Digital Applications in Archaeology and Cultural Heritage 42 (2026) e00554

Contents lists available at ScienceDirect

Digital Applications in Archaeology and Cultural Heritage

journal homepage: www.elsevier.com/locate/daach

HARF: A human–AI collaborative framework for cultural heritage reconstruction with expert-guided multi-platform generative AI and systematic prompt engineering

Kawsar Arzomand [a,*], Tatiana Kalganova [b], Michael Rustell [c]

[a] Department of Civil and Environmental Engineering, Brunel University, London, United Kingdom
[b] Department of Electronic and Computer Engineering, Brunel University, London, United Kingdom
[c] Inframatic Company, United Kingdom

ARTICLE INFO

Keywords:
Cultural heritage reconstruction
Generative artificial intelligence
Expert-in-the-loop validation
Buddhas of Bamiyan
Prompt engineering
Digital preservation

ABSTRACT

The destruction of cultural heritage through conflict and environmental degradation has created urgent needs for reliable digital reconstruction, yet current generative AI approaches often prioritise aesthetic plausibility over cultural authenticity, a risk exemplified by the loss of the Western Buddha of Bamiyan in 2001. This study introduces and validates the Heritage-Aligned Reconstruction Framework (HARF), a reproducible method that integrates archaeological evidence, dimensional analysis, and domain expertise to constrain AI-mediated reconstruction. HARF is structured around three components; a blueprint that encodes verified archaeological and iconographic information into prompts, a quantitative index assessing prompt completeness, and a geometric pre-evaluation stage that screens outputs against site-specific tolerances. Validation proceeded in two phases involving multidisciplinary feedback from 33 international specialists and targeted scoring by Bamiyan experts.

HARF consistently improved evidentiary alignment compared with unstructured prompting, with the strongest candidates reaching the minor-revision band of historical plausibility, though none attained exemplar-level fidelity. Expert assessments identified dimensional accuracy and niche geometry as the most reliable predictors of trust, while persistent challenges remained in head morphology and drapery articulation. By operationalising the principles of the London Charter and Seville Principles (Denard, 2009; López-Menchero Bendicho and Grande, 2017) within a transparent, auditable workflow, HARF provides a culturally grounded and methodologically accountable approach to reconstructing destroyed monuments. It offers a transferable model for guiding AI-generated heritage imagery toward evidential rigour, interpretive clarity, and cultural sensitivity, and can be applied to other destroyed or endangered monuments where archival imagery, dimensional documentation, and basic iconographic evidence are available.

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 and Scott, 2022; Veysel, 2020). Among the most emblematic cases is the destruction of the sixth-century 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; 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

* Corresponding author. Brunel University London Kingston Lane, Uxbridge, UB8 3PH, United Kingdom.
E-mail addresses: karzomand1@gmail.com (K. Arzomand), Tatiana.Kalganova@brunel.ac.uk (T. Kalganova), mike@inframatic.ai (M. Rustell).

https://doi.org/10.1016/j.daach.2026.e00554
Received 10 December 2025; Received in revised form 7 April 2026; Accepted 1 June 2026
Available online 3 June 2026

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 or replaced anachronism (Foka and Griffin, 2024a; Muldoon and 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 and Koutsabasis, 2025) promote inclusivity but are seldom integrated within computational workflows. Across this literature, three methodological deficiencies remain unresolved:

(1) prompt design and evaluation are inconsistent and often opaque,
(2) domain knowledge is loosely coupled with algorithmic scoring; and
(3) 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; 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 is built around three methodological components that structure the entire reconstruction workflow. First, the Dynamic Prompt Blueprint (DPB) translates verified architectural dimensions, iconographic and contextual evidence into layered prompt structure. Second, the Prompt Sufficiency Index (PSI) provides a quantitative measure of prompt completeness before generation, reducing reliance on trial-and-error iteration. Third, a site-specific quantitative pre-evaluation pipeline combining Oriented FAST and Rotated BRIEF (ORB) and Random Sample Consensus (RANSAC) feature matching, the Structural Similarity Index Measure (SSIM), and Histogram of Oriented Gradients (HOG) descriptors anchor generated outputs to pixel-metre calibration and world-scale acceptance of approximately one to three percent. Together, these components establish the evidentiary, quantitative, and interpretive foundations that guide all subsequent stages of reconstruction.

HARF was validated through a multi-phase expert study using a representative case from endangered world heritage. Thirty-three specialists reviewed DPB-structured prompts, PSI scores, and algorithmic indicators, assessing their historical plausibility and identifying context-specific inaccuracies. Their feedback informed iterative refinements that balanced quantitative optimisation with cultural authenticity.

