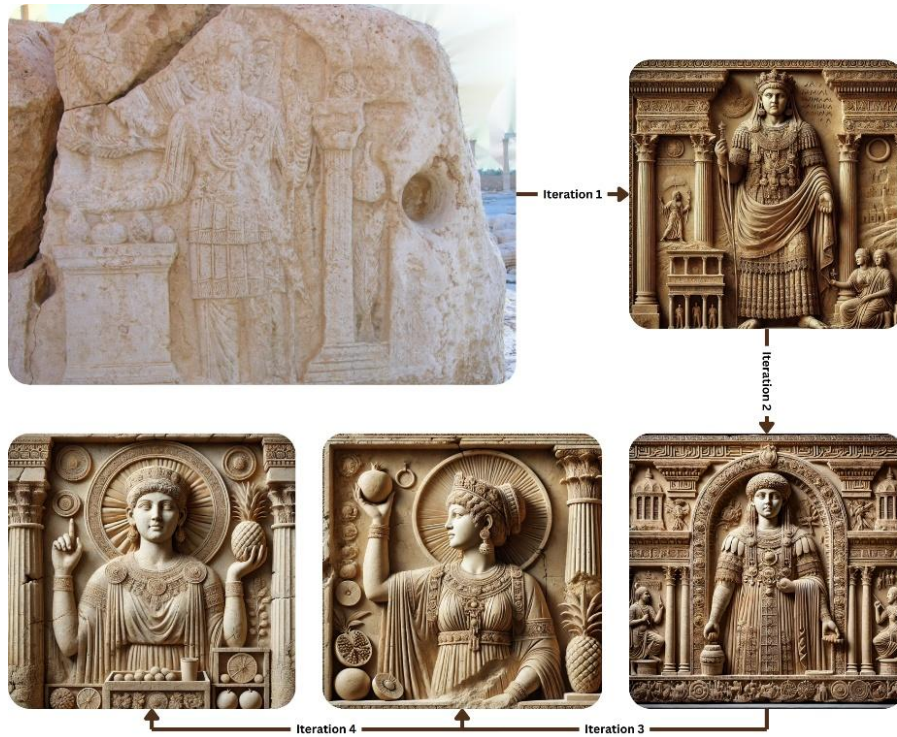


Figure 43. Iterative-refinement stimuli for Temple of Bel relief iconography. Top left: reference photograph of a Palmyra relief (Francesco Bandarin/UNESCO, 2007); remaining panels: AI-generated iterations used as evaluation stimuli.



Figure 44. Comparison of documented Temple of Bel ruins and AI-generated reconstruction stimuli. Image 11: Pre-2015 reference photograph of the Temple of Bel ruins (Getty Images via BBC News). Image 12: Pre-2015 reference photograph of the Temple of Bel ruins (Ian Plumb, CC BY 2.0). Images 13–14: AI-generated evaluation stimuli.

7.2.3 Methodological Justification for Framework-Level Evaluation

The Palmyra evaluation tests framework transferability rather than the historical truth of the generated reconstructions. Keyword sufficiency is therefore treated as a scalar indicator of evidential coverage: it records whether the prompt contained enough site-specific information to constrain generation, while the accompanying keyword-domain item identifies where that coverage remained weak. It is not a full recalculation of the Bamiyan PSI.

Likert aggregation is used descriptively rather than psychometrically. The items are not treated as a single validity scale; each captures a distinct evidential domain, including structural fidelity, proportional coherence, columnar representation, ruin-state comparison, restored detail and keyword sufficiency. Reporting means, medians, interquartile ranges and threshold proportions provides a transparent account of how the panel evaluated each domain.

HARF therefore evaluates whether the framework can be re-instantiated, whether prompt sufficiency can be assessed and whether expert review can locate instability domains. It does not, by itself, establish the historical validity of a reconstruction. Historical plausibility remains dependent on documentary evidence, specialist archaeological authority and, where appropriate, community-grounded interpretation.

7.2.4 Respondents and Analytical Strategy

The Palmyra block was completed by thirty-two respondents drawn from the Phase I interdisciplinary expert panel profiled in Figure 17. That profile records respondent country, education level and professional-field experience for the wider panel; familiarity with the Temple of Bel was elicited separately within the Palmyra block. Of the thirty-two Palmyra completers, ten reported no familiarity with the Temple of Bel, ten slight familiarity, six moderate familiarity and six very high familiarity. The findings are therefore interpreted as interdisciplinary framework-level evidence rather than as Palmyrene specialist judgement.

The respondent profile conditions the interpretation of the accuracy ratings. The monument briefing and image plates gave the panel a shared evidentiary baseline. For the twenty respondents who reported no or slight prior familiarity with the Temple of Bel, the ratings partly reflect conformity between the candidates and that supplied baseline, not independent specialist knowledge of the monument alone. The ratings therefore support a framework-level finding: the transferred HARF instrument made architectural and decorative instability visible to an interdisciplinary panel. The

archaeological status of those instability domains would require subsequent Palmyrene specialist assessment.

Throughout this research, n denotes the number of valid respondents for the relevant item, not the number of images. Unless otherwise stated, $n = 32$ for the Palmyra block. For multi-select items, percentages use the valid respondent count as the denominator and totals may exceed n because respondents could select more than one option.

Likert items are reported as means and standard deviations (SD), medians and interquartile ranges (IQR) and the proportion rated four or five. Multi-select and single-choice categorical items are reported as counts and percentages. Wilson 95% confidence intervals (CI) are reported for headline proportions (Wilson, 1927).

No computational similarity pipeline (SSIM, ORB+RANSAC, proportional similarity) was conducted on the Palmyra candidates. The Bamiyan composite-score procedure (Arzomand & Kalganova, 2026b) was not extended here and the cross-site evaluation is restricted to questionnaire-based framework transferability. The implications of this restriction are addressed in the limitations.

The resulting transferability claim is restricted to the HARF components that were re-instantiated at Palmyra. The DPB schema was repopulated with Palmyra-specific architectural, decorative, material and inscriptional evidence and the Phase I questionnaire block was adapted to the Temple of Bel. The construction-stage PSI defined in Chapter 4 was not recalculated for Palmyra; keyword sufficiency was assessed through a narrower questionnaire item. The Bamiyan computational pipeline, Phase II specialist panel, community-grounded authenticity layer and exhibited paradata package were not repeated at Palmyra. Palmyra therefore tests schema-level transfer and questionnaire-based evaluation, not full replication of the Bamiyan validation.

7.3 Results

7.3.1 Architectural Fidelity

Ratings of architectural and structural accuracy clustered in the moderate band (Table 20). The structural-fidelity rating returned a mean of 3.28 (SD 1.20), with 46.9% of respondents (95% CI 30.9–63.6%) endorsing the candidates as Mostly or Highly Accurate; the proportional-accuracy rating returned 3.22 (SD 1.13) with the same proportion above threshold; the Corinthian-column rating was the highest of the macro-architectural items at 3.47 (SD 1.02), with half the panel above threshold (95% CI 33.6–66.4%). The modal response on the structural-fidelity and proportional-accuracy items

was Mostly Accurate; the Corinthian-column item was split evenly between Mostly Accurate and Moderately Accurate. The pattern indicates cautious endorsement of macro-architectural plausibility rather than strong expert consensus.

Table 20. Likert ratings of the AI reconstructions of the Temple of Bel (n = 32). Counts denote the number of respondents in each scale category. Wilson 95% confidence intervals are reported for the proportion rated ≥ 4 .

Rating dimension	Mean (SD)	Median (IQR)	5	4	3	2	1	% ≥ 4 [95% CI]
Structural fidelity	3.28 (1.20)	3.0 (3–4)	5	10	9	5	3	46.9% [30.9–63.6%]
Proportions of podium, staircase, colonnade	3.22 (1.13)	3.0 (3–4)	3	12	9	5	3	46.9% [30.9–63.6%]
Corinthian columns	3.47 (1.02)	3.5 (3–4)	5	11	11	4	1	50.0% [33.6–66.4%]
Remaining columns, walls, roofline	2.88 (1.18)	3.0 (2–4)	2	9	9	7	5	34.4% [20.4–51.7%]
Restored relief, inscriptions, textures	2.47 (1.11)	2.0 (2–3)	1	5	9	10	7	18.8% [8.9–35.3%]
Keyword sufficiency	3.41 (1.10)	3.5 (3–4)	5	11	10	4	2	50.0% [33.6–66.4%]

The candidate-selection item permitted multiple selections. Image 4 was selected by 16 of the 32 respondents, with 54 selections returned across the ten candidates presented in Figure 42. These frequencies therefore represent multiple candidate selections rather than an exclusive preference for a single reconstruction. Architectural features identified as missing or inaccurate concentrated on column placement and height (40.6%, 95% CI 25.5–57.7%), interior *thalamoi* (37.5%), roofline merlons (34.4%), zodiac ceiling reliefs (31.2%) and architrave relief carvings (28.1%). Corinthian-column representation was therefore rated more favourably than restored relief and texture detail, but column placement and height remained a non-trivial source of expert concern.

This result has a practical implication for future applications of HARF. Macro-structural recognisability should not be treated as sufficient evidence of reconstruction adequacy. A generative output may preserve the broad silhouette or architectural massing of a site while failing in features that carry archaeological or cultural specificity. When HARF is transferred to another artefact, monument, or evidence-rich reconstruction context, the prompt-design and expert-review stages should first identify the culturally diagnostic features through which the object’s site-specific identity is expressed. In the Palmyra case, these features include relief carving, surface treatment, roofline form, inscriptions and interior cult-architectural elements; in another case, the corresponding high-risk features would need to be defined from the documentary record before evaluation begins.

7.3.2 Iterative Refinement

Iteration did not resolve expert-panel concerns about relief iconography. As shown in Figure 45, 56.2% of respondents (18/32; 95% CI 39.3–71.8%) selected “none of the iterations are historically accurate” when asked to identify the iteration most historically accurate in its depiction of the original relief. The remaining responses were distributed across the four iterations, with no single iteration commanding more than 15.6% of the panel.

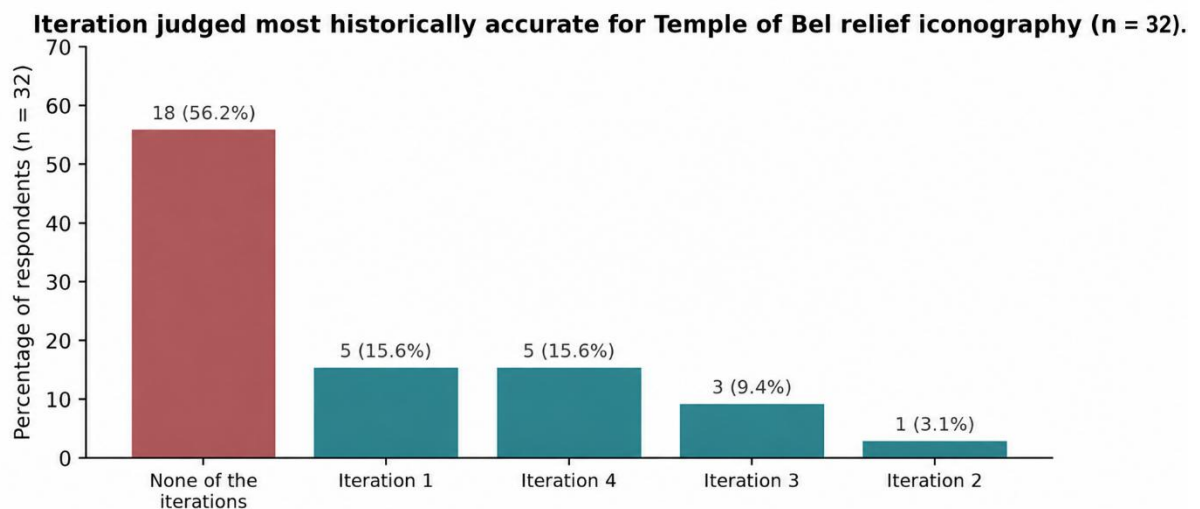


Figure 45. Iteration judged most historically accurate for Temple of Bel relief iconography. The modal response was “None of the iterations are historically accurate” (56.2%; 95% CI 39.3–71.8%), indicating that iterative refinement did not resolve relief-iconographic concerns. Author-generated chart based on Palmyra questionnaire responses.

The result translates into a design constraint for generative reconstruction workflows. Relief iconography should be treated as a high-risk evidential domain rather than as a detail expected to improve through general iterative refinement. Iteration may improve surface coherence, but it does not necessarily improve historical or iconographic adequacy. In future HARF applications, iconographic, inscriptional and culturally specific surface features should be isolated as explicit evaluation targets from the outset, rather than left to late-stage visual refinement.

On the iteration that most convincingly reconstructed the symbolic objects held by figures within the relief, the panel split evenly between “none” and the most refined iteration (10/32 each, 31.2%). Across iterations, respondents recognised gains in background architectural elements (46.9%) and decorative jewellery (43.8%) more readily than in inscriptions and iconography (34.4%) or facial expressions and proportions (31.2%). Iterative refinement, on this evidence, did not consistently translate visual gains into iconographic plausibility. The implication is that refinement cycles can be retained within HARF, but only as controlled diagnostic stages: each iteration should be assessed against predefined evidential criteria rather than against general visual improvement.

7.3.3 Comparison with the Photographic Record

Ratings for the items concerning the surviving ruin and restored decorative detail were lower than those for the preceding macro-architectural items. The proportional-fidelity rating against the remaining columns, walls and roofline returned a mean of 2.88 (SD 1.18); the realism rating for restored relief, inscription and surface detail was the lowest of any item in the block at 2.47 (SD 1.11), with under 20% of respondents above threshold (Table 20). The categories most frequently identified as divergent from the photographic record are reported in Figure 46. Relief sculptures and wall carvings were selected by 62.5% of respondents (95% CI 45.3–77.1%), followed by surface textures and weathering (59.4%), roofline and decorative merlons (53.1%), entrance proportions (37.5%) and column height and fluting patterns (31.2%). Surface, relief and roofline detail therefore dominate the divergence profile; columnar geometry, while present in the distribution, sits below them.

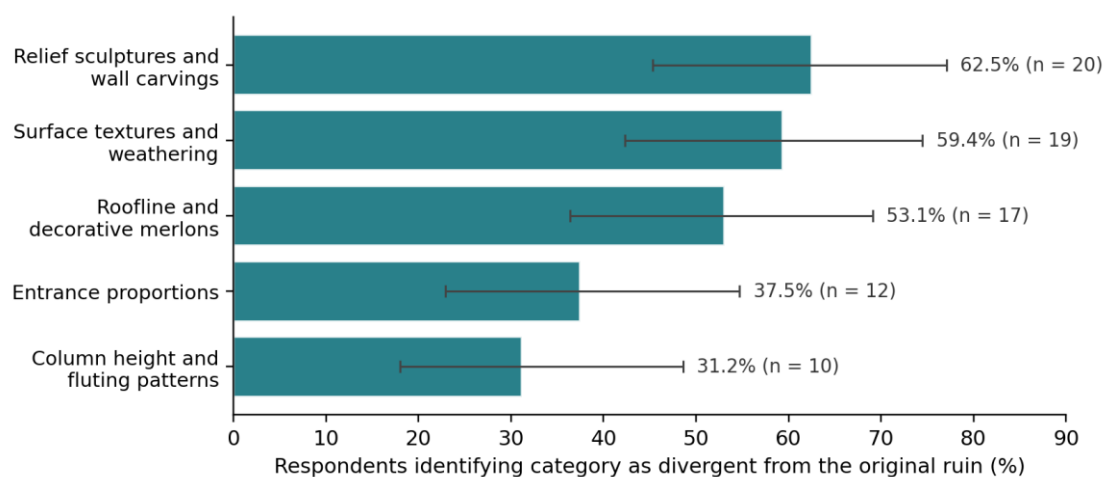


Figure 46. Categories identified as most divergent from the documented Temple of Bel ruins. Bars denote the proportion of respondents flagging each category; error bars denote Wilson 95% confidence intervals. Surface, relief and roofline detail concentrate at the top of the distribution. Author-generated chart based on Palmyra questionnaire responses.

7.3.4 Prompt-Keyword Sufficiency

The keyword-sufficiency rating returned a mean of 3.41 (SD 1.10), with half of the panel rating the keyword set as Mostly or Highly Sufficient and 6.2% as Insufficient. This is broadly consistent with the expert-rated prompt-sufficiency value reported for the Bamiyan case (3.67/5; Section 4.5), although the Palmyra measure is methodologically narrower because it is based on a single keyword-sufficiency item rather than the multi-domain construction-stage index defined in Chapter 4 (Arzomand et al., 2026). Domains where additional keyword coverage was requested concentrated on environmental and surrounding structures (53.1%, 95% CI 36.4–69.1%), architectural proportions and layout (46.9%),

artistic elements such as friezes, frescoes and relief carvings (37.5%), cultural and religious symbolism (37.5%) and material and texture detail (31.2%). Four respondents (12.5%) considered the keyword set fully sufficient. Nine respondents provided substantive free-text suggestions; these recommended explicit specification of the *cella* footprint, removal of evocative or impressionistic phrasing in favour of concrete descriptors, the addition of weathering and patina vocabulary, the explicit inclusion of stepped merlons and the grouping of keywords for use across iterative refinement steps. The pattern is consistent with the principle, articulated for Bamiyan, that prompt completeness is improved through site-specific evidential anchoring rather than through stylistic embellishment (Arzomand et al., 2026).

7.4 Discussion

The Palmyra evaluation supports a framework-level interpretation of HARF's transferability. The principal finding is not that the Temple of Bel outputs were historically accurate; they were not consistently judged as such. Rather, HARF's schema remained usable when transferred from a figural Gandhāran sculpture to a Graeco-Roman architectural monument: the DPB could be re-instantiated, the questionnaire could be extended to a new site and the expert panel could identify domains of evidential instability. Macro-architectural features were assessed more favourably than relief, inscription, surface and roofline detail; 56.2% of respondents judged that none of the relief iterations was historically accurate. The Palmyra case therefore does not validate the Temple of Bel reconstructions as historically adequate. It supports a narrower claim: HARF can structure the evaluation of a different heritage typology and expose where generative outputs fail.

Within this bounded cross-site comparison, the findings indicate a typology-dependent instability pattern. The pattern concerns the different domains in which generative instability was most evident in the Bamiyan and Palmyra cases. At Bamiyan, this instability concentrated in figural micro-grammar: drapery articulation, cranial morphology and hand gesture (Arzomand & Kalganova, 2026b). At Palmyra, it shifted to architectural and decorative micro-grammar: relief carving, surface texture, inscription, roofline treatment and interior cult-architectural features. The schema remained stable, but the failure layer changed. This matters because framework transferability does not imply output reliability. A transferable evaluation framework should be judged by whether it can locate the new failure domains produced by a different object type, not by whether its outputs appear visually coherent.

The Palmyra instability profile has the same status as the wider transferability test. It identifies the main areas of concern raised by the transferred HARF instrument: relief carving, surface treatment,

inscription, roofline form and interior cult-architectural detail. These areas matter because framework transfer did not remove generative instability; it relocated it from bodily form and gesture to architectural surface, inscription and decorative detail. A subsequent Palmyrene specialist assessment would be needed to determine how far this profile corresponds to the archaeological record of the Temple of Bel.

The cross-site evidence can be represented as a three-layer model of transferability.

Table 21. Three-layer model of HARF transferability and typology-dependent instability patterns across Bamiyan and Palmyra.

Layer	Role in the evaluation	Bamiyan	Palmyra	Implication
Stable framework layer	Prompt structure, paradata, sufficiency assessment, expert review	Preserved	Preserved	Can transfer without schema redesign
Site-specific evidence layer	Material, geometry, iconography, environmental context	Gandhāran sculpture, mud-straw plaster, drapery, mudrā	Limestone temple, Corinthian colonnade, thalamoi, inscriptions	Must be re-specified for each case
Typology-dependent instability layer	Features most likely to fail under generation	Drapery, cranial morphology, hand gesture	Relief, surface, inscription, roofline	Must be identified early and evaluated explicitly

This model keeps the contribution of the Palmyra case bounded. The study does not show that HARF produces historically reliable reconstructions across heritage sites. It shows that, when applied to a new typology, HARF can preserve its evaluative architecture while revealing a different pattern of failure. A future application should therefore not begin from a blank methodological position, but neither should it assume that a successful Bamiyan prompt structure will automatically produce a valid reconstruction elsewhere. The schema can be retained; the descriptors and high-risk evidential domains must be redefined from the documentary record.

The Palmyra results also qualify the role of iterative refinement. The iteration sequence did not resolve relief-iconographic concerns, despite some perceived gains in background architecture and decorative detail. Iteration should therefore be treated as a controlled diagnostic stage, not as an automatic route to historical improvement. Each refinement cycle should be assessed against predefined evidential criteria, particularly where iconography, inscription, surface weathering, or roofline form is archaeologically significant. This is consistent with the wider problem identified in generative-AI heritage work: visual coherence can improve while evidential adequacy remains weak, especially where outputs depend on culturally specific surface features that may be under-represented or distorted in training data (Foka & Griffin, 2024a).

The study therefore establishes Phase I transferability rather than complete validation. Phase I shows that HARF's schema and evaluative logic can be re-instantiated across one major typological transition,

from figural sculpture to architectural monument. Phase II remains necessary where site-specific archaeological authority, community grounding, or computational verification is required. HARF is consequently best understood as a diagnostic instrument for responsible generative-AI evaluation, not as a reconstruction engine.