This study makes four 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 reproducible metric for prompt adequacy. Third, it establishes a quantitative pre-evaluation protocol that evaluates generative outputs against geometric tolerances derived from archival measurements. The protocol is reliable for detecting proportional inconsistencies and large-scale geometric drift, but it is less sensitive to fine anatomical or stylistic deviations, which therefore remain dependent on expert review. 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 and Grande, 2011) within a transparent reconstruction workflow.

2. Methodology (Framework architecture and implementation workflow)

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 (Fig. 1).

2.1. Case study and evidentiary corpus

The Western Buddha of Bamiyan serves as the primary case study through which the HARF was developed and validated. 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 balance historical specificity with methodological transparency. The statue occupied the western niche of a monastic cave complex carved into the Bamiyan Valley's sandstone cliffs and surrounded by sculptural programmes associated with the Gandhāran–Central Asian tradition (Fig. 2).

A structured evidentiary corpus was assembled to support dimensional grounding, environmental context, and iconographic analysis. The corpus integrates nearly a century of archaeological, architectural, and conservation documentation and draws on institutional collections held by UNESCO's World Heritage Centre, the Afghanistan Institute of Archaeology, the Society for the Preservation of Afghanistan's Cultural Heritage (SPACH) and associated academic repositories. Early photographic campaigns by DAFA (Petzet, 2009; Toubekis et al., 2017) and the German Archaeological Institute (1920s–1960s) record proportional relationships, drapery rhythms, and niche morphology (Rienjang and Stewart, 2023). These visual records were supplemented by the Japanese cultural missions of the 1970s, whose measured drawings and architectural surveys remain the most precise records of the niche's geometry, including vaulted curvature, pedestal organisation, and wall contours. Scholarly analysis, that of (Klimburg-Salter, 1989), clarifies robe structuring, layering conventions, and stylistic lineage within the region's sculptural idiom. Together, these sources define the dimensional and iconographic evidence base on which historically grounded reconstruction depends (Andrew and Mark, 2016; UNESCO World Heritage Centre, 2011).

Post-destruction photogrammetric and laser-scanning campaigns coordinated by UNESCO preserve stable architectural boundaries such as the plinth, shoulder lines and vault outline. To enable quantitative comparison with generated imagery, these datasets were aligned with pre-2001 photographs and measured drawings to establish a unified spatial reference system. Archival measurements for niche height, shoulder width, pedestal proportions, and drapery fall lines were extracted from questionnaire data, cross-validated across independent sources and standardised within this framework. Reference outlines traced from the clearest pre-2001 imagery serve as geometric anchors for later algorithmic evaluation.

Where records diverged, the most consistently attested or technically precise values were retained, and all remaining uncertainties were