The cross-site comparison can therefore be stated precisely. HARF transfers at the level of schema and evaluative logic; historical plausibility remains dependent on site-specific evidence, specialist judgement and, where relevant, community interpretation. Table 21 summarises this distinction between framework-level transferability and site-specific evidential validity.

7.5 Limitations

Five limitations qualify these findings. First, the panel is interdisciplinary rather than composed of specialists in Palmyrene archaeology. As reported in Section 7.2.4, twenty of the thirty-two Palmyra completers reported no or only slight familiarity with the Temple of Bel. The structured monument briefing mitigated this constraint but cannot substitute for a Palmyrene specialist panel. Secondly, no Phase II Palmyra specialist evaluation accompanies this chapter, so the depth of site-specific authority achieved at the second phase of the Bamiyan study is not reproduced here (Arzomand & Kalganova, 2026b). Thirdly, no community-grounded Syrian or Palmyrene respondent layer accompanies the study, of the kind developed for Bamiyan in the polyvocal community study extending the wider research programme (Arzomand & Kalganova, 2026a). This is a substantive limitation, particularly because recent work on the Temple of Bel after its destruction stresses the importance of the Palmyrene population and collective memory in the site's future (Abdulkarim & Seigne, 2024). Fourthly, no Palmyra-specific computational similarity pipeline was conducted; the cross-site evaluation is therefore restricted to questionnaire-based Phase I transferability and full computational evaluation remains future work. Fifthly, reproducibility constraints arising from the volatility of generative platforms apply equally here. Paradata logging mitigates the constraint, but does not eliminate it (Huvila, 2022a).

7.6 Conclusion

A bounded cross-site evaluation indicates that HARF's prompt blueprint and evaluative logic can be transferred from a figural Gandhāran sculpture to a Graeco-Roman architectural monument without redesigning the framework's top-level schema. The transfer is evaluative rather than reconstructive: the Palmyra outputs were not consistently judged historically accurate. The observed instability profile

differed between the two cases, shifting from drapery, cranial morphology and hand gesture at Bamiyan to relief, surface, inscription, roofline and interior cult-architectural detail at Palmyra. Within this bounded comparison, this supports a typology-dependent instability pattern for generative heritage reconstruction.

The broader implication is that frameworks such as HARF are best understood as diagnostic instruments rather than reconstruction engines. Their value lies in structuring evidence, recording prompt logic, assessing prompt sufficiency and identifying where generative outputs remain unstable. They do not remove the need for specialist judgement, nor do they convert visual plausibility into historical plausibility. A Phase II evaluation by specialists in Palmyrene archaeology, a Syrian and Palmyrene community layer and a Palmyra-specific computational similarity pipeline remain necessary extensions.

These findings close the empirical sequence of the thesis. Chapter 3 established the initial evidence-constrained generation problem; Chapter 4 formalised HARF; Chapters 5 and 6 validated the framework deeply at Bamiyan; and this chapter has tested its bounded transfer to Palmyra. Chapter 8 now synthesises these findings across the research questions, evaluates the framework as a whole and defines the limits of the thesis claim.

Chapter 8

Discussion and Conclusion

The principal contribution of this thesis is the Heritage-Aligned Reconstruction Framework, developed and validated within a bounded two-dimensional scope as a way of evaluating generative AI heritage reconstruction candidates rather than producing definitive reconstructions of any specific monument. HARF moves from an initial evidence-constrained text-to-image workflow to a formal evaluation procedure organised around the Dynamic Prompt Blueprint, the Prompt Sufficiency Index, paradata, computational pre-evaluation and structured expert review. It is operationalised at the Western Buddha of Bamiyan, extended through a community-grounded authenticity register and re-instantiated at the Temple of Bel, Palmyra, to test bounded cross-site generalisation. This chapter discusses the framework in those terms. It sets out how HARF answers each of the four research questions, evaluates the framework against its bounded scope and limitations and identifies the work that follows.

8.1 HARF and the Research Questions

HARF is the central contribution of this thesis. It responds to the four-part evaluation problem identified in Chapter 2: evidential control, AI-specific evaluation procedure, the relation between computational, expert and community-grounded assessment and bounded transferability across heritage typologies. The four research questions correspond to those four problems and are answered through the framework rather than alongside it.

RQ1 asked how generative AI heritage reconstruction can be constrained by documentary, photographic, archaeological and architectural evidence so that source material, prompt construction, model configuration, refinement choices and output selection form an auditable record. Chapter 3 produced an initial evidence-informed text-to-image workflow that exposed the limits of pre-HARF prompting and informal visual comparison. Chapter 4 converted those limits into the Dynamic Prompt Blueprint, the Prompt Sufficiency Index and the paradata protocol. RQ1 is therefore answered at the level of evidential control. A GenAI reconstruction candidate becomes assessable only when source

selection, prompt construction, model choice, generated output sets and filtering decisions are treated as a single auditable record.

RQ2 asked what integrated evaluation procedure can assess generative AI heritage reconstructions through evidence-anchored prompting, prompt sufficiency assessment, paradata, computational pre-evaluation and structured expert review. Chapter 4 formalised HARF as that procedure and Chapter 5 tested it through a generated corpus of 245 Bamiyan reconstruction candidates, Phase I interdisciplinary screening with thirty-three experts on twenty-three images, Phase II specialist assessment with five Bamiyan specialists on eighteen refined images and paired computational–expert analysis on the eleven full-figure outputs. The comparison between the composite computational ranking and the Phase II mean applicable-criteria specialist scores produced only a weak association, with Spearman’s $\rho = 0.369$ and $p = 0.264$. Given the $n = 11$ basis of this comparison, the result is treated descriptively. Its value lies in showing that computational similarity and specialist judgement did not operate as interchangeable measures within the Bamiyan assessment. RQ2 is therefore answered as a procedural claim rather than as a guarantee of accuracy. HARF makes the basis of evaluation inspectable, so that visual plausibility, evidential sufficiency, measurable structural correspondence and specialist plausibility are separated and failure can be located rather than hidden beneath visual finish.

RQ3 asked how computational similarity, structured expert judgement and community-grounded authenticity assess different dimensions of validity in AI-mediated reconstruction at Bamiyan. Computational similarity identifies measurable correspondence with reference imagery but does not assess iconographic or cultural adequacy. Specialist judgement evaluates archaeological, proportional and stylistic plausibility but does not determine community acceptability. The Afghan respondent study reported in Chapter 6 adds a third register. Sixty-two respondents participated in the study. Of these, twenty-five were based inside Afghanistan, thirty-six were in the diaspora and one did not report location. The majority supported reconstruction in some form, but this support did not amount to an unqualified preference for visual replacement. Respondents favoured forms of representation that kept the fact of destruction visible while allowing reconstruction, interpretation or future-facing display. Holograms appeared as one possible representational modality in the questionnaire, but they were not the preferred mode. RQ3 is therefore answered as a multi-register validity claim. Reconstruction validity at Bamiyan is not a single score. It is a relation between distinct forms of evidence, with computational similarity filtering, specialist judgement assessing disciplinary plausibility and community-grounded authenticity providing evidence about cultural acceptability within the bounds of the respondent sample.

RQ4 asked to what extent an evaluation framework developed through a figural Gandhāran case can be re-instantiated for a monumental architectural case without redesigning its top-level schema and where typology-dependent failure domains emerge. Chapter 7 tested this through the Temple of Bel at Palmyra. Palmyra is not a second full validation case. It is a bounded transferability test. The Dynamic Prompt Blueprint schema and the Palmyra questionnaire block were re-instantiated for a Graeco-Roman architectural typology. The full Bamiyan evaluation pipeline was not repeated: PSI recalculation, the computational pipeline, a Phase II specialist panel, a community-grounded layer and exhibited paradata were not implemented for Palmyra. The Palmyra case therefore tests transfer of the evaluative schema rather than full replication of the Bamiyan validation procedure. Panel familiarity also shaped the interpretation of the Palmyra ratings, since twenty of the thirty-two respondents reported no or only slight familiarity with the Temple of Bel. The accuracy ratings therefore partly reflect conformity to the supplied briefing as well as independent recognition of the site. Within that limit, the case shows that the stable framework layer can be transferred while the site evidence and failure domains must be rebuilt. Bamiyan failures concentrated in figural micro-grammar, including drapery articulation, cranial morphology and hand gesture, while Palmyra failures concentrated in architectural and decorative micro-grammar, including relief carving, surface treatment, inscription, roofline form and interior spatial articulation. RQ4 is therefore answered as a bounded generalisation claim. HARF generalises at the level of evaluative schema, evidentiary logic, prompt-structure transfer, paradata discipline and failure-domain diagnosis. It does not prove that AI outputs are historically accurate across all heritage sites.

HARF changes the status of the generated image. In an unstructured workflow, the image is the primary artefact and evaluation is retrospective. In HARF, the image is one element within an evaluative record: evidence is selected before prompting, prompt construction is structured and assessed, paradata records interpretive decisions, computational pre-evaluation identifies measurable structural drift and expert review addresses the archaeological, stylistic and proportional adequacy that metrics cannot decide. Computational similarity filters but does not adjudicate. Expert judgement assesses disciplinary plausibility but does not determine community acceptability. Community-grounded authenticity provides evidence about cultural acceptability within the bounds of the respondent sample but does not adjudicate archaeological or stylistic accuracy.

The thesis therefore shifts the analytical weight from what GenAI can produce to how its outputs are made reviewable. GenAI reconstruction candidates can be treated as heritage reconstruction evidence only when their evidential basis, computational correspondence, expert assessment, community-grounded acceptability and limits of transfer are made explicit within a framework that keeps those registers distinct while connecting them as evaluative components.

Together, the four research questions establish four connected contributions. First, the thesis has produced an initial evidence-constrained AI reconstruction workflow (Chapter 3) whose value lies in exposing why unstructured prompting and informal visual comparison are insufficient for accountable heritage reconstruction. Second, it has formalised the Heritage-Aligned Reconstruction Framework as an integrated AI-specific evaluation procedure that separates visual plausibility from evidential plausibility (Chapter 4). Third, it has produced a multi-register evaluation of AI-mediated reconstruction at Bamiyan, combining computational, expert and community-grounded registers within one framework rather than collapsing them into a single criterion (Chapters 5 and 6). Fourth, it has produced a bounded cross-site generalisation test at the Temple of Bel, Palmyra, showing that the evaluative schema transfers from a figural Gandhāran case to a monumental architectural case while site evidence and typology-specific failure domains must be redefined for each case (Chapter 7).

8.2 Overall Evaluation of the Framework

HARF supplies a structured evidential pathway from documentary sources to generated output through the Dynamic Prompt Blueprint, the Prompt Sufficiency Index and the paradata protocol; it integrates computational filtering and structured expert review into a single staged procedure rather than treating them as separate methodological settings; and it produces a usable cross-site transfer test rather than relying on internal coherence alone to establish generalisability. The framework's strongest empirical validation lies in the two-dimensional, image-based evaluation of Bamiyan reconstruction candidates and its transferability claim is anchored to the single architectural case at the Temple of Bel. Six conditions define the scope of the claim.

8.2.1 Platform Volatility and Reproducibility

HARF is designed to make the evaluation record reproducible. The generative platforms it evaluates remain only partially reproducible. Version updates, opaque changes to training data and non-deterministic sampling mean that prompt-level traceability does not guarantee identical outputs across time or platforms. Paradata logging mitigates this constraint but does not remove it.

8.2.2 Reference-Material Dependency

Computational similarity depends on the availability and condition of archival reference imagery. Where documentation is fragmentary, the Prompt Sufficiency Index reflects the completeness of the

descriptive schema rather than the completeness of the historical record. Structural similarity scores cannot adjudicate features that fall outside the available reference.

8.2.3 Expert-Panel Scope

Bamiyan received deeper specialist authority than Palmyra. Phase I drew on interdisciplinary experts and Phase II on Bamiyan specialists. The Palmyra block was administered to an interdisciplinary panel rather than a dedicated Palmyrene specialist sub-panel and many respondents had limited prior familiarity with the Temple of Bel. The Palmyra evaluation therefore tested framework-level transferability rather than site-specific specialist authority.

8.2.4 Community-Layer Scope

The community-grounded register was operationalised for Bamiyan and not for Palmyra. The Bamiyan respondent layer is also bounded by convenience sampling, language of administration, security conditions inside Afghanistan and the small Hazara sub-sample. It should not be read as an aggregate account of Afghan or Hazara opinion.

8.2.5 Computational-Pipeline Scope

The full computational pipeline was applied within the Bamiyan case and was not extended as a parallel pipeline at Palmyra. The bounded generalisation claim therefore covers the schema and evaluative logic, not a quantitative metric-based replication.

8.2.6 Two-Dimensional Output Scope

HARF has been validated within the scope of evaluating two-dimensional, image-based GenAI reconstruction candidates. It has not been extended in this thesis to text-to-3D outputs, volumetric reconstruction, Neural Radiance Fields, Gaussian splatting or immersive heritage environments. This exclusion is methodological. Three-dimensional outputs require additional validation of spatial coherence, geometric tolerance, uncertainty communication and embodied or immersive interpretation.

8.3 Future Work

Future work follows from these limitations. The first direction is further empirical testing at Palmyra. A Phase II Palmyrene specialist evaluation would test whether the failure-domain pattern identified by the interdisciplinary panel holds under site-specific expertise. A Syrian and Palmyrene community study would add the community-grounded register that was present at Bamiyan but absent at Palmyra. A Palmyra-specific computational similarity pipeline would test whether the divergence between computational and expert registers also appears in a monumental architectural case.

The second direction is strengthening the Bamiyan community register. A language-stratified replication in Dari, Pashto and Hazaragi would improve reach and reduce dependence on English-mediated participation. Semi-structured interviews with Afghan heritage practitioners, Bamiyan residents and diaspora representatives would test the survey findings against fuller accounts of memory, legitimacy and acceptable reconstruction. A further reconstruction run could apply the proposed conditions of provenance visibility, rupture visibility and governance visibility to test whether they materially change the reconstruction record.

The third direction is refinement of HARF itself. PSI weighting should be tested across heritage domains and against incremental prompt adjustments. FAIR-compliant repositories for prompts, model versions and paradata would strengthen reproducibility as platforms evolve. Domain-specific evidence corpora and evaluation datasets would help address recurring bias patterns identified in the thesis, including symmetrical drapery, stylised hair morphology and Petra-like architectural elements. AI-specific guidance aligned with the London Charter and the Seville Principles would also be needed for disclosure, metadata retention, prompt provenance and community participation in prompt-based reconstruction (Denard, 2009; López-Mencheró Bendicho & Grande, 2017).

The fourth direction is conditional extension to three-dimensional generative outputs. Future work should examine the extension of HARF from two-dimensional image-based reconstruction into text-to-3D, volumetric reconstruction, Neural Radiance Fields, Gaussian splatting and immersive heritage environments, but only after the evidential, expert and community-grounded evaluation logic established in this thesis has been adapted for three-dimensional outputs. Three-dimensional extension introduces distinct challenges concerning spatial coherence, lighting stability, uncertainty communication and the calibration of geometric tolerances. These challenges should not be addressed by re-using the present framework directly. Each evaluative register would need to be adapted for the new output type while retaining the framework's separation between visual plausibility and evidential plausibility.

8.4 Closing Statement

Generative AI has made images of destroyed monuments easier to produce than to justify as heritage reconstruction. This thesis has addressed that imbalance by formulating, operationalising and testing HARF as a framework through which generated reconstruction candidates can be evaluated as heritage claims rather than as visual content. HARF places evidence before generation, treats validity as a multi-register property and tests framework-level transfer across a second heritage typology. Its claim is deliberately bounded. The thesis does not claim that GenAI outputs are historically accurate across heritage sites. It makes a methodological claim about how such outputs should be documented, constrained and evaluated before they are treated as reconstruction evidence. The contribution of the thesis is therefore not a reconstructed Bamiyan or Palmyra image, but an accountable procedure for deciding how such images should be assessed before they are accepted as heritage reconstruction.

Appendix A. Dynamic Prompt Blueprint (DPB): Site-Agnostic Schema

This appendix defines the generic DPB schema used to construct evidence-aligned prompts in heritage reconstruction workflows. The schema specifies the core tiers required for reproducible, cross-site application. Placeholders in *[brackets]* denote variables to be instantiated for each specific monument or cultural context.

A.1. DPB Schema Specification

Tier	Descriptor	Fields / Placeholders
T1. Monument identifier	Primary classification	<i>[Name]; [Location]; [Period]; [Functional type]</i>
T2. Dimensional anchors	Geometric constraints	Height $\approx$ <i>[H]</i> ; shoulders $\approx$ <i>[S]</i> ; head $\approx$ <i>[HH]</i> ; pose/gesture = <i>[POSE]</i> ; proportional locks = <i>[Source]</i>
T3. Iconography: head & face	Stylistic constraints	<i>[Facial features]; [Hairstyle];</i> exclude <i>[Unattested traits]</i>
T4. Drapery & material grammar	Textile conventions	<i>[Drapery tradition];</i> fold types = <i>[FOLD_TYPES]</i> ; materials = <i>[MATERIAL]</i> ; pigments = <i>[PIGMENT_INFO]</i> ; exclude <i>[Banned terms]</i>
T5. Material & surface condition	Structural evidence	<i>[Stone/Plaster/Brick]; [Weathering/Erosion/Loss];</i> record <i>[Cracks/Repairs/Patina]</i>
T6. Architectural context	Spatial embedding	Niche/cavity = <i>[SHAPE $\times$ SIZE]</i> ; stratigraphy/openings = <i>[STRUCTURAL DATA]</i> ; peripheral structures = <i>[STRUCTURES]</i>
T7. Environmental setting	Physical environment	<i>[Topography]; [Orientation]; [Lighting conditions]</i>
T8. Cultural & symbolic filters	Inclusion/exclusion	Include <i>[Verified symbols]</i> ; exclude <i>[Anachronisms]</i>
T9. Technical controls	System parameters	Pose/scale locks = <i>[CONTROL]</i> ; sampling/guidance = <i>[PARAMETERS]</i> ; negative prompts = <i>[LIST]</i> ; paradata tag = <i>[TAG]</i>

A.2. Implementation Notes

1. Populate each tier only with descriptors supported by archaeological, architectural, or art-historical evidence.
2. Tiers should be concatenated in the sequence T1 $\rightarrow$ T9 to preserve a structural-first, stylistic-second, environmental-third logic.

3. The schema is model-agnostic; platform-specific settings (e.g., seeds, output resolution) should be documented separately.
4. Each prompt instantiation should include a *paradata tag* linking descriptive elements to their evidential sources.

A.3. Abstract Example (Non-site-specific)

(Illustrative format only.)

Reconstruct *[Monument]* located in *[Region]*, dated to *[Period]*. Maintain proportional anchors (height $\approx [H]$) based on *[Source]*. Apply *[Facial style]* and *[Hairstyle]* aligned with *[Tradition]*; exclude *[Unattested features]*. Use *[Drapery grammar]* with fold type *[FOLD_TYPE]* and material *[MATERIAL]*. Embed within *[Architectural context]* and represent surface conditions consistent with *[Weathering]*. Situate within *[Environmental setting]* under *[Lighting]*. Apply control settings *[PARAMETERS]* and record metadata as *[PARADATA]*.

Appendix B. Phase I Expert Evaluation Instrument (Bamiyan Case Study)

B.1 Scope and Design

The Phase I expert evaluation instrument operationalises the core assessment domains defined within the Heritage-Aligned Reconstruction Framework (HARF): structural fidelity, iconographic accuracy, environmental plausibility and prompt sufficiency. These domains correspond directly to the analytical structure of Interdisciplinary Expert Screening (IES) and provide a consistent basis for evaluating generative reconstruction outputs.

The instrument was administered as a structured online questionnaire to domain experts. The version presented here documents the Bamiyan-specific implementation and is provided to ensure methodological transparency. The same instrument underpins the empirical evaluation reported in Chapters 4 and 5. It is included here to document the evaluation structure rather than to reproduce the full empirical results.