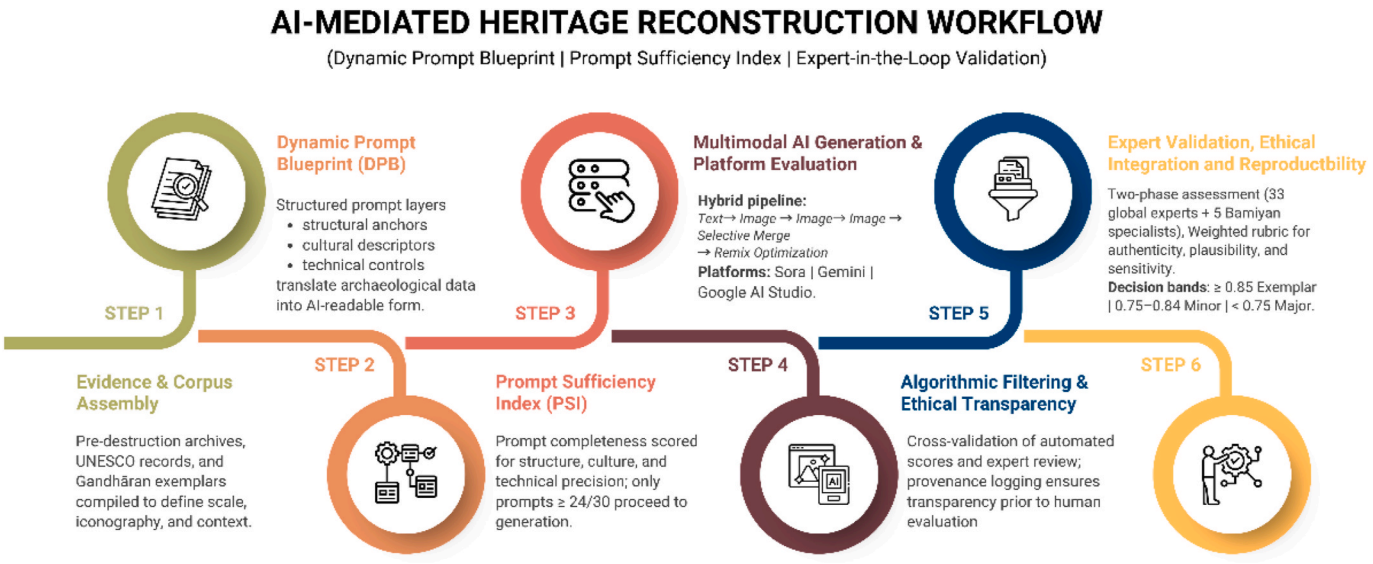

Fig. 1. Overview of the Heritage-Aligned Reconstruction Framework (HARF) and its six interconnected stages for the reconstruction of the Western Buddha of Bamiyan.

Fig. 2. Pre-2001 view of the Western Buddha of Bamiyan showing the niche architecture, drapery contours, and surrounding monastic caves. (Source: Khan Academy @UNESCO).

documented as paradata in line with the London Charter (Denard, 2009) and subsequent work on interpretive accountability in digital heritage (Bentkowska-Kafel and Denard, 2012; Richards-Rissetto and 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 interoperability across HARF's analytical stages, all corpus materials were digitised and standardised. Archival photographs were harmonised in resolution and annotated with camera-orientation metadata to support consistent spatial alignment. Architectural drawings were vectorised, georeferenced, and registered to a shared coordinate system to enable cross-mission comparison and dimensional verification. Textual documentation, including UNESCO mission reports, Japanese survey notes, and archaeological monographs, was encoded in JSON-LD (W3C, 2020) within a repository structured according to FAIR (Findable, Accessible, Interoperable, Reusable) Principles (Wilkinson et al., 2016) with persistent identifiers and explicit separation between descriptive metadata and paradata. Each record was catalogued with metadata specifying provenance, date, viewpoint, technical quality, and relevance to architectural or iconographic features.

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 algorithmic 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, 2022; 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 enables HARF to be applied to other heritage contexts: any monument with archival imagery, dimensional documentation, and descriptive iconography can be processed through the same pre-generation validation steps. The corpus therefore provides the evidentiary parameters within which reconstruction proceeds and ensures that all subsequent methodological stages remain accountable to the historical record.

2.2. Dynamic Prompt Blueprint (DPB) architecture

Prompt engineering in this study was guided by the DPB architecture, developed specifically for cultural heritage reconstruction. DPB represents one of the first structured prompting models designed to integrate archaeological, stylistic, and technical knowledge into the generative process. It was conceived to overcome the limitations of generic prompts that often misrepresent non-Western sites by relying on Western defaults. The model operates through three interconnected layers that balance scientific precision with cultural sensitivity.

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 (Liu and Chilton, 2022; Foka and Griffin, 2024a).

2.3. Prompt Sufficiency Index (PSI)

To ensure consistency and rigour in prompt formulation, PSI was introduced as a pre-generation quality control mechanism. The PSI evaluates the completeness of each prompt across the three DPB domains: structural, cultural, and technical. Each domain is scored on a ten-point scale, assessing dimensional accuracy, iconographic authenticity, and parameter specificity. Prompts reaching a total of 24 or above out of 30 qualify for generation, establishing a systematic threshold for quality and reducing the likelihood of algorithmic hallucination. This approach represents a methodological innovation in cultural heritage AI, applying principles of quantitative validation to the qualitative process of prompt design.

2.4. Multimodal generation workflow and platform evaluation

To produce reconstructions that maintained visual coherence and historical fidelity, the framework implemented a multimodal generative workflow extending beyond conventional single-prompt text-to-image synthesis. The approach integrated text generation, image-to-image refinement, and cross-platform compositing into an iterative workflow. Rather than using the AI platforms independently, the workflow integrated them sequentially so that the strengths of one system informed the next, allowing the models to function as collaborators. In this configuration, each model contributed distinct aesthetic and structural capabilities to the reconstruction process (Brooks and others, 2023; L. Zhang and Agrawala, 2023).