B.2 Participant Profile

Participants reported the following attributes:

- disciplinary specialisation
- institutional affiliation
- country of residence
- highest academic qualification
- prior familiarity with the Bamiyan Buddhas

B.3 Evaluation Scale

All quantitative items were assessed using a five-point Likert scale:

- 1 — Very poor (substantial deviation from archaeological or visual plausibility)
- 2 — Poor (major inaccuracies affecting interpretability)
- 3 — Moderate (partial correspondence with identifiable inconsistencies)

4 — Good (minor deviations with overall acceptable fidelity)

5 — Excellent (high correspondence with expected archaeological and visual characteristics)

B.4 Structural Fidelity and Surface Realism (Images 1–7)

Table B1. Structural fidelity and surface realism assessment items (Phase I, Images 1–7)

Item	Assessment Focus	Response Type
B4.1	Overall photorealism	Likert scale (1–5)
B4.2	Archaeological and historical accuracy	Likert scale (1–5)
B4.3	Accuracy of niche and cave structure	Likert scale (1–5)
B4.4	Proportional consistency (statue–niche)	Likert scale (1–5)
B4.5	Material and surface texture realism	Likert scale (1–5)
B4.6	Selection of most accurate reconstruction (Images 1–7)	Single-choice selection
B4.7	Identification of structural or geometric inconsistencies	Open-ended response
B4.8	Proposed corrections to structural features	Open-ended response

B.5 Iconographic and Stylistic Accuracy (Images 8–17)

Table B2. Iconographic and stylistic assessment items (Phase I, Images 8–17)

Item	Assessment Focus	Response Type
B5.1	Facial expression and stylistic accuracy	Likert scale (1–5)
B5.2	Authenticity of hair representation	Likert scale (1–5)
B5.3	Accuracy of hand gestures (mudrā)	Likert scale (1–5)
B5.4	Drapery consistency (Gandhāran conventions)	Likert scale (1–5)
B5.5	Overall iconographic plausibility	Likert scale (1–5)
B5.6	Selection of most accurate reconstruction (Images 8–17)	Single-choice selection
B5.7	Identification of stylistic inconsistencies	Open-ended response
B5.8	Proposed corrections to iconographic features	Open-ended response

B.6 Environmental Plausibility (Images 18–23)

Table B3. Environmental plausibility assessment items (Phase I, Images 18–23)

Item	Assessment Focus	Response Type
B6.1	Accuracy of cliffside geological context	Likert scale (1–5)
B6.2	Environmental integration of the statue	Likert scale (1–5)
B6.3	Consistency with documented site conditions	Likert scale (1–5)
B6.4	Selection of most accurate reconstruction (Images 18–23)	Single-choice selection
B6.5	Identification of environmental inconsistencies	Open-ended response
B6.6	Proposed contextual improvements	Open-ended response

B.7 Iterative Reconstruction Assessment

Table B4. Iterative reconstruction assessment items (Phase I)

Item	Assessment Focus	Response Type
B7.1	Selection of more accurate reconstruction (baseline vs refined)	Single-choice selection
B7.2	Justification of selection	Open-ended response
B7.3	Stage at which facial improvement becomes evident	Categorical stage selection
B7.4	Stage at which hand gesture improvement becomes evident	Categorical stage selection
B7.5	Stage at which drapery improvement becomes evident	Categorical stage selection

Categorical stage selection refers to predefined stages of the generative refinement process presented to participants during evaluation.

B.8 Prompt Sufficiency

Table B5. Prompt sufficiency assessment items (Phase I)

Item	Assessment Focus	Response Type
B8.1	Sufficiency of descriptive detail	Likert scale (1–5)
B8.2	Clarity of structural and dimensional instructions	Likert scale (1–5)
B8.3	Alignment between prompt specification and generated output	Likert scale (1–5)
B8.4	Identification of missing prompt elements	Open-ended response
B8.5	Proposed refinements to prompt structure	Open-ended response

B.9 Use of Responses

Quantitative responses were aggregated at item and domain levels to enable comparative evaluation across candidate reconstructions. Domain-level aggregation supported the identification of high-performing outputs and recurrent error patterns. Qualitative responses were used to contextualise scoring distributions and to inform iterative refinement within the HARF workflow.

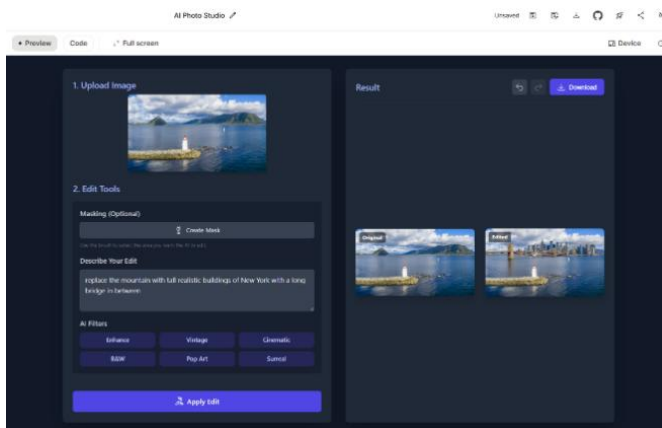
Appendix C. Bamiyan Candidate Corpus, Visual Stimuli and Paradata



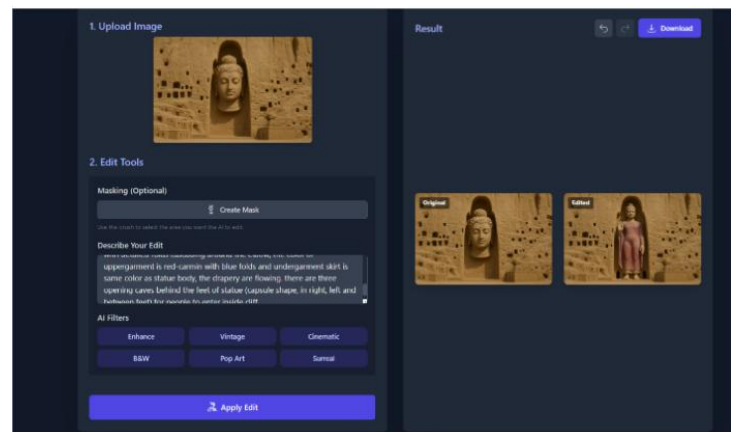
C. 1. Hybrid refinement workflow combining Sora base generation with Gemini (Google AI Studio) regional masked editing to correct localised anatomical and drapery distortions while preserving global statue proportions.



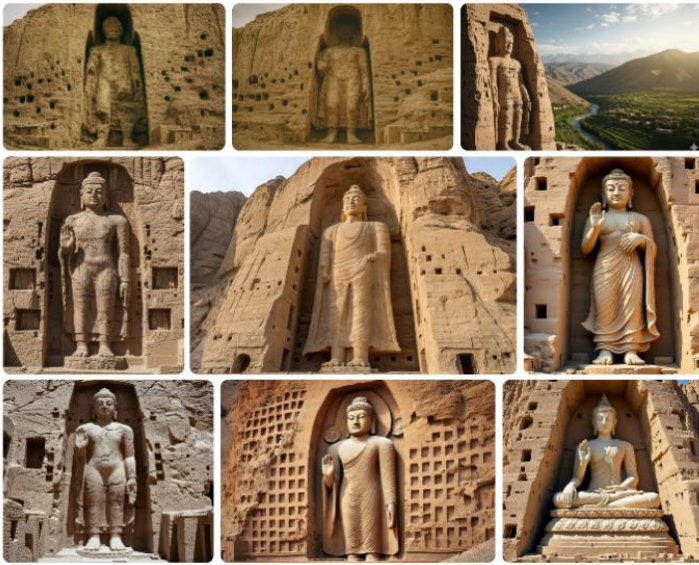
C. 2. Gemini-only generations (prompt-only; excluded). Gemini generations produced from text prompts without Sora inputs. Outputs often exhibited high surface realism but introduced non-evidenced architectural insertions and/or proportion drift and were therefore excluded from the candidate pool.



C. 3. Google AI Studio API-based masked editing workflow. It shows general API testing on non-heritage images to verify regional editing functionality.



C. 4. Google AI Studio API-based masked editing. It illustrates heritage-specific application to Bamiyan reconstruction outputs.



C. 5. Initial Sora-generated outputs produced from a single-shot prompt. Observed failures include mirrored hand gestures, incorrect mudra interpretation, exaggerated cranial proportions and drapery folds inconsistent with Gandhāran “wet-fold” conventions.



C. 6. Sora outputs under iterative text-only prompt revision. Sora generations produced by repeatedly revising the textual prompt to correct specific failures (posture, facial morphology, drapery structure). Despite successive revisions, several outputs retained structural or stylistic instability (e.g., residual hand errors, inconsistent drapery grammar, or niche-context distortions)

Domain	Strengths Observed	Weaknesses / Inaccuracies
Facial Expression	Some outputs conveyed serenity and composure consistent with Buddhist iconography.	Faces frequently over-smoothed or “plastic”, lacking erosion, asymmetry and surface irregularities characteristic of aged stone.
Hand Gestures	Occasional plausible mudrā configurations aligned with canonical forms.	Hands often mirrored, incomplete, or anatomically implausible (e.g., excess digits, distorted proportions).
Drapery & Garments	Certain reconstructions suggested gravity-driven folds and coherent garment layering.	Drapery frequently stylised, overly symmetrical, or cinematic, inconsistent with Gandhāran “wet-cloth” articulation.
Hair Texture	Limited instances reflected wavy strand patterns associated with Gandhāran sculpture.	Over-detailed or stylised “snail curls” inconsistent with historical precedent.
Pose & Anatomy	Balanced posture and proportional alignment occasionally achieved.	Missing limbs, mirrored anatomy and surface smoothing disrupted sculptural realism.
Lighting & Texture	Depth occasionally conveyed through balanced illumination and shadow.	Artificial lighting, glossy finishes and inconsistent light sources produced unrealistic appearance.
Scale & Composition	Some outputs maintained proportional relations between statue, niche and surrounding context.	Disproportionate scaling and weak spatial coherence between figure and environment.

Terrain & Rock Texture	Several outputs captured cliff stratification and rock texture plausibly.	Exaggerated terrain forms and jagged or fantastical peaks inconsistent with Bamiyan geology.
Caves & Cliff Features	Occasional attempts to integrate cave systems into cliff morphology.	Unrealistic cave number, scale and placement; misrepresentation of monastic complexes as decorative rock-cut forms.
Vegetation & Land Use	Some outputs suggested agricultural patterns consistent with regional land use.	Dense vegetation and lush landscapes inconsistent with the arid Bamiyan environment.
Architectural Context	Limited avoidance of extraneous structures in some reconstructions.	Introduction of Petra-like façades, temples and staircases without archaeological basis.
Creative Generation	Certain outputs were visually compelling and compositionally striking.	Over-generation of speculative elements (e.g., halos, ornate interiors) reduced historical plausibility.
Expert Suggestions	Recommendations included use of structural constraints (e.g., pose estimation, negative prompting) to improve anatomical and stylistic accuracy.	Persistent reliance on visually plausible but archaeologically unsupported features highlighted limitations of current generative approaches.

C. 7. Expert observations on strengths and inaccuracies identified during Phase I — Interdisciplinary Expert Screening (IES), covering both statue reconstruction outputs (Images 1–7; Figure 26) and environmental context reconstructions (Images 18–23, Figure 34).

ID	Subset	Platform-of-origin	Merge / notes
Image 1	Full figure	Sora-primary	Anchors overall proportions (58 m / 17 m / 8 m); minor drapery drift.
Image 2	Full figure	Gemini/Google	Surface photorealism; niche distorted, proportions off.
Image 3	Full figure	Hybrid	Sora base + Gemini refinements; improved niche but drapery unresolved.
Image 4	Full figure	Sora-primary	Preserved mudrā; minor facial over-smoothing.
Image 5	Full figure	Sora-primary	Realistic weathering; speculative halo artefact.
Image 6	Full figure	Hybrid	Cross-pass; corrected shoulder width; robe folds incomplete.
Image 7	Full figure	Gemini/Google	Dimensional anchors preserved; toes visible; robe monotone.
Image 8	Full figure	Sora-primary	Textural realism; incorrect hem logic.
Image 9	Full figure	Sora-primary	Proportional fidelity; polychromy lock retained; minor gesture drift.
Image 10	Full figure	Sora-primary	High niche fidelity; hand articulation unresolved.
Image 11	Full figure	Hybrid	Stronger robe shading; moustache hallucination suppressed.
drapery1	Drapery study	Sora-primary	First-step pleats; robe gravity misaligned.
drapery2	Drapery study	Gemini/Google	Surface folds improved; hem curvature distorted.
drapery3	Drapery study	Hybrid	Best convergence: stepped pleats, under-robe visibility.
head1	Head study	Sora-primary	Good cranial height; wave morphology unresolved.
head2	Head study	Gemini/Google	Smooth texture; speculative curls.
head3	Head study	Hybrid	Consensus leader: cranial canon, bun silhouette, facial fidelity.
head4	Head study	Sora-primary	Stable canon; minor ear/nose proportional drift.

C. 8. Provenance of the 18 shortlisted candidates, showing subset (full figure, drapery, head), platform of origin (Sora, Gemini/Google, hybrid) and targeted merges/corrective edits.

C.9 DPB-Derived Bamiyan Prompt Record

C.9.1 Scope and status

This section records the Bamiyan prompt material used during platform comparison, image generation and iterative refinement. It is included as paradata for the HARF workflow. The prompt record should not be read as a fixed benchmark set or as fifty standardised prompts. It documents the iterative prompt history of the Bamiyan workflow, including full-figure prompts, paired Western/Eastern Buddha prompts, component-specific prompts, viewpoint prompts, reference-image-conditioned instructions, masked or localised edit instructions, condensed versions, master versions and expert-informed refinements.

The prompts were developed through the Dynamic Prompt Blueprint (DPB). Initial prompts translated documentary, dimensional, material, iconographic and environmental evidence into generation instructions. Later prompts were refined after inspection of model failures and expert feedback. Refinement concentrated on recurrent instability domains: niche depth and outline, statue proportion, hand orientation, feet and toes, robe-fold grammar, material weathering, cave distribution, surface patina, pigment restraint and avoidance of unsupported visual additions.

The platform-selection comparison reported in Chapter 4 drew on this DPB-derived prompt archive. The resulting platform scores should therefore be interpreted as platform-selection judgements, not as benchmark measurements from a fixed, identical prompt set. After platform selection, the retained HARF configuration was used to produce the Bamiyan image corpus analysed in Chapter 5. This prompt record documents the broader generation and refinement history behind that corpus. It does not replace the Chapter 5 dataset map, which remains the authoritative account of the 245 generated Bamiyan images and the 18 shortlisted candidates used for specialist evaluation. Reference-image-conditioned and masked-edit instructions are retained because they formed part of the still-image refinement process. Prompt strings are reproduced as issued and therefore retain their original dating wording where applicable, including references to the “6th and 7th centuries”.

C.9.2 Prompt record

C.9.2.1 Initial paired Western/Eastern Buddha reconstruction

Status: *Early prompt-history record*

"Reconstruct a pair of monumental ancient Buddha statues carved into sandstone cliffs in Bamiyan, Afghanistan. The Western Buddha, standing at 58 metres, has broad shoulders (17 metres wide), a serene face and intricate flowing robes. Its head is 8 metres wide and the base spans 21 metres. The surrounding niche is tall and rectangular with faint traces of decorative patterns. The Eastern Buddha, standing at 38 metres, is more slender with shoulders 12 metres wide, a head 6 metres across and a base 16 metres wide. It also features flowing robes and a smaller ornate niche. Both statues were intricately detailed, representing Gandhara and Hellenistic artistic styles, painted in vibrant colours. The background consists of the rugged sandstone cliffs of Bamiyan, with eroded details revealing the remnants of their grandeur. Include a peaceful, majestic atmosphere as the statues stand as cultural and spiritual icons."

C.9.2.2 Initial Western Buddha version

Status: *Early prompt-history record*

"Reconstruct the Western Buddha, a monumental 58-metre-tall statue carved into the sandstone cliffs of Bamiyan, Afghanistan. The statue stands majestically in a tall, rectangular niche measuring 21 metres wide at the base. It features broad shoulders, 17 metres wide and a serene face with a head 8 metres across. The robes flow gracefully with intricate folds, symbolising peace and wisdom. The niche surrounding the statue is slightly eroded but retains traces of ornate carvings and patterns. The sandstone cliff forms the backdrop, capturing the rugged yet spiritual atmosphere of the sixth- to seventh-century CE Buddhist heritage. Render the statue with its original grandeur, blending Gandhara and Hellenistic artistic styles with vibrant, earthy tones and fine details in the robe and facial expression."

C.9.2.3 Initial Eastern Buddha version

Status: *Contextual prompt-history record*

"Reconstruct the Eastern Buddha, a 38-metre-tall statue carved into the sandstone cliffs of Bamiyan, Afghanistan. The statue is more slender than the Western Buddha, with shoulders 12 metres wide, a head 6 metres across and a base 16 metres wide. Standing in a narrow and slightly shorter niche, the statue radiates balance and tranquillity, with delicately carved robes flowing naturally. The niche retains subtle hints of intricate decorative patterns, though eroded over time. The sandstone cliff forms a striking contrast, emphasising the spiritual and cultural significance of this sixth- to seventh-century CE Buddhist heritage. Render the statue in its original splendour with Gandhara and Hellenistic artistic influences, using vibrant colours and fine detailing in the robes and posture."

C.9.2.4 Standing Buddha with early material and inscription descriptors

Status: *Superseded exploratory prompt*

"Reconstruct the standing Buddha of Bamiyan, 180 feet tall, carved into a sandstone cliff. Depict the Buddha with intricate robes influenced by Hellenistic and Indian styles, painted with molten bronze highlights. Show the serene face with wooden mask elements and details crafted in mud-straw and lime plaster. Surround the statue with the rugged Bamiyan Valley, emphasising the historical niches and carvings adorned with Kharosthi scripts."

C.9.2.5 Hyper-realistic Western Buddha with early colour descriptors

Status: *Superseded exploratory prompt*

"Render a hyper-realistic depiction of the western Bamiyan Buddha in its original form. The 180-foot figure stands within a sandstone niche, wearing carmine-red robes. Showcase the detailed folds, wavy curls of hair influenced by Gandhara art and gem-embellished adornments as described by Xuanzang. The mountainous terrain serves as a backdrop."

C.9.2.6 Bamiyan Valley and cave context

Status: *Early environmental prompt*

"Illustrate the Bamiyan Valley with its giant Buddhas carved into the cliffs, surrounded by smaller niches featuring brightly coloured frescoes. Highlight the blending of Greek, Persian and Indian artistic traditions with detailed architectural accuracy."

C.9.2.7 Western Buddha with dimensional and material anchors

Status: *Development prompt*

"Reconstruct the 58-metre-tall Western Buddha of Bamiyan, intricately carved into the sandstone cliffs in sixth- to seventh-century CE Afghanistan. Depict the statue standing majestically within a grand 21-metre-wide niche. Emphasise the 8-metre-tall head with its serene, meditative expression and detailed wavy curls of hair influenced by Gandharan art. The body features layers of robes, carefully modelled using a mud-straw mixture coated with lime plaster, painted in carmine red to highlight its regal appearance. Include details of the 17-metre-wide shoulders and the lower section of arms constructed with wooden armatures. Surround the niche with smaller carved caves and niches adorned with Kharosthi inscriptions and brightly coloured frescoes. Set this monumental statue against the rugged, mountainous terrain of the Bamiyan Valley to capture its cultural and geographical significance."

C.9.2.8 Eastern Buddha with dimensional anchors

Status: *Contextual prompt-history record*

"Recreate the 38-metre-tall Eastern Buddha of Bamiyan, housed within a 16-metre-wide sandstone niche. Highlight the 6-metre-tall head, crafted with intricate details such as serene eyes and delicate curls, reflecting a blend of Hellenistic and Indian artistic influences. Depict the 12-metre-wide shoulders and flowing robes modelled with a traditional mud-straw mixture, coated with lime plaster and painted in vibrant colours. Accentuate the serene, celestial posture of the Buddha, symbolising transcendence, with the surrounding niche featuring smaller caves carved into the cliffs. Include decorative elements, such as intricate carvings and remnants of the painted frescoes, which reflect the cultural amalgamation of Indian, Persian and Greek artistic styles. The Bamiyan Valley's rugged landscape serves as a dramatic backdrop, emphasising the historical and spiritual importance of the statue."

C.9.2.9 Bamiyan Valley setting with both statues

Status: *Environmental prompt*

"Depict the Bamiyan Valley with its iconic sandstone cliffs and carved niches, showcasing the two monumental Buddha statues in their original grandeur. Highlight the Western Buddha at 58 metres and the Eastern Buddha at 38 metres, emphasising their intricate niches and detailed construction. Surround the scene with smaller caves decorated with brightly coloured frescoes and Kharosthi inscriptions. Capture the texture and colouration of the sandstone, the expansive valley and the dramatic interplay of light and shadow to create an immersive view of this cultural heritage site."

C.9.2.10 Unified Western/Eastern Buddha

Status: *Development prompt*

"Reconstruct the two standing Buddhas of Bamiyan in their original grandeur, carved into sandstone cliffs in the Bamiyan Valley of Afghanistan during the 6th and 7th centuries. The Western Buddha, named Salsal (meaning 'Light Shines Through the Universe'), stands an imposing 58 metres tall within a 21-metre-wide niche, with shoulders spanning 17 metres. Its 8-metre-tall head is adorned with wavy curls of hair inspired by Gandharan art, reflecting the influence of Hellenistic, Indian and Persian styles. The statue wears flowing robes modelled using a mixture of mud and straw, coated with lime plaster and painted a vivid carmine red, with wooden armatures supporting the arms. The Eastern Buddha, known as Shahmama (the 'Queen Mother'), is 38 metres tall and stands in a 16-metre-wide niche, with shoulders measuring 12 metres and a 6-metre-tall head. The robes are

painted in vibrant multi-coloured hues and the serene expression captures the transcendent nature of the celestial Buddha.

Both statues are set within intricately carved niches, surrounded by smaller caves that served as monastic cells and were adorned with brightly coloured frescoes depicting Buddhist imagery. Include remnants of Kharosthi inscriptions and gems, copper and brass embellishments described in historical accounts. The robes flow naturally, with layers accentuating the delicate craftsmanship of the mud-straw mix, while the facial features, possibly crafted with wooden masks, exude calmness and spirituality.

The rugged sandstone cliffs of the Bamiyan Valley, with their natural ochre tones, rise majestically around the statues. Smaller niches and caves dot the surrounding cliffs, showcasing the vibrant artistic traditions of the Gandhara culture. The statues are depicted standing tall, reflecting their original monumental scale and their role as spiritual and cultural symbols. Capture the interplay of sunlight on the statues and cliffs, emphasising the historical and geographical context. The entire scene conveys the grandeur and cultural significance of the Bamiyan Buddhas, restored to their prime."

C.9.2.11 Revised unified with distance relation

Status: *Development prompt*

"Reconstruct the two monumental standing Buddhas of Bamiyan, carved into the sandstone cliffs of the Bamiyan Valley, Afghanistan, during the 6th and 7th centuries. The two Buddhas are positioned approximately 800 metres apart, reflecting their distinct spiritual and cultural significance.

The Western Buddha, named Salsal (meaning 'Light Shines Through the Universe'), stands at an imposing 58 metres (175 feet) tall within a 21-metre-wide niche. Its shoulders measure 17 metres and the 8-metre-tall head is adorned with wavy curls of hair inspired by Gandharan art, showcasing the fusion of Hellenistic, Indian and Persian artistic styles. The statue wears flowing robes modelled using a mixture of mud and straw, coated with lime plaster and painted in a vivid carmine red. The arms, constructed with wooden armatures, are detailed and lifelike. The serene expression of the face conveys its spiritual significance as the celestial Buddha.

The Eastern Buddha, known as Shahmama ('Queen Mother'), stands at 38 metres (120 feet) tall within a 16-metre-wide niche, located 800 metres east of the Western Buddha. Its shoulders measure 12 metres and the 6-metre-tall head reflects a calm, meditative expression. The robes, similarly crafted using mud and straw, are coated with lime plaster and painted in vibrant, multi-

coloured hues. Historical accounts suggest that parts of the statues may have been adorned with metals like copper and brass and their faces possibly featured wooden masks supported by rows of pegs visible in the rock.

Both statues are surrounded by an intricate network of smaller caves and niches, carved into the sandstone cliffs, which once served as monastic cells. These caves are embellished with brightly coloured frescoes and Kharosthi inscriptions, reflecting the Gandhara culture. The statues stand tall in the rugged Bamiyan Valley, with the cliffs showcasing their natural ochre tones, adding to the dramatic backdrop of the scene. The valley's monumental scale is emphasised by the sunlight highlighting the statues and their surrounding carvings, creating a vivid representation of the Bamiyan Buddhas in their prime.

This scene captures the spiritual, cultural and architectural grandeur of the Bamiyan Buddhas, restored to their original glory, with every detail, including their distance from each other, meticulously incorporated."

C.9.2.12 Western Buddha focused version

Status: *Development prompt*

"Reconstruct the Western Buddha of Bamiyan, known as Salsal (meaning 'Light Shines Through the Universe'), standing at a majestic 58 metres (175 feet) tall. The statue is carved into a 21-metre-wide niche in the sandstone cliffs of the Bamiyan Valley. Its shoulders measure 17 metres and the 8-metre-tall head is adorned with wavy curls of hair inspired by Gandharan art, reflecting a blend of Hellenistic, Indian and Persian artistic influences. The statue is depicted in a standing posture, wearing flowing robes modelled using a mixture of mud and straw, coated with lime plaster and painted a vivid carmine red. The arms, constructed with wooden armatures, are detailed and lifelike, emphasising the craftsmanship of the era.

The statue's serene face conveys a sense of spirituality and calm. The surrounding niche features intricately carved details and smaller caves that served as monastic cells, some adorned with brightly coloured frescoes and Kharosthi inscriptions. The sandstone cliffs of the Bamiyan Valley, with their natural ochre tones, rise dramatically around the statue, creating a monumental backdrop. The interplay of light and shadow highlights the intricate details, capturing the Western Buddha in its full grandeur as a symbol of spiritual and cultural significance."

C.9.2.13 Eastern Buddha focused version

Status: *Contextual prompt-history record*

"Recreate the Eastern Buddha of Bamiyan, known as Shahmama ('Queen Mother'), standing at an impressive 38 metres (120 feet) tall within a 16-metre-wide niche. This statue is located 800 metres east of the Western Buddha and depicts a standing Buddha with a serene, meditative expression. Its shoulders measure 12 metres and the 6-metre-tall head features wavy curls of hair, showcasing influences from Gandharan art and blending Hellenistic, Indian and Persian artistic traditions.

The statue's robes are modelled using a traditional mixture of mud and straw, coated with lime plaster and painted in vibrant, multi-coloured hues. The niche is surrounded by a network of smaller caves, carved into the sandstone cliffs, which were once adorned with brightly coloured frescoes and Kharosthi inscriptions. The serene setting of the Bamiyan Valley forms the backdrop, with its rugged cliffs emphasising the monumental scale of the Eastern Buddha. The sunlight plays across the textured surfaces, capturing the intricate craftsmanship and cultural significance of this transcendent statue."

C.9.2.14 Full standing Western Buddha correction

Status: *Development prompt*

"Reconstruct the Western Buddha of Bamiyan, known as Salsal (meaning 'Light Shines Through the Universe'), standing at an imposing 58 metres (175 feet) tall in its full standing posture. The statue is carved into a 21-metre-wide niche in the sandstone cliffs of the Bamiyan Valley, with its shoulders spanning 17 metres. Its 8-metre-tall head is adorned with intricate wavy curls of hair inspired by Gandharan art, blending Hellenistic, Indian and Persian artistic influences.

The Buddha wears a Hellenistic-style cloak draped over a long tunic, with flowing robes modelled using a mud-straw mixture, coated with lime plaster and painted a vivid carmine red. The statue's right hand points upward in a symbolic gesture, while the left hand points downward or holds an object, reflecting its spiritual and artistic significance. The arms, supported by wooden armatures, are crafted to appear detailed and lifelike, emphasising the remarkable craftsmanship of the era.

The image should show the Buddha as a complete standing figure from head to toes within the niche. The scene should emphasise the folds of the robes, the carved textures of the sandstone, the lime-plaster finish and the relationship between the statue and the surrounding cliff."

C.9.2.15 Head morphology

Status: *Regional prompt: head/hair*

"Generate a detailed image of the head of the Giant Buddha of Bamiyan, standing 58 metres tall, with an 8-metre-tall head. Show the serene, meditative expression and the intricate wavy curls of

hair, inspired by Gandharan art, blending Hellenistic, Indian and Persian artistic styles. Focus on the fine details of the lime plaster texture and carvings."

C.9.2.16 Right-hand

Status: *Regional prompt: right hand*

"Depict the right hand of the Giant Buddha of Bamiyan, carved with precision in a symbolic gesture pointing upward. Highlight the texture of the sandstone, the craftsmanship of the hand and the mud-straw mixture coated with lime plaster."

C.9.2.17 Left-hand

Status: *Regional prompt: left hand*

"Create a detailed image of the left hand of the Giant Buddha of Bamiyan, pointing downward or holding an object, as depicted in Gandharan art. Show the lifelike modelling supported by wooden armatures and coated with lime plaster."

C.9.2.18 Torso and robe

Status: *Regional prompt: drapery*

"Illustrate the flowing robes of the Giant Buddha of Bamiyan, draped in a Hellenistic cloak over a tunic, modelled using mud-straw and coated with lime plaster. Focus on the intricate folds and the vibrant carmine red paint."

C.9.2.19 Feet and toes

Status: *Regional prompt: feet/base*

"Generate a close-up of the feet and toes of the Giant Buddha of Bamiyan, standing on a sandstone pedestal. Highlight the base of the statue, the texture of the material and the fine detailing of the lime plaster."

C.9.2.20 Front view

Status: *Full-figure viewpoint prompt*

"Depict the Giant Buddha of Bamiyan in its full standing form, carved into a 21-metre-wide niche in the sandstone cliffs. Show the statue from a front-facing view, highlighting the right hand pointing upward, the left hand downward and the flowing Hellenistic cloak over a tunic, painted in carmine red."

C.9.2.21 Side view

Status: *Full-figure viewpoint prompt*

"Generate a side view of the Giant Buddha of Bamiyan, standing at 58 metres tall. Highlight the depth of the sandstone carving, the folds of the robes and the overall posture of the standing Buddha."

C.9.2.22 Back view

Status: *Full-figure viewpoint prompt*

"Create a rear view of the Giant Buddha of Bamiyan, carved into sandstone. Emphasise the integration of the statue into the cliff, showing the niche's depth and the texture of the sandstone."

C.9.2.23 45-degree view

Status: *Full-figure viewpoint prompt*

"Illustrate the Giant Buddha of Bamiyan from a 45-degree angle, capturing both the frontal details and the side depth. Highlight the statue's height, flowing robes and carved niche surroundings."

C.9.2.24 Niche and statue

Status: *Environmental / full-figure context prompt*

"Depict the Giant Buddha of Bamiyan standing within its 21-metre-wide niche, carved into the sandstone cliffs of the Bamiyan Valley. Show the full niche, including its arch-like design, with the statue standing inside, surrounded by detailed carvings."

C.9.2.25 Niches and surrounding caves

Status: *Environmental context prompt*

"Generate a zoomed-out view of the Giant Buddha of Bamiyan, surrounded by smaller niches and caves carved into the sandstone cliffs. Highlight the vibrant frescoes and carvings within the caves, reflecting the Buddhist monastic tradition."

C.9.2.26 Panoramic Bamiyan Valley

Status: *Environmental context prompt*

"Create a panoramic image of the Bamiyan Valley, showcasing the Giant Buddha within its niche, surrounded by the rugged sandstone cliffs and smaller monastic caves. Emphasise the natural ochre tones of the cliffs and the monumental scale of the statue."

C.9.2.27 Right hand, standing posture

Status: *Regional prompt: right hand*

"Generate a close-up of the right hand of the Giant Buddha of Bamiyan, in a standing posture, positioned with all fingers pointing upward in a symbolic gesture. Show the lifelike details of the sandstone carving, the texture of the mud-straw mixture coated with lime plaster and the proportions of the hand relative to the full standing statue. The background should hint at the vertical posture of the statue within the niche."

C.9.2.28 Left hand, standing posture

Status: *Regional prompt: left hand*

"Create a close-up of the left hand of the Giant Buddha of Bamiyan, in a standing posture, positioned pointing downward or holding an object, as depicted in Gandharan art. Highlight the texture of the sandstone, the lime plaster coating and the gesture's symbolic significance, ensuring the vertical alignment matches the standing figure."

C.9.2.29 Right hand from multiple angles

Status: *Regional viewpoint prompt*

"Generate a close-up of the right hand of the Giant Buddha of Bamiyan, seen from the front. The hand is positioned in a standing gesture with fingers pointing upward. Focus on the intricate details of the sandstone texture, lime plaster coating and proportions of the hand relative to the body."

"Create a detailed image of the right hand of the Giant Buddha of Bamiyan, viewed from the left side. Emphasise the curve of the palm and fingers from the side perspective, showing the standing posture of the Buddha and the detailed folds of the robes in the background."

"Illustrate the right hand of the Giant Buddha of Bamiyan, viewed from the right side. Highlight the contours of the fingers and palm, showing the hand in its upward symbolic gesture. Include details of the sandstone carvings and the lime plaster surface from this angle."

"Generate a close-up of the back view of the right hand of the Giant Buddha of Bamiyan. Emphasise the rear texture of the palm, fingers and wrist, showing the intricate carvings and the statue's integration into the niche. Ensure the standing posture of the statue is evident."

C.9.2.30 Back and side views of both hands

Status: *Regional viewpoint prompt*

"Generate a close-up of the back of the right hand of the Giant Buddha of Bamiyan, as seen from a backward angle (180 degrees). The hand is part of a standing Buddha with the fingers pointing upward in a symbolic gesture. Highlight the texture of the sandstone, the rear of the palm and the integration with the statue's arm."

"Create a side view of the right hand of the Giant Buddha of Bamiyan, showing the hand in a symbolic upward gesture from a side angle (90 degrees). Emphasise the profile of the palm, fingers and the connection to the arm, maintaining the standing posture of the statue."

"Illustrate a close-up of the back of the left hand of the Giant Buddha of Bamiyan, as seen from a backward angle (180 degrees). The hand points downward in a symbolic gesture, with the rear of the palm and fingers clearly visible. Highlight the sandstone texture and proportionality."

"Generate a side view of the left hand of the Giant Buddha of Bamiyan, showing it from a side angle (90 degrees). Focus on the profile of the hand and its downward gesture, emphasising the details of the palm and fingers from this perspective."

C.9.2.31 Back view from Buddha perspective

Status: *Full-figure viewpoint prompt*

"Generate a detailed image of the Giant Buddha of Bamiyan, viewed from behind as if seen from the Buddha's own eye level looking backward at the niche and surrounding sandstone cliffs. Maintain the statue's standing posture, focusing on the rear of the head, shoulders and back. Show the texture of the sandstone and the lime plaster, with the surrounding niche and caves visible in the background. The arms and hands should retain their original positions but seen from behind, reflecting the full 180-degree perspective."

C.9.2.32 Back side of right hand

Status: *Regional viewpoint prompt*

"Generate an image of the right hand of the Giant Buddha of Bamiyan, viewed from the back side (180-degree perspective). The hand should be positioned upward in a symbolic gesture, with all fingers pointing upward as part of the standing Buddha statue. Show the back of the hand and fingers, emphasising the texture of the sandstone, lime plaster coating and the hand's proportionality to the statue's standing posture. Exclude any frontal elements or sitting gestures, focusing solely on the rear view of the hand in its upright position."

C.9.2.33 Sandstone cliffs and monastic caves

Status: *Environmental context prompt*

"Generate a highly detailed and realistic depiction of the sandstone cliffs and caves of the Bamiyan Valley, showcasing the vast network of ancient rock-cut monastic caves surrounding the niches where the Giant Buddhas once stood. Emphasise the rugged textured cliffs, naturally eroded over centuries, with layers of weathered sandstone. The caves should vary in size and depth, some appearing as small openings, while others feature intricate carved entrances and niches. Include subtle traces of Kharosthi inscriptions and faint remnants of Buddhist murals inside some of the caves, reflecting their historical and spiritual significance. The scene should be bathed in natural lighting, with soft shadows cast by the cliffs, capturing the grandeur and serenity of this UNESCO World Heritage site."

C.9.2.34 Evidence-constrained Western Buddha

Status: *Refined still-image prompt*

"Reconstruct the Western Buddha of Bamiyan, a monumental 58-metre-tall statue carved into the weathered sandstone cliffs of Bamiyan, Afghanistan. The statue stands gracefully in a tall, deep, rectangular niche approximately 21 metres wide at its base. Depict broad, anatomically accurate shoulders measuring about 17 metres across and a calm, symmetrical face with a head spanning 8 metres. The robes should flow naturally, featuring deeply carved, layered folds with subtle erosion and texture, evoking movement and craftsmanship. Render the surrounding cliff as naturally rugged and eroded, with remnants of ornate Gandharan and Hellenistic carvings faintly visible on the inner walls of the niche. Ensure the stone texture reflects realistic ageing, weathering and shadow interplay. Use warm, earthy tones like ochre, terracotta and faded sienna, avoiding oversaturation. Highlight subtle surface imperfections, nuanced lighting and a sense of scale. The right hand is raised in a teaching gesture, the left hand facing downward, both carved with realistic anatomical detail. The overall composition should evoke the lost grandeur of the sixth- to seventh-century CE spiritual heritage."

C.9.2.35 Structured expert-informed example

Status: *Refined still-image prompt*

"Visualise the Western Buddha of Bamiyan as it stood in the 6th–7th century CE. Depict the 180-foot-tall Buddha within its sandstone niche, showcasing detailed Gandharan features including wavy hair without spiral tufts, no moustache and garments in Greco-Buddhist style with stiff, clearly defined folds resembling a wet cloth. Position the right hand raised in a teaching gesture and the left hand holding drapery downward. Include smooth niche walls with natural erosion and subtle

remnants of carvings. Surround the Buddha with accurately represented sub-rectangular cave openings, realistic geological features and historically documented monastery structures. Reflect accurate pigmentation and realistic weathering, ensuring a composition that evokes the authentic, lost grandeur of this significant heritage site."

C.9.2.36 Cultural visualisation

Status: *Refined still-image prompt*

"Generate a historically grounded visualisation of the Western Buddha of Bamiyan (Salsal), Afghanistan, dated to the 6th–7th century CE. The statue stands upright in Abhaya mudra pose, with accurate proportions (55 metres height, 17 metres shoulder width, 8 metres head height). Position the figure within a tall, unevenly carved sandstone niche (approximately 58 metres high and 21 metres wide), integrated into the vertical cliff face of the Bamiyan Valley. Depict the Buddha with wavy Gandharan-style hair, a composed facial expression reflecting Buddhist serenity and flowing Hellenistic-style robes with realistic, stepped folds, avoiding modern or stylised interpretations. Include historically verified surface features such as faded polychromy, mud-straw plaster, signs of erosion, patina, weathering cracks and missing facial or hand elements where appropriate.

Surround the niche with approximately 750 sub-rectangular caves of varied sizes (monastic cells), unevenly distributed across the cliff, avoiding symmetrical placement. The environment should reflect the rugged terrain of the Bamiyan Valley: dry, mountainous and stratified, without exoticised or fictional additions such as Petra-like architecture, domes, or vegetation. Lighting should mimic natural shadow interplay with the cliff and statue, based on early morning or late afternoon sun angles.

Avoid hallucinated features: no carved halos, overly smooth surfaces, speculative frescoes, or symmetrical temple façades. Use visual cues from surviving documentation, 1970s photogrammetry, ICOMOS reports and comparative Gandharan sculptures. This is a cultural visualisation, not a literal reconstruction. Emphasise accuracy over photorealism."

C.9.2.37 Full-body photorealistic archival-style

Status: *Refined still-image prompt*

"Full-body photorealistic image of the Western Bamiyan Buddha (Salsal), Afghanistan, 6th–7th century CE carved into a deep sandstone niche. The statue must look like an ancient eroded monument, not a modern restoration.

Head and hair: serene Gandharan face, elongated earlobes, no moustache. Hair carved only as parallel wavy lines, absolutely no snail-shell curls. Face weathered with softened lips and nose edges, showing erosion.

Hands: right hand raised in Abhaya mudra. Left hand clearly holds a fold of the robe downward. Fingers blunt and worn, not smooth or human-like.

Garments: Greco-Buddhist wet drapery style with stiff, crisp stone folds clinging to the body. Heavy layered robe with cracks, tool-mark textures and weathering.

Feet: five toes on each foot, blunt, squared, carved into stone, resting on shallow raised stone steps within the niche.

Niche and surroundings: deep, shadowed sandstone recess. Cliff face gritty and rough, with stratified layers. Honeycomb of rectangular caves, some collapsed. Small, weathered mud-brick monastic houses at the base, sun-bleached and dusty.

Colour and texture: patchy sandstone with cracks. Surface should show faint, faded polychromy: thin white gesso undercoat with very subtle traces of red-orange pigment, flaking and uneven. Avoid smooth uniform beige.

Lighting: bright midday sun, deep shadows in folds and caves. Photographed like a 35mm film archival photo with subtle film grain and slight dust haze.

Do not generate: snail-shell hair curls, smooth modern restoration, polished surfaces, cartoon/CGI, lotus base, halo, vibrant colours, living human skin."

C.9.2.38 Condensed detail-retention version

Status: *Refined still-image prompt*

"Full-body Western Bamiyan Buddha (Salsal), Afghanistan, 6th–7th century CE inside a deep arched niche. Gandharan face, elongated ears, no moustache; hair = wavy carved lines, not snail curls; lips and nose edges eroded. Right hand Abhaya, left hand holding robe fold downward. Greco-Buddhist wet drapery with crisp, stiff folds like stone cloth; tool-mark texture and small cracks. Feet with five toes each, on raised stone steps. Cliff with honeycomb caves and small mud-brick houses at the base. Material: buff sandstone with white gesso and faded red-orange traces, patchy and weathered; no bright carmine, no gloss. Midday sun, high contrast, slight dust haze; 35mm film look. Exclude modern restorations, halo, pedestal, cartoon/CGI, saturated colours, snail-shell hair, moustache, living-person skin."

C.9.2.39 Master full-fidelity version

Status: *Refined still-image prompt*

"Photorealistic archival-style photograph of the Western Bamiyan Buddha (Salsal), Afghanistan, 6th–7th century CE before 2001, shown full body and centred inside a deep arched sandstone niche, with no crop. Monumental carved stone statue, not a living person.

Head and face: Gandharan style, calm/serene monumentality; no moustache. Hair = wavy parallel carved lines, not snail-shell curls, with a subtle hair band. Elongated earlobes, lips slightly weathered; nose and lip edges softened by erosion; no glossy or skin-like details.

Hands and pose: right hand raised in Abhaya mudra at chest height. Left hand holds a fold of the robe downward at the hip/thigh. Fingers worn and less defined by erosion.

Garments: Greco-Buddhist wet drapery. Stiff, crisp Hellenistic folds that cling to the body as wet cloth, with two layered drapery zones across chest and legs. Heavy stone-cloth appearance with minor cracks, chipped edges and tool-mark textures.

Feet: five toes on each foot, blunt and squared as stone carving. Feet stand on slightly raised double stone steps inside the niche, not flat ground.

Niche and surroundings: niche walls rough, irregular, stratified sandstone; deep shadowed recess. Faint, fragmentary mural traces on the upper arch only. Cliff face around the niche shows a honeycomb of sub-rectangular caves of varied sizes; some collapsed openings; deep shadows inside. At the base, small modest mud-brick monastic buildings, sun-bleached, dusty, flat-roofed and integrated into the talus slope.