Initial generations were produced through a tri-platform configuration; OpenAI Sora, Gemini PRO, and Google AI Studio API, using DPB to balance fidelity and interpretive flexibility. The initial outputs served as visual hypotheses that guided subsequent decisions about composition and structure. Together, the sequential use of these platforms enabled the workflow to combine the proportional accuracy of Sora with the textural depth of Gemini, ensuring that no single system determined the reconstruction outcome.

Cross-platform refinement adjusted surface tone, volumetric balance, and lighting coherence while preserving geometric consistency. High-quality elements from each stage were composited through semantic segmentation and colour harmonisation (Avrahami and others, 2023; Lugmayr and others, 2022). Edge blending and atmospheric balancing ensured that transitions between merged regions retained perceptual continuity. Each synthesis stage included internal verification of geometric stability and dataset consistency before expert review, establishing a reproducible pipeline that could be applied to other heritage contexts. Quality control throughout the workflow prioritised architectural integrity, iconographic consistency, and algorithmic transparency. All generations were checked for conformance with DPB-encoded parameters. They were then evaluated against quantitative proportion thresholds and contextual descriptors derived from the PSI. Each iteration was versioned and archived with paradata documenting seed, token length, guidance scale, and compositing logic.

This multimodal synthesis operates as both a technical workflow and an evidence-interpretation strategy, integrating computational analysis with verified cultural documentation. By combining algorithmic precision with human interpretive oversight, the workflow positions generative AI as a complementary analytical tool rather than a replacement for historical expertise (Bau et al., 2020; Gal and others, 2023).

2.4.1. AI model evaluation and comparative analysis

To operationalise the workflow, seven major text-to-image and multimodal generation platforms were initially assessed: Midjourney v6 (Don-Yehiya et al., 2023), Stable Diffusion XL (Rombach et al., 2022) DALL-E 3 (Nichol et al., 2021; OpenAI. Dalle: Introducing Outpainting, 2023; Ramesh et al., 2022), Gemini PRO and Google AI Studio API (Google DeepMind, 2024; Saharia et al., 2022), Meta Imagine (Sun et al., 2023), and OpenAI Sora (OpenAI, 2024). Each system was pinned to a stable release to ensure reproducibility and tested under identical prompting conditions. Comparative assessments were conducted across four criteria central to cultural-heritage reconstruction: interpretive accuracy, historical-detail fidelity, generation speed, and stability across iterative refinements (Birhane et al., 2023; C. Zhang et al., 2023). Consistent with recent analysis of multimodal generation in cultural context (Cetinic, 2023), substantial variation was observed across models, with no platform demonstrating uniform superiority; instead, each exhibited distinctive strengths and limitations relevant to different phases of the reconstruction workflow.

Cross-platform trials revealed complementary strengths rather than a single optimal engine. OpenAI Sora produced the strongest results architectural precision and proportional coherence, particularly for iconographically dense compositions, though generation times averaged (45-60 s per image). Gemini PRO and Google AI Studio API provided faster inference times (7–15 s per image) and greater responsiveness to prompt refinements, making them suitable for expert-feedback cycles and iterative validation (Anil and others, 2023; Google DeepMind, 2024). DALL·E 3, Stable Diffusion XL, and Midjourney v6 served as comparative baselines, illustrating common diffusion-model tendencies toward stylistic drift and texture-level artefacts, while Meta Imagine achieved competitive visual quality but inconsistent geometric stability.

Accordingly, the research adopted a tri-platform configuration to balance fidelity and responsiveness. High-complexity reconstructions tasks, including primary reconstruction passes, iconographically critical regions, and final presentation outputs, were executed in *OpenAI Sora*, due to its superior proportional control; Rapid testing and refinement were undertaken in *Gemini PRO* and *Google AI Studio API*, whose shorter cycle times enabled prompt adjustments, alternative compositions, and expert-driven iteration without compromising evidentiary alignment.

Cross-platform consistency checks were subsequently applied to identify model-specific distortions or omissions that could undermine historical or iconographic integrity. This combined configuration provided high accuracy, fast iteration, and resilience against platform-specific biases or service variability (Hou and others, 2023) (Table 1). summarizes comparative performance across 50 standardised reconstruction prompts anchored in the Bamiyan corpus.

All platform versions, API configurations, and generation parameters were documented to ensure reproducibility and mitigate issues related to prompt drift and version instability in heritage-AI applications. In the final workflow, Sora v1.2, Gemini PRO v1.5, and Google AI Studio v2.1 were employed. Prompt windows were tested up to 300 tokens, guidance controls remained at platform defaults, and random seeds were fixed where exposed. Prompt construction followed the DPB, using consistent terminology, structured tiering, and systematic negative prompting. The DPB ensures that all descriptive elements are derived from verifiable evidence and are appended in a reproducible sequence. A full site-agnostic specification of the DPB schema, including all prompt fields and their documentation structure, is provided in Appendix A, which serves as the standard reference for prompt design within the HARF workflow. Output verification combined quantitative proportion checks (head:height, shoulder:height, base:height) with expert review and assessment of cultural and contextual coherence. All generated artefacts were archived with persistent metadata in a controlled, FAIR-aligned institutional repository, to ensure full traceability to their prompt configuration and methodological parameters. Table 2 summarizes the core model versions and generation parameters used across all HARF experiments.

Table 2
Summary of model versions and generation parameters used in the HARF workflow. Settings apply to all experiments unless otherwise indicated.

Component	Specification
Final platforms used	OpenAI Sora v1.2; Gemini Pro v1.5; Google AI Studio API v2.1
Generation mode	Tri-platform workflow: Sora for high-fidelity reconstruction; Gemini Pro and Google AI Studio for rapid testing and refinement
Prompt window	Up to 300 tokens per DPB-structured prompt
Aspect ratio	16:9
Image resolution	Platform maximum HD resolution
Guidance/control settings	Platform defaults
Random seeds	Fixed where exposed
Prompt construction standard	Dynamic Prompt Blueprint (DPB); consistent terminology; systematic negative prompting

2.5. Generation, pre-selection, and transparency

Using fifteen DPB-based prompt variants, over two hundred images were generated across Sora, Gemini PRO, and Google AI Studio under controlled parameters, including a 16:9 aspect ratio, maximum resolution, and fixed seed values. Image selection followed a two-stage process integrating computational assessment with expert review. A Python-based system first evaluated proportional alignment and compositional stability, after which domain specialists assessed anatomical plausibility, stylistic coherence, and the presence of anachronistic or culturally inconsistent elements. Several candidates that performed well under automated scoring were subsequently rejected during expert review, underscoring the limits of algorithmic metrics and the continuing necessity of informed human judgment in cultural-heritage reconstruction (Conte and others, 2023).

Post-processing was deliberately minimal, limited to framing adjustments and tonal normalisation to support cross-image comparison. No speculative or missing elements were manually corrected, unresolved gaps instead triggered prompt revision and re-generation. This ensured full paradata transparency, preserved traceability to algorithmic origin, and aligned the workflow with principles of evidentiary restraint in digital heritage (Huvila, 2022; López-Menchero Bendicho and Grande, 2017). The resulting corpus therefore reflects methodological precision and epistemic integrity, demonstrating how transparency functions as an ethical requirement in AI-mediated reconstruction.

2.6. Expert validation, ethical integration and reproducibility

The evaluation framework followed a sequential two-phase design that combined broad methodological validation with deep site-specific expertise. This structure addressed a longstanding challenge in digital heritage research: balancing general technical competency with culturally grounded contextual knowledge (Agapiou, 2021; Morgan and 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 wide geographic, disciplinary, and institutional diversity (Fig. 3).

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 min. 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

Table 1
Performance ratings based on systematic evaluation using 50 standardized heritage reconstruction prompts. Ratings represent mean expert assessments (n = 5 evaluators) on 10-point scales. Gen. time (s) = Generation time in seconds. Architectural Precision = expert scoring of geometric and proportional fidelity against archival measurements (e.g., niche geometry, height ratios, shoulder width). Qualitative bands correspond to the following score ranges: Excellent = 9.0–10.0, Very Good = 8.0–8.9, Good = 7.0–7.9, Fair = 5.0–6.9. All ratings apply the same 10-point evaluation scale, where 10 indicates highest performance. Relative cost categories are based on average API/subscription pricing at study time (Low < $0.05, Medium $0.05–0.10, High > $0.10 per HD image generation.