Material, colour and weathering: buff sandstone/stucco with tool marks, cracks and chips; granular texture. Patchy original polychromy: thin white gesso undercoat with subtle, faded reddish-pink/orange traces, not vibrant. Avoid carmine-red, metallic, glossy or uniform colouring.

Lighting and camera: midday sun, high contrast; deep shadows in caves and folds. Shot like a 35mm film photo, approximately 50 mm, f/8, ISO 100, with subtle film grain and slight dry dust haze for scale. Front view, full body from head to feet and steps; include cliff and houses on both sides.

Do not generate: freestanding statue, pedestal or lotus base, halo/aureole, restorations, scaffolding, signage, tourists, vegetation, cartoon/CGI, saturated colours, Thai/Chinese iconography, snail-shell hair, moustache, polished smooth new surface, or living human skin pores."

C.9.2.40 Reference-fusion instruction

Status: *Refined still-image / reference-conditioned instruction*

"Generate a rigorously accurate, photoreal full-body reconstruction of the Western Buddha of Bamiyan inside its niche, using the head/hair reference for hair waves, bun silhouette and facial proportions while restoring the statue's damaged areas. Produce a real-location look, not a studio render, with correct scale and geology for Bamiyan Valley.

Reference-fusion rules: use the head reference only for hair waves, bun silhouette and facial proportions. Use the environment plate only for cliff geology, cave distribution, camera height and colour palette. Ignore all statue damage in the environment plate. Do not inherit blast pits, missing surfaces or broken folds from the environment plate. Overwrite the statue region with the reconstructed geometry described below.

Measured geometry: statue height = 58 m; head height = 8 m; shoulder width = 17 m; niche height = 58 m; niche base width = 21 m; niche depth = 18 m. Preserve the proportional ratios: head/height approximately 0.138; shoulders/height approximately 0.293; base/height approximately 0.362; depth/height approximately 0.310.

Statue design: 6th–7th century CE; Greco-Buddhist/Gandharan style. Material: matte grey-beige sandstone; granular surface with fine chisel/tool marks; realistic erosion on edges only; no glossy or wet shader; no human skin pores. Face: serene symmetry, thick lips, elongated ears, centred urna, three shallow neck lines, no moustache. Hair: rigid parallel S-waves swept upward into a compact rounded bun; no spiral/snail curls or braids. Robes: Hellenistic stepped folds that cling like wet cloth; two undergarment layers visible under the right arm and at the lower hem. Hands: right hand in Abhaya mudra with planar, stylised stone geometry; left hand grips a descending robe fold, back of hand visible, elongated stiff fingers integrated with the fold mass. Feet: both feet fully visible on the base; five toes per foot, clearly separated by shallow carved grooves.

Niche, cliff and valley: deep arched recess; smooth/eroded walls; subtle remnant carvings; white preparatory layer around the head. Layered sandstone with horizontal bedding planes and vertical joints; talus ledges; ambient occlusion in cavities; no melted sand-castle look. Caves: sub-rectangular openings with dark interiors, scaled consistently to the 58 m statue; some wooden lintels or doors. Settlement: small mud-brick houses at the base and a low monastery complex forward but not occluding the toes.

Camera and output: 3840 x 2160, 16:9. Strict front, orthographic feel; tripod-locked. Entire niche plus adjacent cliff edges and valley floor visible; both feet and all ten toes unobstructed. Lens: 35 mm equivalent, minimal distortion. Midday sun around 5200K, strong but natural contrast, soft bounce from cliff, slight warm dust haze, subtle 35 mm film grain.

Negative constraints: no moustache; no spiral curls or braids; no glossy/wet stone; no human skin texture; no modern scaffolding, crowds or signage; no tilted camera; no HDR halos or monochrome brown cast; no sand-sculpture look; do not copy statue damage from the environment plate."

C.9.2.41 Localised correction instruction

Status: *Refinement paradata*

"Keep the statue, face, hair, niche, scale and proportions exactly the same. Correct the missing parts: sculpt the left hand gripping a vertical robe fold, with the back of the hand visible and elongated carved fingers. Refine the robe folds into Hellenistic stepped folds that cling like wet cloth, with at least two under-layers visible near the right arm and lower hem. Stylise both hands and toes with planar, stone-carved geometry, not fleshy anatomy. Show five toes per foot with crisp chisel edges. Make the caves sub-rectangular, arranged in rows, with some wooden lintels and gates. Add a few mud-brick houses and a small monastery on the valley floor, placed forward but not covering the feet. Add a faint white preparatory layer around the head and upper niche. Lighting, sandstone texture and camera framing stay exactly the same."

Appendix D. Supplementary Materials for Chapter 6 (Community-Grounded Authenticity at Bamiyan)

Purpose

This supplementary appendix supports Chapter 6, Community-Grounded Authenticity at Bamiyan. Its purpose is to make the statistical claims reported in Chapter 6 independently reproducible and to provide the demographic, cross-tabulation, coding and reproducibility detail that is too extensive for the main chapter. Section S1 documents the ethics approval and consent procedure. Section S2 reproduces the questionnaire instrument in full. Section S3 presents the complete demographic composition of the sample, including current-location detail. Section S4 reports full descriptive tables with Wilson 95% confidence intervals for every closed item. Section S5 reports multi-select frequency tables with 95% confidence intervals. Section S6 reports subgroup contrasts with chi-square statistics, p-values and bias-corrected Cramér's V. Section S7 reports the population-representativeness comparison. Section S8 gives the Hazara sub-sample breakdown. Section S9 describes the thematic coding scheme and theme frequencies. Section S10 lists the accompanying reproducibility files.

S1. Ethics approval, consent procedure and anonymisation

Ethics approval: The study was approved by the Brunel University of London Research Ethics Committee under reference 50187-LR-Jan/2025-53624-2. The same approval covers the expert questionnaires used in the paired HARF evaluation (Arzomand et al., 2026; Arzomand & Kalganova, 2026b), so that all questionnaire components of the wider research programme operate under a single approved protocol.

Consent procedure: Respondents encountered an information page describing the study purpose, estimated completion time (15–25 minutes), voluntary participation, right to withdraw before submission and the investigator's contact details. Consent was indicated by continuing past that page. Responses were stored in Brunel's secure research repository under UK GDPR, with access restricted to the research team.

Anonymisation measures applied to the released dataset: The anonymised CSV released with this supplement (dataset_anonymised.csv) differs from the raw export in three ways: (i) the submission timestamp has been removed, as it could narrow respondents to small temporal cohorts; (ii) the raw country-of-residence field has been grouped to four categories — Inside Afghanistan, Diaspora: Europe, Diaspora: North America, Diaspora: Other — to reduce re-identification risk for respondents from countries represented by only one or two respondents; (iii) open-ended response fields have been truncated to 500 characters per item. The anonymised dataset contains a synthetic respondent identifier in random order that does not map to any external identifier.

Quotation policy: No individual-level identifier strings, that is, combinations of ethnicity, gender and country of residence, are attached to quotations in either the main thesis chapters or this supplement. Cell sizes at these intersections are too small to permit individualising attribution without risking re-identification. Quotations used in the study are reported with a theme tag and a frequency count only.

S2. Questionnaire instrument

The instrument was titled *Memory, Absence and Digital Reconstruction of Lost Cultural Heritage in the Bamiyan Valley* and was hosted on Google Forms. It comprised forty-six items across six thematic blocks. Item formats were single-select, multi-select, or short open-ended. The full item list is reproduced below with response options, in the order presented to respondents. Items marked [multi] permitted multiple selections; items marked [open] received free-text responses.

Block A — Demographic context

- S1.** What is your ethnic identity? (single-select; Pashtun / Hazara / Tajik / Arab / Sadaat / It does not matter / [other])
- S2.** Where did you grow up? (single-select; Kabul or another major city / Bamiyan / Other Hazara-majority area (Daikundi, Ghazni, Samangan) / Herat / Laghman / Nangarhar / Khost / Outside Afghanistan / [other])
- S3.** Your age group (single-select; under 20 / 21-35 / 36-50 / over 50)
- S4.** Gender (optional; Male / Female / [other] / prefer not to say)
- S5.** Where do you currently live? (free text country)

Block B — Connection and memory

- S1.** In which of the following ways, if any, do you feel a connection to the Bamiyan Valley? [multi] (Culturally (stories, ancestry, traditions) / Emotionally (grief, pride, love) / Personally (family history, visits, lived experience) / Politically (site of injustice or identity struggle) / Spiritually (sacred meanings, religious or mythic value) / I do not feel a strong connection)
- S2.** Do you know these statues? (Yes / No)
- S3.** What do these statues represent for you? [open]
- S4.** By what names or identities do you know the Bamiyan Valley statues? [open]
- S5.** Which of the following best reflects how you identify the statues? (single-select)
- S6.** Who are Salsal and Shahmama? [open]
- S7.** Were these statues ever discussed in your home, schools, or media? [open]
- S8.** Have you ever visited the site? (Yes, before destruction / Yes, after destruction / No)
- S9.** How did you feel when you were there or saw it? [open, optional]
- S10.** Did your parents, grandparents, or community share stories about the statues? (Yes and I still remember them / Yes, but vaguely / No / I am not sure)
- S11.** If yes, please share an example or story [open]

S12. Have you heard oral stories, myths, or spiritual beliefs linked to the statues? [open]

S13. What do the statues of Bamiyan mean to you personally, historically, or spiritually? [open]

Block C — Interpretations of destruction

S1. In your opinion, what was the intended target or message behind the destruction? [open]

S2. What do you think Afghanistan or Bamiyan lost economically, if anything? [open]

S3. What do you think was lost when the statues were destroyed in 2001? [open]

S4. Do you believe anything survived, despite the destruction? [open]

S5. What do you think the Taliban wanted to communicate by destroying the statues? [multi] (e.g. a religious statement / a political message to the international community / targeting the Hazara community / asserting ideological control / other)

S6. Do you believe the international community cared more about the statues than Afghan people? (Yes / No / Not sure)

Block D — Reconstruction preferences and ethics

S1. Do you think the statues should be reconstructed? (Yes / No / Not sure)

S2. In which form should they be reconstructed? [multi] (Physically / Digitally (images) / Virtually (3D/VR/AR) / Holograms / I have no preference)

S3. What exactly should be reconstructed? [multi] (Materiality as close as possible to the original / Materiality not necessarily identical / Skills and knowledge associated with their creation / Their destruction and its historical context / What they represented (meaning, symbolism) / Nothing should be reconstructed)

S4. How should absence or destruction be represented in future reconstructions? [multi] (Fully restored digitally / A combination of past and present (e.g. holograms) / Only the ruined niches, untouched / I don't support any reconstruction)

S5. Do you feel Hazara voices are included in reconstruction discussions? (Yes / No / I don't know)

S6. Please explain your choice [open]

S7. Whose voices have been missing in global or national discussions? [open]

S8. Should digital reconstruction aim to heal the past or expose its wounds? Or both? Or neither?

S9. Who should lead or guide future reconstruction efforts? [multi] (International heritage organisations (e.g. UNESCO) / Afghan government / Afghan archaeologists and cultural historians / Heritage scholars / Local Hazara community / Buddhist monks / All equally)

S10. If the statues are rebuilt, what should be the guiding principle? [multi] (Historical accuracy / Community healing / Cultural resistance or protest / Tourism and economy / Memorialisation of trauma)

S11. Should reconstructions acknowledge multiple meanings? (Yes — both Buddhist and Hazara stories / Yes — but only Buddhist history / No / Not sure)

Block E — Intangible and sensory dimensions

- S1.** If digital reconstructions are made, what should they include? [multi] (Accurate visual appearance / The caves and murals / Buddhist spiritual symbolism / Economic impact / Hazara oral histories / Environmental and sensory experiences (wind, sound, light) / Ruins and the trauma of destruction / The Taliban's justification and international reactions / Soundscapes of the valley / Nothing — they should remain absent)
- S2.** How do you feel about using AI or digital technologies to reconstruct the statues? [open]
- S3.** What risks do you see in digital reconstruction? [open]
- S4.** What is impossible to reconstruct digitally? [open]
- S5.** What concerns do you have about using AI to recreate destroyed heritage? [multi] (AI might oversimplify or misrepresent cultural meanings / may impose Western views of heritage / may ignore local voices and knowledge / could create fake or misleading versions of history / I have no concerns)
- S6.** What parts of your heritage do you think AI cannot or should not try to reconstruct? [open]
- S7.** Do you think intangible or sensory elements are relevant and worth including in digital reconstructions? [open]
- S8.** Should sensory elements be included in digital reconstructions? (Yes / No / Not sure)
- S9.** What would you prioritise? (Echo of Prayer / Wind in the valley / Scent of Earth / Light and shadow)

Block F — Open reflections

- S1.** What message do you want the world to remember (if any) about the statues of Bamiyan? [open]

S3. Sample composition (full demographic breakdown)

Section 6.1 of the main chapter gives a summary of the sample composition. Tables S1–S4 below provide the full distributions, including the current-location detail. All figures are counts over $N = 62$ unless otherwise noted; two respondents did not provide an ethnicity response and one did not provide a gender response.

Table S1. Ethnic self-identification ($N = 62$).

Category	n	%
Pashtun	26	41.9%
Hazara	13	21.0%
Tajik	12	19.4%
It does not matter	6	9.7%
Arab	2	3.2%
No response	2	3.2%
Sadaat	1	1.6%

Table S2. Age band (N = 62).

Age band	n	%
Under 20	1	1.6%
21-35	35	56.5%
36-50	18	29.0%
Over 50	8	12.9%

Table S3. Gender (N = 62).

Gender	n	%
Male	43	69.4%
Female	18	29.0%
NaN	1	1.6%

Table S4. Current country of residence (N = 62). Inside/diaspora categorisation used elsewhere is derived from this column.

Country	n	%
Afghanistan	25	40.3%
USA	11	17.7%
UK	8	12.9%
Germany	3	4.8%
Turkey	3	4.8%
Canada	2	3.2%
Ireland	2	3.2%
France	1	1.6%
Brazil	1	1.6%
Portugal	1	1.6%
Iran	1	1.6%
Belgium	1	1.6%
Australia	1	1.6%
Romania	1	1.6%
Not Reported / Blank	1	1.6%

Table S5. Place of upbringing (N = 62). Several respondents grew up in Kabul or major cities and later moved to the diaspora.

Region of upbringing	n	%
Kabul or another major city	49	79.0%
Bamiyan	3	4.8%

Region of upbringing	n	%
Other Hazara-majority area (e.g., Daikundi, Ghazni, Samangan)	3	4.8%
NaN	2	3.2%
Herat	1	1.6%
Laghman	1	1.6%
Nangarhar	1	1.6%
Outside Afghanistan	1	1.6%
Khost	1	1.6%

S4. Full single-select items with Wilson 95% confidence intervals

All closed single-select items with 95% Wilson (score) confidence intervals. The Wilson interval is used rather than the Wald approximation because it provides more reliable coverage at extreme proportions ($\leq 10\%$ or $\geq 90\%$), several of which appear below.

Table S6. Single-select responses with Wilson 95% CIs (N = 62).

Item	n	Proportion	95% CI
Support reconstruction (Yes)	51	82.3%	71.0-89.8%
Oppose reconstruction (No)	5	8.1%	3.5-17.5%
Uncertain about reconstruction	6	9.7%	4.5-19.5%
Sensory inclusion (Yes)	38	61.3%	48.8-72.4%
Sensory inclusion (Not sure)	17	27.4%	17.9-39.6%
Sensory inclusion (No)	7	11.3%	5.6-21.5%
Hazara voices: Yes	19	30.6%	20.6-43.0%
Hazara voices: No	12	19.4%	11.4-30.9%
Hazara voices: Don't know	31	50.0%	37.9-62.1%
Knew statues before destruction (Yes)	59	95.2%	86.7-98.3%
Heal/expose: Both	37	59.7%	47.3-71.0%
Heal/expose: Heal the past	18	29.0%	19.2-41.3%
Heal/expose: Expose its wounds	2	3.2%	0.9-11.0%
Heal/expose: Neither	5	8.1%	3.5-17.5%
Intl cared more: No	20	32.3%	22.0-44.6%
Intl cared more: Not sure	23	37.1%	26.2-49.5%
Intl cared more: Yes	17	27.4%	17.9-39.6%

S5. Multi-select item frequencies with 95% CIs

Multi-select items permitted respondents to select any number of options, including none. Total selections therefore exceed N = 62 by design and proportions below denote the proportion of respondents who selected the option, not a share of total selections. Confidence intervals are Wilson 95%. Items selected by only one respondent (likely write-in responses or idiosyncratic answers) are not tabulated.

Table S7. Modes of connection to the Bamiyan Buddhas (multi-select; N = 62).

Option	n	%	95% CI
Culturally (stories, ancestry, traditions)	48	77.4%	65.6-86.0%
Emotionally (grief, pride, love)	24	38.7%	27.6-51.2%
Personally (family history, visits, lived experience)	18	29.0%	19.2-41.3%
Politically (site of injustice or identity struggle)	15	24.2%	15.2-36.2%
Spiritually (sacred meanings, religious or mythic value)	10	16.1%	9.0-27.2%
I do not feel a strong connection	3	4.8%	1.7-13.3%

Table S8. What should be reconstructed (multi-select; N = 62).

Option	n	%	95% CI
The materiality of the statues as close as possible to the original	52	83.9%	72.8-91.0%
Skills and knowledge associated with their creation	24	38.7%	27.6-51.2%
Their destruction and its historical context	21	33.9%	23.3-46.3%
What they represent	20	32.3%	22.0-44.6%
The materiality of the statues	8	12.9%	6.7-23.4%
not necessarily identical to the original	8	12.9%	6.7-23.4%
Nothing	4	6.5%	2.5-15.4%

Table S9. Preferred form of reconstruction (multi-select; N = 62).

Option	n	%	95% CI
Physically	49	79.0%	67.4-87.3%
Digitally (images)	19	30.6%	20.6-43.0%
Virtually (3D/VR/AR)	18	29.0%	19.2-41.3%
Holograms	14	22.6%	14.0-34.4%
I have no preference	11	17.7%	10.2-29.0%

Table S10. How absence should be represented (multi-select; N = 62).

Option	n	%	95% CI
A combination of past and present (e.g., holograms)	37	59.7%	47.3-71.0%
Fully restored digitally	30	48.4%	36.4-60.6%
Only the ruined niches	8	12.9%	6.7-23.4%

Option	n	%	95% CI
untouched	8	12.9%	6.7-23.4%
I don't support any reconstruction	3	4.8%	1.7-13.3%

Table S11. Who should lead reconstruction (multi-select; N = 62).

Option	n	%	95% CI
International heritage organizations (e.g., UNESCO)	39	62.9%	50.5-73.8%
Afghan archaeologists and cultural historians	37	59.7%	47.3-71.0%
Heritage scholars	26	41.9%	30.5-54.3%
Afghan government	25	40.3%	29.0-52.7%
Local Hazara community	18	29.0%	19.2-41.3%
All equally	16	25.8%	16.6-37.9%
Buddhist monks	13	21.0%	12.7-32.6%

Table S12. Guiding principles for reconstruction (multi-select; N = 62).

Option	n	%	95% CI
Historical accuracy	52	83.9%	72.8-91.0%
Tourism and economy	40	64.5%	52.1-75.3%
Cultural resistance or protest	28	45.2%	33.4-57.5%
Community healing	24	38.7%	27.6-51.2%
Memorialization of trauma	18	29.0%	19.2-41.3%

Table S13. What digital reconstructions should include (multi-select; N = 62).

Option	n	%	95% CI
Accurate visual appearance of the statues	53	85.5%	74.7-92.2%
The caves and murals	31	50.0%	37.9-62.1%
Buddhist spiritual symbolism	26	41.9%	30.5-54.3%
Economic impact (e.g., loss of tourism)	26	41.9%	30.5-54.3%
Hazara oral histories	22	35.5%	24.7-47.9%
Environmental and sensory experiences (wind, sound, light)	15	24.2%	15.2-36.2%
Ruins and the trauma of destruction	13	21.0%	12.7-32.6%
The Taliban's justification and international reactions	11	17.7%	10.2-29.0%
Soundscapes of the valley (prayer echoes, silence, wind)	8	12.9%	6.7-23.4%
Nothing — they should remain absent	2	3.2%	0.9-11.0%

Table S14. Concerns about AI-mediated reconstruction (multi-select; N = 62).

Option	n	%	95% CI
AI might oversimplify or misrepresent cultural meanings	36	58.1%	45.7-69.5%
It may ignore local voices and knowledge	19	30.6%	20.6-43.0%
It could create fake or misleading versions of history	18	29.0%	19.2-41.3%
I have no concerns	16	25.8%	16.6-37.9%
It may impose Western views of heritage	13	21.0%	12.7-32.6%

Table S15. Perceived Taliban message in destroying the statues (multi-select; N = 62).

Option	n	%	95% CI
A religious act (against “idols” or shrines of unbelievers)	42	67.7%	55.4-78.0%
A rejection of non-Islamic (Buddhist) heritage in Afghanistan	29	46.8%	34.9-59.0%
A political message to the international community	25	40.3%	29.0-52.7%
A rejection of Western cultural values and the “cult of art”	15	24.2%	15.2-36.2%
A punishment or intimidation targeting the Hazara community	15	24.2%	15.2-36.2%
An act designed for global media attention and symbolism	15	24.2%	15.2-36.2%
I don’t know	9	14.5%	7.8-25.3%
A protest against international sanctions and global hypocrisy	8	12.9%	6.7-23.4%

S6. Full sub-group contrasts

All sub-group contrasts reported in the main chapter plus several additional contrasts not reported there. For 2x2 tables the Fisher exact test is used; for larger tables the Pearson chi-square test is reported, with minimum expected cell count reported to warn where chi-square assumptions are stressed. Cramér's V is reported with the Bergsma (Bergsma, 2013) bias correction. No contrast below reaches $\alpha = .05$.

Table S16. All sub-group contrasts tested (nonsignificant at $\alpha = .05$).

Contrast	Statistic	p	Cramer's V	Min expected
Reconstruct x Ethnicity	chi2(4) =5.42	0.247	0.12	0.71
Reconstruct x Gender	chi2(2) =0.36	0.835	0.00	1.48
Reconstruct x Location	chi2(2) =2.25	0.325	0.06	1.64
Reconstruct x Age band	chi2(6) =3.20	0.784	0.00	0.08
Sensory x Ethnicity	chi2(4) =6.51	0.164	0.16	1.18
Sensory x Gender	chi2(2) =2.65	0.265	0.10	2.07
Sensory x Location	chi2(2) =1.72	0.424	0.00	2.82
Hazara-voice x Ethnicity	chi2(4) =4.34	0.361	0.05	2.59
Hazara-voice x Gender	chi2(2) =1.69	0.429	0.00	3.54

Contrast	Statistic	p	Cramer's V	Min expected
Hazara-voice x Age band	chi2(6) =10.16	0.118	0.18	0.19
Heal/expose x Ethnicity	chi2(6) =4.06	0.668	0.00	0.47
Heal/expose x Gender	chi2(3) =2.22	0.529	0.00	0.59
Intl cared more x Ethnicity	chi2(4) =1.18	0.881	0.00	3.43
Intl cared more x Gender	chi2(2) =4.49	0.106	0.21	4.61

Fisher exact — Hazara-voice perception (Yes vs No/Don't know) by Location (Inside/Diaspora). Inside Afghanistan: Yes = 11, No/DK = 14. Diaspora: Yes = 8, No/DK = 28. Odds ratio = 2.85, $p = 0.092$. This is the single contrast that approaches but does not reach conventional significance; it is noted for transparency rather than interpreted substantively and the direction (respondents inside Afghanistan somewhat more likely to report that Hazara voices are included) would require an interview-based follow-up to investigate.

S7. Sample vs population representativeness

The main chapter reports that Hazara respondents are over-represented in the sample relative to rough Afghanistan-wide population estimates. Those estimates are indicative, not definitive — the most recent authoritative Afghanistan-wide census was conducted decades ago and contemporary ethnic-share estimates vary across sources. The goodness-of-fit test below is offered to quantify the over-representation, not to certify the estimates themselves.

Table S17. Chi-square goodness-of-fit, named-ethnicity portion of the sample (n = 51) against rough Afghanistan-wide ethnic-share estimates.

Ethnicity	Observed (sample)	Expected (normalised to n = 51)
Pashtun	26	27.5
Tajik	12	17.7
Hazara	13	5.9

Chi-square = 10.49, $df = 2$, $p = 0.0053$. Interpretation: the sample over-represents Hazara respondents relative to rough population shares. This is consistent with the sampling strategy (purposive and snowball recruitment, with the Bamiyan case salient for Hazara communities) and is the basis for the main chapter's explicit refusal to claim population-level representativeness.

S8. Hazara sub-sample breakdown (n = 13)

The Hazara respondents ($n = 13$) are not a statistically sufficient sub-sample to sustain claims about Hazara opinion in aggregate. It shows internal disagreement within the Hazara sub-sample that is itself informative — particularly on the question of whether Hazara voices are included in reconstruction discussions.

Table S18. Hazara sub-sample — demographic and core-item breakdown (n = 13).

Variable	Breakdown
Location (Inside/Diaspora)	Diaspora: 7; Inside: 6
Gender	Male: 9; Female: 4
Age band	21-35: 11; 36-50: 2
Should reconstruct?	Yes: 10; No: 2; Not sure: 1
Hazara voices included?	I don't know: 5; No: 4; Yes: 4
Sensory inclusion?	Yes: 9; Not sure: 4
Heal / Expose / Both / Neither	Both: 9; Heal the past: 2; Neither: 1; Expose its wounds: 1

Within the Hazara sub-sample, views on whether Hazara voices are currently included in reconstruction discussions divide evenly (Yes / No / Don't know). Substantive claims about Hazara opinion in aggregate are not made in the main chapter, for the reason implicit in this sub-sample size.

S9. Thematic coding scheme

Open-ended responses were coded following Braun and Clarke (Braun & Clarke, 2006) under a critical-realist framing. The first author conducted an initial inductive coding pass, generating candidate themes without reference to demographic variables. A second pass cross-referenced the themes against the closed-item findings and revised the theme definitions. A subset of transcripts (n = 20, selected by stratified random sampling across respondents) was independently coded by a co-author blinded to the initial coding. Agreement on the seven themes was substantial (Cohen's κ = 0.74); disagreements were resolved by discussion. Themes are not mutually exclusive: a single open-ended response can contribute to more than one theme.

Table S19. Thematic coding scheme and approximate frequencies across 62 respondents. Frequencies are approximate because themes are not mutually exclusive.

Theme	Description	Approx. frequency
T1. Reconstruction as identity work rather than devotion	Respondents frame the statues as markers of Afghan civilisational identity; cultural > spiritual register.	n ≈ 41
T2. Destruction as political project	The 2001 act is interpreted as politically motivated, targeting ethnic plurality; religious justification disputed.	n ≈ 37
T3. Pragmatic reconstruction with livelihood dimensions	Reconstruction framed as a basis for tourism, employment and economic recovery at Bamiyan.	n ≈ 34
T4. Resistance to the technical/community dichotomy	Material fidelity and intangible/meaning dimensions treated as complementary, not as alternatives.	n ≈ 28
T5. Scepticism about AI mediation	Principal concern: oversimplification of cultural meaning; conditional openness given transparency and community consultation.	n ≈ 36
T6. Representational absence rather than presence	Preference for reconstructions that preserve rather than hide the rupture.	n ≈ 22

Theme	Description	Approx. frequency
T7. Governance and the inclusion question	Identification of missing voices (Hazara, women, local residents); endorsement of distributed leadership.	n ≈ 24

S10. Files and reproduction recipe

The following files are released alongside the main chapter and this supplement:

File	Description
dataset_anonymised.csv	Anonymised responses (N = 62, 46 columns)
codebook.txt	Column key for the anonymised dataset
analysis.py	Python script reproducing every number in Tables S4-S8
analysis_output.txt	Output of analysis.py on the anonymised dataset

Reproduction recipe. With Python 3.12 or newer and the packages pandas, numpy and scipy installed, place the raw XLSX export in the same directory as analysis.py and run:

```
python3 analysis.py
```

The script prints a log to the console and writes the same log to analysis_output.txt. The numbers reported in Tables S4, S5, S6, S7 and S8 of this supplement are emitted verbatim by this script. The script is also a sufficient specification of the analytical pipeline: any user can verify that no hidden decisions lie between the raw responses and the reported statistics.

Version controls. The anonymised dataset and the analysis script are archived alongside this supplement. If downstream analyses amend either, the amendment should be versioned (e.g. dataset_anonymised_v2.csv, analysis_v2.py) so that the correspondence between chapter claims and reproducibility artefacts remains traceable.

Reference

- Abdulkarim, M., & Seigne, J. (2024). The Future of the Temple of Bel in Palmyra after Its Destruction. *Bulletin of the American Society of Overseas Research*, 391, 93–106. <https://doi.org/10.1086/729835>
- Acabado, S., & Martin, M. (2020). Decolonizing the past, empowering the future: Community-led heritage conservation in Ifugao, Philippines. *Journal of Community Archaeology & Heritage*, 7(3), 171–186. <https://doi.org/10.1080/20518196.2020.1767383>
- Adetayo, A. J. (2024). Reimagining Learning Through AI Art: The Promise of DALL-E and MidJourney for Education and Libraries. *Library Hi Tech News*. <https://doi.org/10.1108/lhtn-01-2024-0005>
- Agapiou, A. (2021). UNESCO World Heritage properties in changing and dynamic environments: change detection methods using optical and radar satellite data. *Heritage Science*, 9(1), 64. <https://doi.org/10.1186/s40494-021-00542-z>
- Ajuzieogu, U. (2021). *Cultural Heritage Reconstruction and Preservation Through Generative AI*. <https://doi.org/10.13140/RG.2.2.24236.17281>
- Angouri, J., Paraskevaidi, M., & Wodak, R. (2017). Discourses of cultural heritage in times of crisis: The case of the Parthenon Marbles. *Journal of Sociolinguistics*, 21(2), 208–237. <https://doi.org/10.1111/josl.12232>
- Apollonio, F. I., Basilissi, V., Callieri, M., Dellepiane, M., Gaiani, M., Ponchio, F., Rizzo, F., Rubino, A. R., Scopigno, R., & Sobra', G. (2018). A 3D-centered information system for the documentation of a complex restoration intervention. *Journal of Cultural Heritage*, 29, 89–99. <https://doi.org/10.1016/j.culher.2017.07.010>
- Arzomand, K., & Kalganova, T. (2026a). Community-Grounded Authenticity at Bamiyan: Afghan Perspectives on AI-mediated Heritage Reconstruction [Preprint]. *Zenodo*. <https://doi.org/10.5281/zenodo.20435358>
- Arzomand, K., & Kalganova, T. (2026b). Evaluating Generative AI Reconstructions of the Bamiyan Buddhas: Computational Similarity and Expert Archaeological Assessment [Preprint]. *Zenodo*. <https://doi.org/10.5281/zenodo.20435265>
- Arzomand, K., & Kalganova, T. (2026c). Generalising HARF Beyond Bamiyan: A Bounded Cross-Site Evaluation of a Generative-AI Heritage Reconstruction Framework at the Temple of Bel, Palmyra. *Heritage*, 9(8), 331. <https://doi.org/10.3390/heritage9080331>
- Arzomand, K., Kalganova, T., & Rustell, M. (2026). HARF: A human–AI collaborative framework for cultural heritage reconstruction with expert-guided multi-platform generative AI and systematic prompt engineering. *Digital Applications in Archaeology and Cultural Heritage*, 42, e00554. <https://doi.org/10.1016/j.daach.2026.e00554>

- Arzomand, K., Rustell, M., & Kalganova, T. (2024). From Ruins to Reconstruction: Harnessing Text-to-Image AI for Restoring Historical Architectures. *Challenge Journal of Structural Mechanics*, 10(2), 69. <https://doi.org/10.20528/cjsmec.2024.02.004>
- Atalay, S. (2012). *Research with, by, and for Indigenous and Local Communities* (1st ed.). University of California Press. <http://www.jstor.org/stable/10.1525/j.ctt1pn9dj>
- Aubry, M., Berggren, W. A., Dupuis, C., Ghaly, H., Ward, D., King, C., Knox, R. W. O., Ouda, K., Youssef, M., & Galal, W. F. (2009). Pharaonic necrostratigraphy: a review of geological and archaeological studies in the Theban Necropolis, Luxor, West Bank, Egypt. *Terra Nova*, 21(4), 237–256. <https://doi.org/10.1111/j.1365-3121.2009.00872.x>
- Avrahami, O., & others. (2023). Blended latent diffusion. *ACM SIGGRAPH 2023*.
- Bao, H., & Bowen, J. P. (2025). From Material Conservation to Digital Presence: Reconstructing Visitors' Heritage Experience and Meaning-Making Through Digital Dunhuang. *Heritage*, 8(12), 534. <https://doi.org/10.3390/heritage8120534>
- Bassier, M., Yousefzadeh, M., & Vergauwen, M. (2020). Comparison of 2D and 3D wall reconstruction algorithms from point cloud data for as-built BIM. *Journal of Information Technology in Construction*, 25, 173–192. <https://doi.org/10.36680/j.itcon.2020.011>
- Bau, D., & others. (2023). Rewriting a deep generative model. *ICCV 2023*.
- Becker, C., & Laycock, R. (2023). Embracing Deepfakes and AI-generated Images in Neuroscience Research. *European Journal of Neuroscience*, 58(3), 2657–2661. <https://doi.org/10.1111/ejn.16052>
- Bender, A., Guerreiro, M., Agapito, D., Dias Sequeira, B., & Mendes, J. (2023). Sensory experiences in heritage contexts: A qualitative approach. *European Journal of Tourism Research*, 36, 3604. <https://doi.org/10.54055/ejtr.v36i.3060>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. <https://doi.org/10.1145/3442188.3445922>
- Bennoui-Ladraa, B., & Chennaoui, Y. (2018). Use of Photogrammetry for Digital Surveying, Documentation and Communication of the Cultural Heritage. Example Regarding Virtual Reconstruction of the Access Doors for the Nameless Temple of Tipasa (Algeria). *Studies in Digital Heritage*, 2(2), 121–137. <https://doi.org/10.14434/sdh.v2i2.24496>
- Bentkowska-Kafel, A., & Denard, H. (2012). *Paradata and Transparency in Virtual Heritage* (A. Bentkowska-Kafel & Denard Hugh, Eds.; 1st ed.). Ashgate. https://www.researchgate.net/publication/287270234_Paradata_and_transparency_in_virtual_heritage
- Bergsma, W. (2013). A bias-correction for Cramér's and Tschuprow's. *Journal of the Korean Statistical Society*, 42(3), 323–328. <https://doi.org/10.1016/j.jkss.2012.10.002>
- Berto, S., Demetrescu, E., Fanini, B., Bonetto, J., & Salemi, G. (2021). Analysis and Validation of the 3D Reconstructive Process through the Extended Matrix Framework of the Temple of the Roman Forum of Nora (Sardinia, CA). *ArcheoFOSS XIII Workshop—Open Software, Hardware,*

- Processes, Data and Formats in Archaeological Research*, 18.
<https://doi.org/10.3390/environsciproc2021010018>
- Bevan, A., & Lake, M. (2016). *Computational Approaches to Archaeological Spaces* (A. Bevan & M. Lake, Eds.). Routledge. <https://doi.org/10.4324/9781315431932>
- Bevilacqua, M. G., Caroti, G., Piemonte, A., & Ulivieri, D. (2019). RECONSTRUCTION OF LOST ARCHITECTURAL VOLUMES BY INTEGRATION OF PHOTOGRAMMETRY FROM ARCHIVE IMAGERY WITH 3-D MODELS OF THE STATUS QUO. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W9, 119–125.
<https://doi.org/10.5194/isprs-archives-XLII-2-W9-119-2019>
- Bharati, P. (2023). Leveraging AI for Archaeological Insights: Generating Visual Representations of Historical Events - A Case Study of the Hastinapur Flood during the Reign of King Nichakshu. *International Journal for Research in Applied Science and Engineering Technology*, 11(10), 142–145. <https://doi.org/10.22214/ijraset.2023.55951>
- Biljecki, F., Stoter, J., Ledoux, H., Zlatanova, S., & Çöltekin, A. (2015). Applications of 3D City Models: State of the Art Review. *ISPRS International Journal of Geo-Information*, 4(4), 2842–2889.
<https://doi.org/10.3390/ijgi4042842>
- Birhane, A., Prabhu, V., & Kahembwe, E. (2023). The values encoded in machine vision. *Patterns*, 4(2), 100779. <https://doi.org/10.1016/j.patter.2023.100779>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp063oa>
- Brisco, R., Hay, L., & Dhami, S. (2023). EXPLORING THE ROLE OF TEXT-TO-IMAGE AI IN CONCEPT GENERATION. *Proceedings of the Design Society*, 3, 1835–1844.
<https://doi.org/10.1017/pds.2023.184>
- British Museum, & Zwalf, W. (1996). A Catalogue of the Gandhāra Sculpture in the British Museum. In *Journal of the American Oriental Society* (Vol. 119, Number 2). For the Trustees of British Museum by British Museum Press. <https://catalog.hathitrust.org/Record/004005790>
- Brooks, T., & others. (2023). InstructPix2Pix: Learning to Edit Images from Instructions. *CVPR 2023*.
- Brosché, J., Legnér, M., Kreutz, J., & Ijla, A. (2017). Heritage under attack: motives for targeting cultural property during armed conflict. *International Journal of Heritage Studies*, 23(3), 248–260. <https://doi.org/10.1080/13527258.2016.1261918>
- Cao, Y., Bowker, M. A., Delgado-Baquerizo, M., & Xiao, B. (2023). Biocrusts protect the Great Wall of China from erosion. *Science Advances*, 9(49). <https://doi.org/10.1126/sciadv.adk5892>
- Carter, A. K., Stark, M. T., Quintus, S., Zhuang, Y., Wang, H., Heng, P., & Chhay, R. (2019). Temple occupation and the tempo of collapse at Angkor Wat, Cambodia. *Proceedings of the National Academy of Sciences*, 116(25), 12226–12231. <https://doi.org/10.1073/pnas.1821879116>
- Champion, E., Rahaman, H., Bekele, M., & Mcmeekin, D. (2021). *Virtual Heritage: A Guide*.
<https://doi.org/10.5334/bck>
- Chang, K. K., Cramer, M., Soni, S., & Bamman, D. (2023). *Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4*.

- Chase, A. F., Chase, D. Z., Weishampel, J. F., Drake, J. B., Shrestha, R. L., Slatton, K. C., Awe, J. J., & Carter, W. E. (2011). Airborne LiDAR, archaeology, and the ancient Maya landscape at Caracol, Belize. *Journal of Archaeological Science*, 38(2), 387–398. <https://doi.org/10.1016/j.jas.2010.09.018>
- Chen, C. (2021). Angkor Wat: A transcultural history of heritages. *Journal of Southeast Asian Studies*, 52(1), 133–140. <https://doi.org/10.1017/S0022463421000230>
- Chen, C., Leask, A., & Phou, S. (2016). Symbolic, Experiential and Functional Consumptions of Heritage Tourism Destinations: The Case of Angkor World Heritage Site, Cambodia. *International Journal of Tourism Research*, 18(6), 602–611. <https://doi.org/10.1002/jtr.2077>
- Clini, P., Nespeca, R., Angeloni, R., & Coppetta, L. (2024). 3D representation of Architectural Heritage: a comparative analysis of NeRF, Gaussian Splatting, and SfM-MVS reconstructions using low-cost sensors. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-2/W8-2024, 93–99. <https://doi.org/10.5194/isprs-archives-XLVIII-2-W8-2024-93-2024>
- Cobb, P. J. (2023). Large Language Models and Generative AI, Oh My! *Advances in Archaeological Practice*, 11(3), 363–369. <https://doi.org/10.1017/aap.2023.20>
- Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1), 37–46. <https://doi.org/10.1177/001316446002000104>
- Colace, F., Gaeta, R., Lorusso, A., Pellegrino, M., & Santaniello, D. (2025). New AI challenges for cultural heritage protection: A general overview. *Journal of Cultural Heritage*, 75, 168–193. <https://doi.org/10.1016/j.culher.2025.07.019>
- Connerton, P. (2008). Seven types of forgetting. *Memory Studies*, 1(1), 59–71. <https://doi.org/10.1177/1750698007083889>
- Conte, C., & others. (2023). AI in cultural heritage: Bias and evaluation. *Heritage Science*, 11(1), 103.
- Costantino, D., Pepe, M., & Restuccia, A. G. (2023). Scan-to-HBIM for conservation and preservation of Cultural Heritage building: the case study of San Nicola in Montedoro church (Italy). *Applied Geomatics*, 15(3), 607–621. <https://doi.org/10.1007/s12518-021-00359-2>
- Cui, H., & Wu, J. (2025). How Architectural Heritage Is Moving to Smart: A Systematic Review of HBIM. *Buildings*, 15(15), 2664. <https://doi.org/10.3390/buildings15152664>
- Cunliffe, E., Muhesen, N., & Lostal, M. (2016). The Destruction of Cultural Property in the Syrian Conflict: Legal Implications and Obligations. *International Journal of Cultural Property*, 23(1), 1–31. <https://doi.org/10.1017/S0940739116000011>
- Cusack-McVeigh, H. (2014). Community-Based Archaeology: Research with, by, and for Indigenous and Local Communities by Sonya Atalay (review). *Great Plains Research*, 24(2), 214–215. <https://doi.org/10.1353/gpr.2014.0043>
- Dalal, N., & Triggs, B. (2005). Histograms of Oriented Gradients for Human Detection. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, 886–893. <https://doi.org/10.1109/CVPR.2005.177>