Platform	Gen. time (s)	Prompt Capacity	Historical Detail	Cultural Context	Arch. Precision	Relative per-image cost
OpenAI Sora	45-60 s	Excellent (>300 tokens)	Excellent (9.2/10)	Excellent (9.1/10)	Excellent (9.3/10)	Medium
Gemini Pro	8-15 s	Very Good (200-250 tokens)	Very Good (8.1/10)	Very Good (8.3/10)	Very Good (8.2/10)	High
Google AI Studio API	7-11 s	Very Good (200-250 tokens)	Very Good (8.0/10)	Very Good (8.2/10)	Very Good (8.1/10)	Very High
DALL·E 3	15-25 s	Limited (150-180 tokens)	Good (7.2/10)	Good (7.1/10)	Good (7.3/10)	Medium
Midjourney v6	20-35 s	Limited (100-150 tokens)	Good (7.0/10)	Fair (6.8/10)	Good (7.1/10)	Medium
Stable Diffusion XL	12-20 s	Good (180-220 tokens)	Fair (6.5/10)	Fair (6.3/10)	Fair (6.7/10)	Very High
Meta Imagine	10-18 s	Good (170-200 tokens)	Fair (6.8/10)	Fair (6.5/10)	Fair (6.9/10)	High

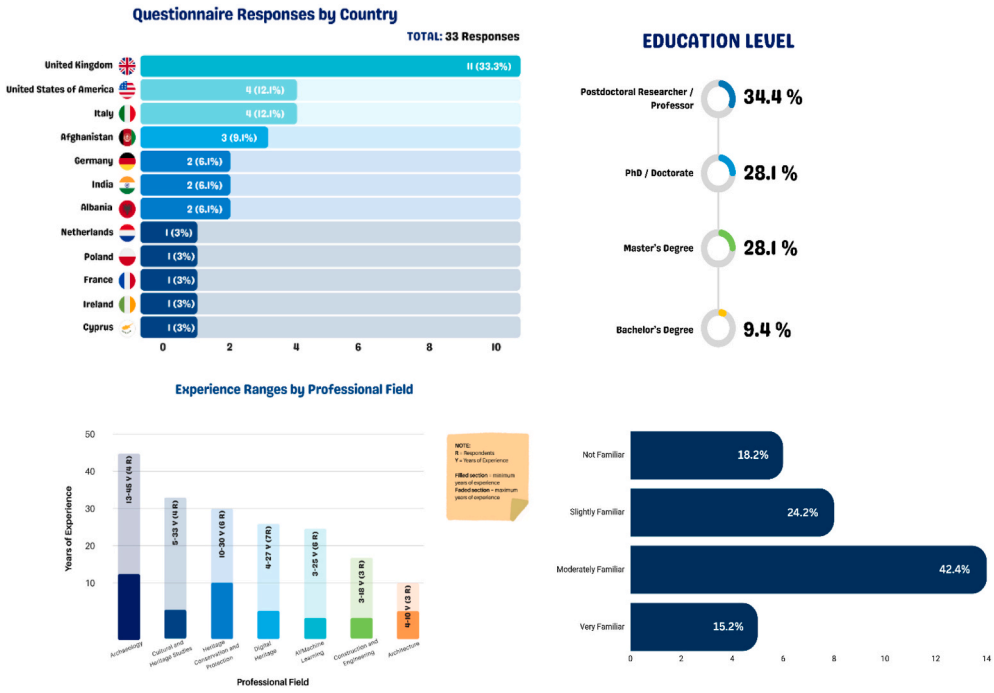

Fig. 3. Demographic profile of the Phase I expert panel (n = 33), showing country distribution, educational attainment, years of professional experience, and familiarity with the Buddhas of Bamiyan.

examples of the reconstruction candidates are shown in Fig. 4 to illustrate the evaluation material provided to experts. The complete set of evaluated reconstruction outputs and detailed image-level analysis are reported in a separate empirical study (Arzomand and Kalganova, 2026, under review), which presents the full candidate corpus and expert evaluation results.

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 and 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.

Fig. 4. Representative reconstruction candidates evaluated in Phase I.
(a) structural reconstruction (Image 1 - left) (b) iconographic detail (Image 13 - middle) (c) environmental context (Image 21 - right).

2.6.1. Data collection tools and procedures

Data collection employed two complementary instruments aligned with the comprehensive and specialist phases of evaluation. The design prioritised clarity, cross-disciplinary accessibility, and methodological transparency to ensure that both quantitative ratings and qualitative reflections contributed to the iterative reconstruction process.

Phase I utilised a 53-item instrument combining five-point Likert scales with open-ended prompts. Developed with reference to established digital-heritage assessment frameworks(Pietroni, 2021), the tool was piloted with five colleagues and supervisors to confirm interpretive clarity and establish an average completion time of 47 min (standard deviation, SD = 8). 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. All 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 employed a structured evaluation rubric applied to eighteen final-generation images refined through Phase I feedback. Although implemented in Excel for consistency, the rubric was based on established best-practice models in digital-heritage assessment and tailored to Bamiyan's iconographic and material context. Historical authenticity carried the highest weighting (25 points) and assessed facial attributes, mudrā gestures, drapery organisation, and the use of period-appropriate materials. Structural plausibility (20 points) measured proportional integrity and engineering feasibility against surviving architectural features. Cultural sensitivity (15 points) evaluated avoidance of speculative reconstruction and appropriate handling of the site's religious significance. Technical quality (15 points) assessed lighting coherence, atmospheric realism, and resolution suitability for scholarly interpretation. Scores were normalised to a 0–1 scale by dividing achieved points by the applicable weight, with interpretive thresholds set at $\geq$0.85 (exemplar), 0.75–0.84 (minor revision), and $<$0.75 (major revision). Items deemed un-assessable were marked as N/A, and normalisation was adjusted to preserve consistent weighting across reviewers.

2.6.2. 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 (mean, median, and dispersion measures) described scoring patterns within each domain and identified where reviewers converged or diverged. Subgroup comparisons by disciplinary orientation, site familiarity, seniority, and institutional affiliation examined whether reviewer characteristics shaped assessments of historical plausibility or authenticity. Reliability of ordinal ratings in Phase I was estimated using Krippendorff's alpha, while agreement among Phase II specialists was assessed using Kendall's W and Kendall's τ. Bootstrap resampling ($n = 1000$) generated confidence intervals, and multiple comparisons were adjusted using Benjamini–Hochberg procedures.

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.

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 all methodological decisions remained explicitly linked to their evidentiary basis.

2.6.3. Ethical considerations and data management

The study complied with institutional ethics requirements and the British Academy of Management Ethics Guidelines (2021), ensuring informed consent, anonymity, and full withdrawal rights. Given the sensitive history of iconoclasm in Bamiyan, the analysis incorporated explicit reflexivity regarding the positional and interpretive responsibilities involved in reconstructing a sacred monument. Ethical safeguards emphasised cultural sensitivity and evidentiary restraint: undocumented iconographic or ceremonial features were excluded, speculative components were explicitly identified, and specialist consultation ensured respect for Buddhist and Afghan heritage perspectives. These practices align with emerging ethical frameworks for AI in cultural heritage that emphasise transparency, cultural sensitivity, and accountability (Foka and Griffin, 2024a; Ocón et al., 2023). Data governance followed institutional policy and adhered to FAIR principles, with questionnaire data, reconstruction outputs, and analytic scripts stored in a secure, version-controlled repository. Participant information was processed under UK GDPR through pseudonymisation, restricted access, and encrypted storage.

Legal compliance was maintained across three domains: (1) cultural-property obligations under the 1970 UNESCO Convention and the 1972 World Heritage Convention, ensuring that reconstructions were non-commercial and clearly identified as research interpretations; (2) copyright and licensing restrictions governing third-party archival materials; and (3) adherence to the ethical principles of responsible visualisation in digital heritage. Together, these measures demonstrate that credible digital reconstruction depends not only on technical precision but also on cultural accountability and procedural transparency.

2.6.4. Methodological Limitations, risk mitigation, and reproducibility

AI-assisted reconstruction involves unavoidable interpretive uncertainty. To ensure methodological transparency, the study identified risks that could affect reproducibility, fidelity, or cultural validity. Risks were first detected through pilot testing, discrepancy logs generated during automated pre-selection, and systematic patterns in expert feedback from both evaluation phases. Each risk was mapped to its observable effects withing the workflow, (e.g. changes in geometric rankings, repeated regeneration triggers, or recurrent misinterpretations raised by specialists) allowing its impact to be assessed against the evidentiary constraints defined earlier in the methodology.

Residual severity was assigned using a structured rubric based on (1) frequency, (2) impact on reconstruction outcomes, and (3) cultural or ethical significance. A risk was designated *Critical* when it could still compromise cultural or ethical integrity after mitigation; *High* when it could alter rankings, thresholds, or require major re-analysis; *Moderate* when its effects were local and did not affect overall conclusions; and *Low* when its impact was negligible.

These mitigations, from version pinning and discrepancy logging to culturally stratified analysis and explicit disclaimers, reflect a broader commitment to methodological transparency. Rather than treating uncertainty as error, the workflow incorporates it as evidentiary data, embedding reflexive awareness within technical design. This approach positions reproducibility not only as a procedural requirement but as an ethical obligation, where the visibility of methodological decisions functions as a safeguard for cultural integrity.