- Debevec, P. (2004). The Parthenon. *ACM SIGGRAPH 2004 Computer Animation Festival on - SIGGRAPH '04*, 188. <https://doi.org/10.1145/1186015.1186041>
- Delgado-Martos, E., Montas-Laracuenta, N., Pesqueira-Calvo, C., Intra Sidola, G., Maitín, A., Nogales, A., & García-Tejedor, Á. J. (2026). ARCHER: An integrated approach to artificial intelligence in the virtual reconstruction of architectural heritage. *Digital Applications in Archaeology and Cultural Heritage*, 41, e00542. <https://doi.org/10.1016/j.daach.2026.e00542>
- Denard, H. (2009). *The London Charter for the Computer-Based Visualisation of Cultural Heritage*. King's College London. <https://www.londoncharter.org/>
- Denker, A. (2017). 3D VISUALIZATION AND PHOTO-REALISTIC RECONSTRUCTION OF THE GREAT TEMPLE OF BEL. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W3, 225–229. <https://doi.org/10.5194/isprs-archives-XLII-2-W3-225-2017>
- Dillman, D. A., Smyth, J. D., & Christian, L. M. (2014). *Internet, Phone, Mail, and Mixed-Mode Surveys*. Wiley. <https://doi.org/10.1002/9781394260645>
- Don-Yehiya, S., Choshen, L., & Abend, O. (2023). *Human Learning by Model Feedback: The Dynamics of Iterative Prompting with Midjourney*.
- Dupree, N. H. (2002). Cultural heritage and national identity in Afghanistan. *Third World Quarterly*, 23(5), 977–989. <https://doi.org/10.1080/0143659022000028549>
- Elcheikh, Z. (2019). Palmyra: a story of ruins, struggle(s) and beyond. *Chronos*, 39, 105–123. <https://doi.org/10.31377/chr.v39i0.597>
- Evans, D. H., Fletcher, R. J., Pottier, C., Chevance, J.-B., Soutif, D., Tan, B. S., Im, S., Ea, D., Tin, T., Kim, S., Cromarty, C., De Greef, S., Hanus, K., Bâty, P., Kuszinger, R., Shimoda, I., & Boornazian, G. (2013). Uncovering archaeological landscapes at Angkor using lidar. *Proceedings of the National Academy of Sciences*, 110(31), 12595–12600. <https://doi.org/10.1073/pnas.1306539110>
- Fernandes Dionizio, R., & Dezen-Kempter, E. (2025). From Data and Metadata to HBIM-GIS Integration. *International Journal of Architectural Heritage*, 19(10), 2443–2456. <https://doi.org/10.1080/15583058.2024.2439916>
- Fincham, D. (2012). The Parthenon Sculptures and Cultural Justice. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.2131825>
- Fischler, M. A., & Bolles, R. C. (1981). Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM*, 24(6), 381–395. <https://doi.org/10.1145/358669.358692>
- Flood, F. B. (2002). Between Cult and Culture: Bamiyan, Islamic Iconoclasm, and the Museum. *The Art Bulletin*, 84(4), 641. <https://doi.org/10.2307/3177288>
- Foka, A., & Griffin, G. (2024a). AI, Cultural Heritage, and Bias: Some Key Queries That Arise from the Use of GenAI. *Heritage*, 7(11), 6125–6136. <https://doi.org/10.3390/heritage7110287>
- Foka, A., & Griffin, G. (2024b). Ethics of algorithmic heritage. *AI & Society*. <https://doi.org/10.1007/s00146-024-01742-8>

- Foka, A., Griffin, G., Ortiz Pablo, D., Rajkowska, P., & Badri, S. (2025). Tracing the bias loop: AI, cultural heritage and bias-mitigating in practice. *AI & SOCIETY*, 40(8), 5835–5847. <https://doi.org/10.1007/s00146-025-02349-z>
- Forte, M. (2024). Perceiving Etruscan Art: AI and Visual Perception. *Humans*, 4(4), 409–429. <https://doi.org/10.3390/humans4040027>
- Fulford, M., & Wallace-Hadrill, A. (1999). Towards a history of pre-Roman Pompeii: excavations beneath the House of Amarantus (I.9.11–12), 1995–8. *Papers of the British School at Rome*, 67, 37–144. <https://doi.org/10.1017/S0068246200004529>
- Gal, R., & others. (2023). Textual inversion: Generating personalized images. *CVPR 2023*.
- Gemini Team, Georgiev, P., Lei, V. I., Burnell, R., & et al. (2024). *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context*. <https://deepmind.google/models/gemini/>
- George, P. (2022). AI trends in digital humanities research. *Trends in Computer Science and Information Technology*, 7(2), 026–034. <https://doi.org/10.17352/tcsit.000048>
- Ghaboura, S., More, K., Thawkar, R., Alghallabi, W., Thawakar, O., Khan, F. S., Cholakkal, H., Khan, S., & Anwer, R. M. (2025). *Time Travel: A Comprehensive Benchmark to Evaluate LMMs on Historical and Cultural Artifacts*.
- Gröger, G., & Plümer, L. (2012). CityGML – Interoperable semantic 3D city models. *ISPRS Journal of Photogrammetry and Remote Sensing*, 71, 12–33. <https://doi.org/10.1016/j.isprsjprs.2012.04.004>
- Grün, A., Remondino, F., & Zhang, L. (2004). Photogrammetric Reconstruction of the Great Buddha of Bamiyan, Afghanistan. *The Photogrammetric Record*, 19(107), 177–199. <https://doi.org/10.1111/j.0031-868X.2004.00278.x>
- Gualandi, M. L., Gattiglia, G., & Anichini, F. (2021). An Open System for Collection and Automatic Recognition of Pottery through Neural Network Algorithms. *Heritage*, 4(1), 140–159. <https://doi.org/10.3390/heritage4010008>
- Gwet, K. (2012). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*.
- Gwet, K. L. (2008). Computing Inter-rater Reliability and its Variance in the Presence of High Agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1), 29–48. <https://doi.org/10.1348/000711006X126600>
- Hamilakis, Y. (2002). The Other “Parthenon”: Antiquity and National Memory at Makronisos. *Journal of Modern Greek Studies*, 20(2), 307–338. <https://doi.org/10.1353/mgs.2002.0025>
- Hammer, E., Seifried, R., Franklin, K., & Lauricella, A. (2018). Remote assessments of the archaeological heritage situation in Afghanistan. *Journal of Cultural Heritage*, 33, 125–144. <https://doi.org/10.1016/j.culher.2017.12.008>
- Harrison, R. (2012). *Heritage: Critical Approaches*. Routledge. <https://doi.org/10.4324/9780203108857>

- He, Y., Bai, Y., Lin, M., Zhao, W., Hu, Y., Sheng, J., Yi, R., Li, J., & Liu, Y.-J. (2024). *T³Bench: Benchmarking Current Progress in Text-to-3D Generation*.
- He, Z., Su, J., Chen, L., Wang, T., & LC, R. (2024). *"I Recall the Past": Exploring How People Collaborate with Generative AI to Create Cultural Heritage Narratives*.
- Hegazy, M., & Saleh, A. (2023). Evolution of AI role in architectural design: between parametric exploration and machine hallucination. *MSA Engineering Journal*, 2(2), 262–288. <https://doi.org/10.21608/msaeng.2023.291873>
- Heikkinen, J. (2009). Close-range constrained image sequences. *ISPRS Journal of Photogrammetry and Remote Sensing*, 64(3), 267–274. <https://doi.org/10.1016/j.isprsjprs.2008.10.001>
- Holtorf, C. (2020). Destruction and Reconstruction of Cultural Heritage as Future-Making. In *The Future of the Bamiyan Buddha Statues* (pp. 157–172). Springer International Publishing. https://doi.org/10.1007/978-3-030-51316-0_10
- Hong, Y., Zhang, K., Gu, J., Bi, S., Zhou, Y., Liu, D., Liu, F., Sunkavalli, K., Bui, T., & Tan, H. (2024). *LRM: Large Reconstruction Model for Single Image to 3D*.
- Horn, C., Ivarsson, O., Lindhé, C., Potter, R., Green, A., & Ling, J. (2022). Artificial Intelligence, 3D Documentation, and Rock Art—Approaching and Reflecting on the Automation of Identification and Classification of Rock Art Images. *Journal of Archaeological Method and Theory*, 29(1), 188–213. <https://doi.org/10.1007/s10816-021-09518-6>
- Hosseini, S., Barker, K., & Ramirez-Marquez, J. E. (2016). A review of definitions and measures of system resilience. *Reliability Engineering & System Safety*, 145, 47–61. <https://doi.org/10.1016/j.ress.2015.08.006>
- Hou, J., & others. (2023). Cross-model evaluation of diffusion systems. *CVPR Workshops*.
- Hsu, Y.-C., Yang, Z., & Buehler, M. J. (2022). Generative design, manufacturing, and molecular modeling of 3D architected materials based on natural language input. *APL Materials*, 10(4). <https://doi.org/10.1063/5.0082338>
- Hu, G., Mortazavian, P., Kittler, J., & Christmas, W. (2013). A facial symmetry prior for improved illumination fitting of 3D morphable model. *2013 International Conference on Biometrics (ICB)*, 1–6. <https://doi.org/10.1109/ICB.2013.6613000>
- Huang, C.-T., Geng, T., & Liu, J. (2023). *Capturing the Characteristics of Mis/Disinformation Propagation Over the Internet*. <https://doi.org/10.1117/12.2664036>
- Huvila, I. (2022a). Improving the usefulness of research data with better paradata. *Open Information Science*, 6(1), 28–48. <https://doi.org/10.1515/opis-2022-0129>
- Huvila, I. (2022b). Transparency and ethics in information work. *Journal of Documentation*, 78(7), 1492–1507.
- ICOMOS. (1964). *International Charter for the Conservation and Restoration of Monuments and Sites (The Venice Charter 1964)* (li. I. C. of A. and T. of H. Monuments, Ed.). ICOMOS. https://www.icomos.org/images/DOCUMENTS/Charters/venice_e.pdf

- ICOMOS. (1994). *The Nara Document on Authenticity*. ICOMOS.
https://www.iccom.org/sites/default/files/publications/2020-05/convern8_06_thenaradocu_ing.pdf
- ICOMOS. (2003). *Evaluation of the Cultural Landscape and Archaeological Remains of the Bamiyan Valley (Afghanistan)*. http://whc.unesco.org/archive/advisory_body_evaluation/208rev.pdf
- Jokilehto, J. I. (2020). Reflections on the Case of Bamiyan. In *The Future of the Bamiyan Buddha Statues* (pp. 205–214). Springer International Publishing. https://doi.org/10.1007/978-3-030-51316-0_13
- Jones, S. (2010). Negotiating Authentic Objects and Authentic Selves. *Journal of Material Culture*, 15(2), 181–203. <https://doi.org/10.1177/1359183510364074>
- Kadhim, I., & Abed, F. M. (2023). A Critical Review of Remote Sensing Approaches and Deep Learning Techniques in Archaeology. *Sensors*, 23(6), 2918. <https://doi.org/10.3390/s23062918>
- Kenig, N., Monton Echeverria, J., & Muntaner Vives, A. (2023). Human Beauty according to Artificial Intelligence. *Plastic and Reconstructive Surgery - Global Open*, 11(7), e5153.
<https://doi.org/10.1097/GOX.00000000000005153>
- Kerbl, B., Kopanas, G., Leimkuehler, T., & Drettakis, G. (2023). 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4), 1–14.
<https://doi.org/10.1145/3592433>
- Keyvanfar, A., Shafaghat, A., & Rosley, M. S. (2022). Performance comparison analysis of 3D reconstruction modeling software in construction site visualization and mapping. *International Journal of Architectural Computing*, 20(2), 453–475.
<https://doi.org/10.1177/14780771211066876>
- Kila, J. D. (2019). Iconoclasm and Cultural Heritage Destruction During Contemporary Armed Conflicts. In *The Palgrave Handbook on Art Crime* (pp. 653–683). Palgrave Macmillan UK.
https://doi.org/10.1057/978-1-137-54405-6_30
- Kirshenblatt-gimblett, B. (2004). Intangible Heritage as Metacultural Production1. *Museum International*, 56(1–2), 52–65. <https://doi.org/10.1111/j.1350-0775.2004.00458.x>
- Klimburg-Salter, D. (1989). *The Kingdom of Bamiyan : Buddhist Art and Culture of the Hindu Kush*. Istituto Universitario Orientale / Istituto Italiano per il Medio ed Estremo Oriente.
<https://books.google.co.uk/books?id=XRZuOQAACAAJ>
- Knell, S. (2022). **Experimental museology: institutions, representations, users** ., *Museum Management and Curatorship*, 37(3), 330–332.
<https://doi.org/10.1080/09647775.2022.2068287>
- Ko, J., Ajibefun, J., & Yan, W. (2023). *Experiments on Generative AI-Powered Parametric Modeling and BIM for Architectural Design*.
- Krippendorff, K. (2004). *Content Analysis: An Introduction to Its Methodology*. Sage.
<https://books.google.co.uk/books?id=q657o3M3C8cC>
- Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1), 159. <https://doi.org/10.2307/2529310>

- Lee, H.-H., & Chang, A. X. (2022). Understanding Pure CLIP Guidance for Voxel Grid NeRF Models. *Arxiv*.
- Li, X., Lin, J., & Zhang, X. (2025). Dynamic Transmission and Innovative Transformation of Cultural Heritage: Generative Artificial Intelligence Practices Based on Cultural Cognitive Models. *Applied Sciences*, 15(23), 12651. <https://doi.org/10.3390/app152312651>
- Lin, C.-H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.-Y., & Lin, T.-Y. (2022). Magic3D: High-Resolution Text-to-3D Content Creation. In *ArXiv*. <https://doi.org/10.48550/ARXIV.2211.10440>
- Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., & Vondrick, C. (2023). Zero-1-to-3: Zero-shot One Image to 3D Object. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9264–9275. <https://doi.org/10.1109/ICCV51070.2023.00853>
- Liu, T., Mathrani, A., & Mbachu, J. (2019). Benefits and barriers in uptake of mobile apps in New Zealand construction industry. *Facilities*, 37(5/6), 254–265. <https://doi.org/10.1108/F-08-2017-0078>
- Logan, W., & Reeves, K. (2008). *Places of Pain and Shame* (W. Logan & K. Reeves, Eds.). Routledge. <https://doi.org/10.4324/9780203885031>
- López-Menchero Bendicho, V. M., & Grande, A. (2011). *The Seville Principles: International Principles of Virtual Archaeology (2011)*. International Forum of Virtual Archaeology (IFVA). <https://smartheritage.com/seville-principles/>
- López-Menchero Bendicho, V. M., & Grande, A. (2017). *The Seville Principles (2017 Update): International Principles of Virtual Archaeology*. International Forum of Virtual Archaeology (IFVA). <https://smartheritage.com/seville-principles/>
- Lostal, M., & Vasconcelos Vilaça, G. (2015). The Bamiyazation of Cultural Heritage and the Silk Road Economic Belt: Challenges and Opportunities for China. *The Chinese Journal of Comparative Law*, cxv013. <https://doi.org/10.1093/cjcl/cxv013>
- Lowenthal, D. (1998). *The Heritage Crusade and the Spoils of History*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511523809>
- Lugmayr, A., & others. (2022). RePaint: Inpainting using diffusion models. *CVPR 2022*.
- Lyu, Y., Wang, X., Lin, R., & Wu, J. (2022). Communication in Human–AI Co-Creation: Perceptual Analysis of Paintings Generated by Text-to-Image System. *Applied Sciences*, 12(22), 11312. <https://doi.org/10.3390/app122211312>
- Macdonald, S. (2008). Difficult heritage: Negotiating the Nazi past in Nuremberg and beyond. In *Difficult Heritage: Negotiating the Nazi Past in Nuremberg and Beyond*. <https://doi.org/10.4324/9780203888667>
- Mahmoud, H., Alfons, R., & M. Reffat, R. (2019). Analysis of The Driving Forces of Urban Expansion in Luxor City by Remote Sensing Monitoring. *International Journal of Integrated Engineering*, 11(6). <https://doi.org/10.30880/ijie.2019.11.06.031>
- Maiwald, F., Vietze, T., Schneider, D., Henze, F., Münster, S., & Niebling, F. (2017). PHOTOGRAMMETRIC ANALYSIS OF HISTORICAL IMAGE REPOSITORIES FOR VIRTUAL

- RECONSTRUCTION IN THE FIELD OF DIGITAL HUMANITIES. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-2/W3, 447–452. <https://doi.org/10.5194/isprs-archives-XLII-2-W3-447-2017>
- Manning, J. (2012). Thebes (Diospolis Magna), Ptolemaic and Roman periods. In *The Encyclopedia of Ancient History*. Wiley. <https://doi.org/10.1002/9781444338386.wbeah07107>
- Marwick, B. (2017). Computational Reproducibility in Archaeological Research: Basic Principles and a Case Study of Their Implementation. *Journal of Archaeological Method and Theory*, 24(2), 424–450. <https://doi.org/10.1007/s10816-015-9272-9>
- Marwick, B. (2022). Open science and reproducibility in archaeology. *Advances in Archaeological Practice*, 10(1), 1–13.
- Mazzaglia, A. (2021). The Information System of Pompeii Sustainable Preservation Project. A Tool for the Collection, Management and Sharing of Knowledge Useful for Conservation and Renovation of Archaeological Monuments. *ArcheoFOSS XIII Workshop—Open Software, Hardware, Processes, Data and Formats in Archaeological Research*, 14. <https://doi.org/10.3390/environsciproc2021010014>
- Meister, J., Garbe, P., Trappe, J., Ullmann, T., Es-Senussi, A., Baumhauer, R., Lange-Athinodorou, E., & El-Raouf, A. A. (2021). The Sacred Waterscape of the Temple of Bastet at Ancient Bubastis, Nile Delta (Egypt). *Geosciences*, 11(9), 385. <https://doi.org/10.3390/geosciences11090385>
- Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model Cards for Model Reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229. <https://doi.org/10.1145/3287560.3287596>
- Monsutti, A. (2005). *War and Migration: Social Networks and Economic Strategies of the Hazaras of Afghanistan*. Routledge. <https://doi.org/10.4324/9780203959107>
- Morgan, C., & Eve, S. (2012). DIY and digital archaeology: what are you doing to participate? *World Archaeology*, 44(4), 521–537. <https://doi.org/10.1080/00438243.2012.741810>
- Mousavi, S. A., & Canfield, R. L. (1999). The Hazaras of Afghanistan: An Historical, Cultural, Economic and Political Study (Surrey: Curzon, 1998). Pp. 282. *International Journal of Middle East Studies*, 31(2), 321–322. <https://doi.org/10.1017/S0020743800054325>
- Muldoon, J., & Wu, B. A. (2023). Artificial Intelligence in the Colonial Matrix of Power. *Philosophy & Technology*, 36(4), 80. <https://doi.org/10.1007/s13347-023-00687-8>
- Munawar, N. A. (2017). Reconstructing Cultural Heritage in Conflict Zones: Should Palmyra be Rebuilt? *Ex Novo: Journal of Archaeology*, 2, 33–48. <https://doi.org/10.32028/exnovo.v2i0.388>
- Münster, S., & others. (2024). Digital workflows for cultural heritage reconstruction. *Journal on Computing and Cultural Heritage*.
- Murphy, M., McGovern, E., & Pavia, S. (2013). Historic Building Information Modelling – Adding intelligence to laser and image based surveys of European classical architecture. *ISPRS Journal of Photogrammetry and Remote Sensing*, 76, 89–102. <https://doi.org/10.1016/j.isprsjprs.2012.11.006>

- Murtiyoso, A., & Grussenmeyer, P. (2023). Initial Assessment on the Use of State-of-the-Art NERF Neural Network 3D Reconstruction for Heritage Documentation. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-M-2–2023, 1113–1118. <https://doi.org/10.5194/isprs-archives-XLVIII-M-2-2023-1113-2023>
- Nagaoka, M. (2020). The Future of the Bamiyan Buddha Statues – Evolving Conservation Ethics and Principles concerning Intentionally Destructed Cultural Heritage. In *The Future of the Bamiyan Buddha Statues* (pp. 3–15). Springer International Publishing. https://doi.org/10.1007/978-3-030-51316-0_1
- Natampally, M. (2014). Reconstrucción visual (gráfica, ilustrada y digital) del Templo Hampi. *Virtual Archaeology Review*, 5(10), 117. <https://doi.org/10.4995/var.2014.4225>
- Navarro-Mateu, D., Carrasco, O., & Cortes Nieves, P. (2021). Color-Patterns to Architecture Conversion through Conditional Generative Adversarial Networks. *Biomimetics*, 6(1), 16. <https://doi.org/10.3390/biomimetics6010016>
- Nayak, R., & Balabantaray, B. K. (2021). *Generative Adversarial Network for Heritage Image Super Resolution*. 161–173. https://doi.org/10.1007/978-981-16-1092-9_14
- Newton, A., & Dhole, K. (2023). *Is AI Art Another Industrial Revolution in the Making?* <https://doi.org/10.48550/arxiv.2301.05133>
- Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Sutskever, I., & Chen, M. (2021). *GLIDE: Towards Photorealistic Image Generation and Editing With Text-Guided Diffusion Models*. <https://doi.org/10.48550/arxiv.2112.10741>
- Nichol, A., Jun, H., Dhariwal, P., Mishkin, P., & Chen, M. (2022). Point-E: A System for Generating 3D Point Clouds from Complex Prompts. In *arXiv.org*. <https://doi.org/10.48550/ARXIV.2212.08751>
- Nikolakopoulou, V., & Koutsabasis, P. (2020). *Methods and Practices for Assessing the User Experience of Interactive Systems for Cultural Heritage* (pp. 171–208). <https://doi.org/10.4018/978-1-7998-2871-6.ch009>
- Nikolakopoulou, V., & Koutsabasis, P. (2025). The ‘making’ of participatory and co-design for digital experiences in cultural heritage: a review. *CoDesign*, 1–27. <https://doi.org/10.1080/15710882.2025.2511791>
- Nyaaba, M., Wright, A., & Choi, G. L. (2024). *Generative AI and Digital Neocolonialism in Global Education: Towards an Equitable Framework*.
- Ocón, D., Yin, C., & Luna, J. (2026). Artificial insights or historical fidelity? Crafting an ethical framework for the use of GenAI in the restoration, reconstruction and recreation of movable cultural heritage. *AI & SOCIETY*, 41(1), 121–134. <https://doi.org/10.1007/s00146-025-02454-z>
- Ocón, S., Gómez, E., & Martínez, C. (2023). Ethics and AI for cultural heritage. *Digital Applications in Archaeology and Cultural Heritage*, 31, e00328. <https://doi.org/10.1016/j.daach.2023.e00328>
- OpenAI. (2024a). *OpenAI – Sora System Card*. <https://openai.com/index/sora-system-card/>
- OpenAI. (2024b). *OpenAI Sora System Card*.

- OpenAI. (2025). *OpenAI – Sora 2 System Card*. https://cdn.openai.com/pdf/50d5973c-c4ff-4c2d-986f-c72b5d0ff069/sora_2_system_card.pdf?utm_source=chatgpt.com
- Oppenheim, A. N. (2000). *Questionnaire Design, Interviewing and Attitude Measurement* (2nd ed.). Bloomsbury Academic / Continuum.
- Oppenlaender, J. (2022). The Creativity of Text-to-Image Generation. *Proceedings of the 25th International Academic Mindtrek Conference*, 192–202. <https://doi.org/10.1145/3569219.3569352>
- Orengo, H. A., Krahtopoulou, A., Garcia-Molsosa, A., Palaiochoritis, K., & Stamati, A. (2015). Photogrammetric re-discovery of the hidden long-term landscapes of western Thessaly, central Greece. *Journal of Archaeological Science*, 64, 100–109. <https://doi.org/10.1016/j.jas.2015.10.008>
- Pakkeerappa, P., & Thomas, J. (2006). Strategic Role of Hampi Development Authority in Promoting Tourism in Karnataka: a Study. *Atna - Journal of Tourism Studies*, 1(1), 86–95. <https://doi.org/10.12727/ajts.1.8>
- Pan, J., Li, L., Yamaguchi, H., Hasegawa, K., Thufail, F. I., & Tanaka, S. (2020). FUSED 3D TRANSPARENT VISUALIZATION FOR LARGE-SCALE CULTURAL HERITAGE USING DEEP LEARNING-BASED MONOCULAR RECONSTRUCTION. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, V-2–2020, 989–996. <https://doi.org/10.5194/isprs-annals-V-2-2020-989-2020>
- Paolanti, M., Frontoni, E., & Pierdicca, R. (2026). Towards trustworthy AI in cultural heritage. *Npj Heritage Science*, 14(1), 131. <https://doi.org/10.1038/s40494-026-02403-z>
- Petzet, M. (2009). *The Giant Buddhas of Bamiyan : safeguarding the remains / edited by Michael Petzet in cooperation with RWTH Aachen (Michael Jansen) and TU München (Erwin Emmerling)*. International Council On Monuments and Sites (ICOMOS),.
- Pierrot-Deseilligny, M., De Luca, L., & Remondino, F. (2011). Automated Image-Based Procedures for Accurate Artifacts 3D Modeling and Orthoimage Generation. *Geoinformatics FCE CTU*, 6, 291–299. <https://doi.org/10.14311/gi.6.36>
- Pietroni, E. (2021). The role of evaluation in virtual museums. *Heritage*, 4(3), 1825–1845. <https://doi.org/10.3390/heritage4030100>
- Pietroni, E., & Ferdani, D. (2021). Virtual Restoration and Virtual Reconstruction in Cultural Heritage: Terminology, Methodologies, Visual Representation Techniques and Cognitive Models. *Information*, 12(4), 167. <https://doi.org/10.3390/info12040167>
- Pocobelli, D. P., Boehm, J., Bryan, P., Still, J., & Grau-Bové, J. (2018). BIM for heritage science: a review. *Heritage Science*, 6(1), 30. <https://doi.org/10.1186/s40494-018-0191-4>
- Pollegioni, P., Woeste, K. E., Chiocchini, F., Del Lungo, S., Olimpieri, I., Tortolano, V., Clark, J., Hemery, G. E., Mapelli, S., & Malvolti, M. E. (2015). Ancient Humans Influenced the Current Spatial Genetic Structure of Common Walnut Populations in Asia. *PLOS ONE*, 10(9), e0135980. <https://doi.org/10.1371/journal.pone.0135980>
- Poole, B., Jain, A., Barron, J., & Mildenhall, B. (2022). DreamFusion: Text-to-3D using 2D Diffusion. In *ArXiv*. <https://doi.org/10.48550/ARXIV.2209.14988>

- Powell, S. (2018). Etched in Stone: Sixteenth-century Visual and Material Evidence of Śaiva Ascetics and Yogis in Complex Non-seated Āsanas at Vijayanagara. *Journal of Yoga Studies*, 1, 45–106. <https://doi.org/10.34000/JoYS.2018.V1.003>
- Quattrini, R., Malinverni, E. S., Clini, P., Nespeca, R., & Orlietti, E. (2015). From TLS to HBIM. High Quality Semantically-Aware 3D Modeling of Complex Architecture. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XL-5/W4, 367–374. <https://doi.org/10.5194/isprsarchives-XL-5-W4-367-2015>
- Radanovic, M., Khoshelham, K., & Fraser, C. (2020). Geometric accuracy and semantic richness in heritage BIM: A review. *Digital Applications in Archaeology and Cultural Heritage*, 19, e00166. <https://doi.org/10.1016/j.daach.2020.e00166>
- Raja, R., & Seland, E. H. (2022). The Paradox of Palmyra: An Ancient *anomalopolis* in the Desert. *Journal of Urban Archaeology*, 5, 177–189. <https://doi.org/10.1484/J.JUA.5.129848>
- Rajangam, K., & Sundar, A. (2021). Reading the Entanglements of Nature-culture Conservation and Development in Contemporary India. *Journal of South Asian Development*, 16(1), 7–32. <https://doi.org/10.1177/097317412111013676>
- Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022a). *Hierarchical Text-Conditional Image Generation With CLIP Latents*. <https://doi.org/10.48550/arxiv.2204.06125>
- Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022b). Hierarchical Text-Conditional Image Generation with CLIP Latents. *Arxiv*.
- Reade, J. E. (2002). The ziggurat and temples of Nimrud. *Iraq*, 64, 135–216. <https://doi.org/10.1017/S0021088900003697>
- Remondino, F. (2011). Heritage Recording and 3D Modeling with Photogrammetry and 3D Scanning. *Remote Sensing*, 3(6), 1104–1138. <https://doi.org/10.3390/rs3061104>
- Richards-Rissetto, H., & von Schwerin, J. (2017). Transparency and accountability in digital archaeology. *Journal of Archaeological Method and Theory*, 24, 1–24.
- Rico, T. (2016). *Constructing Destruction: Heritage Narratives in the Tsunami City*. Routledge. <https://doi.org/10.4324/9781315520933>
- Rienjang, W., & Stewart, P. (2023). *Gandhāran Art in Its Buddhist Context* (W. Rienjang & P. Stewart, Eds.). Archaeopress Archaeology. <https://doi.org/10.32028/9781803274737>
- Rihani, N. (2023). Interactive Immersive Experience: Digital Technologies for Reconstruction and Experiencing Temple of Bel Using Crowdsourced Images and 3D Photogrammetric Processes. *International Journal of Architectural Computing*, 21(4), 730–756. <https://doi.org/10.1177/14780771231168224>
- Risam, R. (2018). *New Digital Worlds Postcolonial Digital Humanities in Theory, Praxis, and Pedagogy*. Northwestern University Press. <https://doi.org/10.2307/j.ctv7tq4hg>
- Rombach, R., Blattmann, A., & Ommer, B. (2022). *Text-Guided Synthesis of Artistic Images with Retrieval-Augmented Diffusion Models*.

- Rössler, M. (2020). World Heritage and Reconstruction: An Overview and Lessons Learnt for the Bamiyan Valley. In *The Future of the Bamiyan Buddha Statues* (pp. 99–111). Springer International Publishing. https://doi.org/10.1007/978-3-030-51316-0_6
- Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. (2011). ORB: An efficient alternative to SIFT or SURF. *2011 International Conference on Computer Vision*, 2564–2571. <https://doi.org/10.1109/ICCV.2011.6126544>
- Sadek, M., Calvo, R. A., & Mougenot, C. (2024). Designing value-sensitive AI: a critical review and recommendations for socio-technical design processes. *AI and Ethics*, 4(4), 949–967. <https://doi.org/10.1007/s43681-023-00373-7>
- Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J., & Norouzi, M. (2022). *Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding*.
- Sbrogiò, L. (2022). Parametric approach to the reconstruction of timber structures in Campanian Roman houses. *Virtual Archaeology Review*, 13(26), 45–61. <https://doi.org/10.4995/var.2022.15319>
- Scherer, A. K. (2007). Population structure of the classic period Maya. *American Journal of Physical Anthropology*, 132(3), 367–380. <https://doi.org/10.1002/ajpa.20535>
- Schettino, P. (2016). Successful Strategies for Dealing With New Technology in Museums: A Case Study of Immersive Technology at the Immigration Museum, Melbourne. *Museum International*, 68(1–2), 130–135. <https://doi.org/10.1111/muse.12091>
- Schmidt-Colinet, A. (2019). No temple in Palmyra! Opposing the reconstruction of the Temple of Bel. *Syria Studies*, 11(2), 63–85. <https://ojs.st-andrews.ac.uk/index.php/syria/article/view/2015>
- Schmidt-Colinet, A., & Seigne, J. (2022). The battlements of temples in Hellenistic-Roman Syria. The crenellations of the Sanctuary of Bel in Palmyra reconsidered. *Revue Archéologique*, n° 73(1), 57–77. <https://doi.org/10.3917/arch.221.0057>
- Scorrano, G., Viva, S., Pinotti, T., Fabbri, P. F., Rickards, O., & Macchiardi, F. (2022). Bioarchaeological and palaeogenomic portrait of two Pompeians that died during the eruption of Vesuvius in 79 AD. *Scientific Reports*, 12(1), 6468. <https://doi.org/10.1038/s41598-022-10899-1>
- Senatore, M. R., Ciarallo, A., & Stanley, J. (2014). Pompeii Damaged by Volcaniclastic Debris Flows Triggered Centuries Prior to the 79 A.D. Vesuvius Eruption. *Geoarchaeology*, 29(1), 1–15. <https://doi.org/10.1002/gea.21458>
- Seyrig, H., Amy, R., & Will, E. (1975). Le temple de Bel à Palmyre. Vol. I: Texte et planches; Vol. II: Album. In *Revue Archéologique* (Number 1). Librairie Orientaliste Paul Geuthner. https://www.google.co.uk/books/edition/Le_Temple_de_B%C3%AAI_a_Palmyre/OJtbgtEACAAJ?hl=en
- Shelach-Lavi, G., Wachtel, I., Golan, D., Batzorig, O., Amartuvshin, C., Ellenblum, R., & Honeychurch, W. (2020). Medieval long-wall construction on the Mongolian Steppe during the eleventh to thirteenth centuries AD. *Antiquity*, 94(375), 724–741. <https://doi.org/10.15184/aqy.2020.51>
- Shi, Y., Wang, P., Ye, J., Mai, L., Li, K., & Yang, X. (2024). MVDream: Multi-view Diffusion for 3D Generation. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, & Y. Sun (Eds.),

- International Conference on Learning Representations* (Vol. 2024, pp. 39838–39859).
https://proceedings.iclr.cc/paper_files/paper/2024/file/adbe936993aa7cf41e45054d8b72f183-Paper-Conference.pdf
- Shih, N.-J. (2025). *AI-assisted 3D Preservation and Reconstruction of Temple Arts*.
- Shishido, H., Ito, Y., Kawamura, Y., Matsui, T., Morishima, A., & Kitahara, I. (2017). Proactive preservation of world heritage by crowdsourcing and 3D reconstruction technology. *2017 IEEE International Conference on Big Data (Big Data)*, 4426–4428.
<https://doi.org/10.1109/BigData.2017.8258479>
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420–428. <https://doi.org/10.1037/0033-2909.86.2.420>
- Silberman, N. A. (2012). Heritage interpretation and human rights: documenting diversity, expressing identity, or establishing universal principles? *International Journal of Heritage Studies*, 18(3), 245–256. <https://doi.org/10.1080/13527258.2012.643910>
- Silver, M., Fangi, G., & Denker, A. (2018). *Reviving Palmyra in Multiple Dimensions: Images, Ruins and Cultural Memory*. Whittles Publishing Ltd.
- Singh, K. (2022). When Peace Is Defeat, Reconstruction Is Damage: “Rebuilding” Heritage in Post-conflict Sri Lanka and Afghanistan. In J. CUNO & T. G. WEISS (Eds.), *Cultural Heritage and Mass Atrocities* (pp. 151–167). Getty Publications. <https://doi.org/10.2307/jj.6142257>
- Sloggett, R., & Scott, M. (2022). *Climatic and Environmental Threats to Cultural Heritage*. Routledge. <https://doi.org/10.4324/9781003163312>
- Smith, L. (2006). *Uses of Heritage*. Routledge. <https://doi.org/10.4324/9780203602263>
- Spallone, R., Lamberti, F., Olivieri, L. M., Ronco, F., & Castagna, L. (2022). AR AND VR FOR ENHANCING MUSEUMS’ HERITAGE THROUGH 3D RECONSTRUCTION OF FRAGMENTED STATUE AND ARCHITECTURAL CONTEXT. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLVI-2/W1-2022*, 473–480.
<https://doi.org/10.5194/isprs-archives-XLVI-2-W1-2022-473-2022>
- Spennemann, D. H. R. (2024). Generative Artificial Intelligence, Human Agency and the Future of Cultural Heritage. *Heritage*, 7(7), 3597–3609. <https://doi.org/10.3390/heritage7070170>
- Steinfeld, K. (2023). Clever Little Tricks: A Socio-Technical History of Text-to-Image Generative Models. *International Journal of Architectural Computing*, 21(2), 211–241.
<https://doi.org/10.1177/14780771231168230>
- Stobiecka, M. (2020). Archaeological heritage in the age of digital colonialism. *Archaeological Dialogues*, 27(2), 113–125. <https://doi.org/10.1017/S1380203820000239>
- Sun, T., Zhang, Y., & Fang, L. (2023). Diffusion models for image generation: A survey. *ACM Computing Surveys*, 55(12), 1–38. <https://doi.org/10.1145/3562082>
- Tiribelli, A., Foka, A., & Huggett, J. (2024). Algorithmic agency in cultural heritage. *Heritage*, 7(3), 418–436. <https://doi.org/10.3390/heritage7030049>

- Tiribelli, S., Pansoni, S., Frontoni, E., & Giovanola, B. (2024). Ethics of Artificial Intelligence for Cultural Heritage: Opportunities and Challenges. *IEEE Transactions on Technology and Society*, 5, 293–305. <https://doi.org/10.1109/TTS.2024.3432407>
- Tochilkin, D., Pankratz, D., Liu, Z., Huang, Z., Letts, A., Li, Y., Liang, D., Laforte, C., Jampani, V., & Cao, Y.-P. (2024). *TripoSR: Fast 3D Object Reconstruction from a Single Image*.
- Toubekis, G., Jansen, M., & Jarke, M. (2017). Long-Term Preservation of the Physical Remains of the Destroyed Buddha Figures in Bamiyan (Afghanistan) using Virtual Reality Technologies for Preparation and Evaluation of Restoration Measures. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, IV-2/W2, 271–278. <https://doi.org/10.5194/isprs-annals-IV-2-W2-271-2017>
- UNESCO. (2026). *Site of Palmyra*. UNESCO World Heritage Centre. <https://whc.unesco.org/en/list/23>
- UNESCO World Heritage Centre. (2011). *World Heritage List: Cultural Landscape and Archaeological Remains of the Bamiyan Valley*.
- Veysel Apaydin. (2020). *Critical Perspectives on Cultural Memory and Heritage* (V. Apaydin, Ed.). UCL Press. <https://doi.org/10.14324/111.9781787354845>
- von Schwerin, J., Richards-Rissetto, H., Remondino, F., Agugiaro, G., & Girardi, G. (2013). The MayaArch3D project: A 3D WebGIS for analyzing ancient architecture and landscapes. *Literary and Linguistic Computing*, 28(4), 736–753. <https://doi.org/10.1093/lc/fqt059>
- W3C. (2020). *JSON-LD 1.1: A JSON-based Serialization for Linked Data*.
- Wahbeh, W., Nebiker, S., & Fangi, G. (2016). COMBINING PUBLIC DOMAIN AND PROFESSIONAL PANORAMIC IMAGERY FOR THE ACCURATE AND DENSE 3D RECONSTRUCTION OF THE DESTROYED BEL TEMPLE IN PALMYRA. *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, III–5, 81–88. <https://doi.org/10.5194/isprsannals-III-5-81-2016>
- Wang, X., Lasaponara, R., Luo, L., Chen, F., Wan, H., Yang, R., & Zhen, J. (2020). Digital Heritage. In *Manual of Digital Earth* (pp. 565–591). Springer Singapore. https://doi.org/10.1007/978-981-32-9915-3_17
- Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., & Zhu, J. (2023). *ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation*.
- Waterton, E., & Watson, S. (2013). *Heritage and Community Engagement* (E. Waterton & S. Watson, Eds.). Routledge. <https://doi.org/10.4324/9781315875064>
- Wei, C., Kong, D., Wang, Y., Jia, J., Liu, J., & Wei, Y. (2026). AIGC based digital heritage reconstruction of Qing interior canopies. *Npj Heritage Science*, 14(1), 71. <https://doi.org/10.1038/s40494-025-02280-y>
- Westenberger, P., & Farmaki, D. (2025). Artificial intelligence for cultural heritage research: the challenges in UK copyright law and policy. *European Journal of Cultural Management and Policy*, 15. <https://doi.org/10.3389/ejcmp.2025.14009>
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR

- Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1), 160018. <https://doi.org/10.1038/sdata.2016.18>
- Wilson, E. B. (1927). Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association*, 22(158), 209–212. <https://doi.org/10.2307/2276774>
- Wotzlaw, J.-F., Bastian, L., Guillong, M., Forni, F., Laurent, O., Neukampf, J., Sulpizio, R., Chelle-Michou, C., & Bachmann, O. (2022). Garnet petrochronology reveals the lifetime and dynamics of phonolitic magma chambers at Somma-Vesuvius. *Science Advances*, 8(2). <https://doi.org/10.1126/sciadv.abk2184>
- Wu, S., Li, H., & Zhang, W. (2023). Prompt engineering for domain-specific image generation. *ArXiv Preprint ArXiv:2305.05685*.
- Wu, T., Yang, G., Li, Z., Zhang, K., Liu, Z., Guibas, L., Lin, D., & Wetzstein, G. (2024). *GPT-4V(ision) is a Human-Aligned Evaluator for Text-to-3D Generation*.
- Xie, Y., Pan, Z., Ma, J., Jie, L., & Mei, Q. (2023). A Prompt Log Analysis of Text-to-Image Generation Systems. *Proceedings of the ACM Web Conference 2023*, 3892–3902. <https://doi.org/10.1145/3543507.3587430>
- Xu, J., Cheng, W., Gao, Y., Wang, X., Gao, S., & Shan, Y. (2024). *InstantMesh: Efficient 3D Mesh Generation from a Single Image with Sparse-view Large Reconstruction Models*.
- Xu, J., Yan, L., Zhang, R., & Zhou, M. (2025). A review of the development and application of generative technology in digital museums. *Npj Heritage Science*, 13(1), 589. <https://doi.org/10.1038/s40494-025-02164-1>
- Yamauchi, K., & Bunka, I. K. K. S. (2012). *Documentation of the Bamiyan Buddhas* (Vol. 2). Japan Center for International Cooperation in Conservation. <https://books.google.co.uk/books?id=m5uwzQEACAAJ>
- Yang, M. (2017). Crossing between the Great Wall of China and the “Great” Trump Wall. *Palgrave Communications*, 3(1), 25. <https://doi.org/10.1057/s41599-017-0031-2>
- Yang, X., Grussenmeyer, P., Koehl, M., Macher, H., Murtiyoso, A., & Landes, T. (2020). Review of built heritage modelling: Integration of HBIM and other information techniques. *Journal of Cultural Heritage*, 46, 350–360. <https://doi.org/10.1016/j.culher.2020.05.008>
- Zeng, X., & Jin, T. (2023). 3D reconstruction of buildings based on transformer-MVSNet. In X. Chen & H. M. Srivastava (Eds.), *3rd International Conference on Applied Mathematics, Modelling, and Intelligent Computing (CAMMIC 2023)* (p. 191). SPIE. <https://doi.org/10.1117/12.2686276>
- Zhang, C., Zhang, C., Zhang, M., & Kweon, I.-S. (2023). Text-to-image Diffusion Models in Generative AI: A Survey. In *arXiv.org*. <https://doi.org/10.48550/ARXIV.2303.07909>
- Zhang, L., & Agrawala, M. (2023). *Adding Conditional Control to Text-to-Image Diffusion Models*.
- Zhou Wang, Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13(4), 600–612. <https://doi.org/10.1109/TIP.2003.819861>

Zhou, Y., Li, P., Ye, Z., Yue, L., Gui, L., Jiang, X., Li, X., & Liu, Y. (2022). Building information modeling-based 3D reconstruction and coverage planning enabled automatic painting of interior walls using a novel painting robot in construction. *Journal of Field Robotics*, 39(8), 1178–1204. <https://doi.org/10.1002/rob.22103>