2.6.5. Paradata and metadata capture

To ensure full traceability within HARF, the study distinguished between metadata that recorded the technical provenance of each generative iteration and paradata that documented the interpretive decisions shaping reconstruction. The structure followed London Charter (Denard, 2009) and aligned with later work emphasising accountability and transparency in digital-heritage workflows

(Bentkowska-Kafel and Denard, 2012; Richards-Rissetto and von Schwerin, 2017).

Metadata recorded the technical parameters for reproducibility, including platform and model versions, API settings, seed values when available, generation timestamps, and the DPB and PSI configurations applied at each iteration. Each Output was logged with its quantitative pre-evaluation measures (SSIM, HOG descriptors, region-of-interest masks), providing a reproducible technical record against which variations could be audited. Paradata documented the interpretive pathway that guided iterative refinement. This included expert rationales, explanations for acceptance or rejection of images, prompt-revision notes, iteration-change logs, and records of cases where algorithmic metrics conflicted with expert assessment. Such entries clarified how evidentiary constraints, cultural expertise, and computational indicators were weighed when resolving uncertainty.

3. Results

The evaluation produced three principal findings on the operational performance of the framework. Experts observed that the prompt architecture directly shaped proportional, iconographic, and textural behaviour across reconstruction cycles. PSI scores differentiated prompts that lacked essential structural or cultural content from those that met the threshold for reliable generation. Written feedback further 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. Together, they form the empirical basis for assessing how well-structured prompting and sufficiency scoring support historically grounded reconstruction.

3.1. 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. Almost two-thirds of participants assessed the prompt structure as Highly or Mostly sufficient, though several core descriptive layers required further development, and one-third rated coverage as Moderate (Fig. 5). The PSI mean

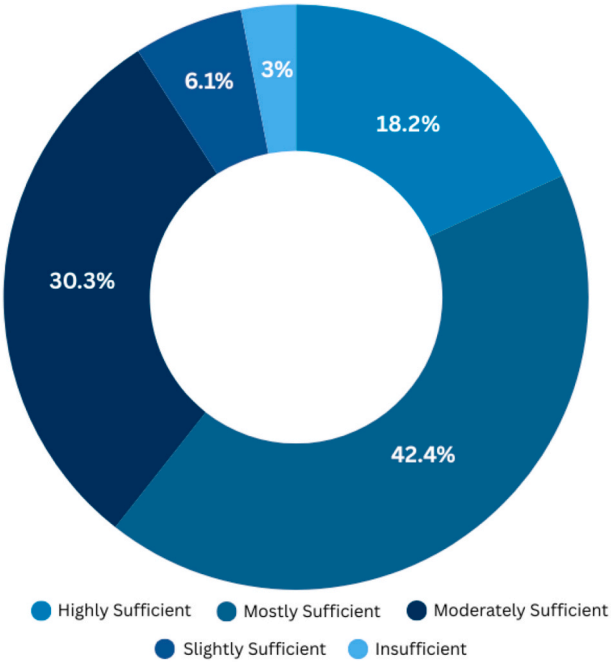

Fig. 5. Expert ratings of baseline prompt sufficiency (n = 33).

averaged 3.6 out of 5, suggesting that the framework captured the essential elements, but several contextual details required clearer and more specific articulation (OECD, E. U. E. C. J. R. C., 2008).

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 (Fig. 6). These domains represented 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.

The six refinement themes (Table 3) 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. These themes indicate where ambiguity or omission was most likely to occur in the baseline prompt design (see Table 4).

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 highlighted that this tiered structure improved reproducibility, since descriptive and ethical decisions were consistently logged in defined fields rather than dispersed across narrative text.

Experts assessed sufficiency according to the completeness and clarity of slot coverage. A prompt is considered sufficient when its core anchors are fully specified, contextual descriptors are grounded in verifiable evidence and technical controls record the reasoning behind inclusions and exclusions. Experts confirmed that the PSI scores reliably highlighted underspecified components and guided targeted refinement.

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 sufficiency but also clarified how architectural formalisation supports transparency, transferability and reproducibility across heritage reconstruction scenarios.

3.2. Quantitative PSI and refinement outcomes

The PSI synthesises expert assessments by weighting each domain's importance (W_i) and multiplying it by the corresponding sufficiency rating (R_i), averaged over n domains:

$$PSI = \frac{\sum_{i=1}^{n} (W_i \cdot R_i)}{n}$$

Where:

- W_i : expert-assigned weight of importance for domain i (scale 1–5)
- R_i: sufficiency rating assigned to that domain's keyword coverage (scale 1–5)
- : total number of assessed domains

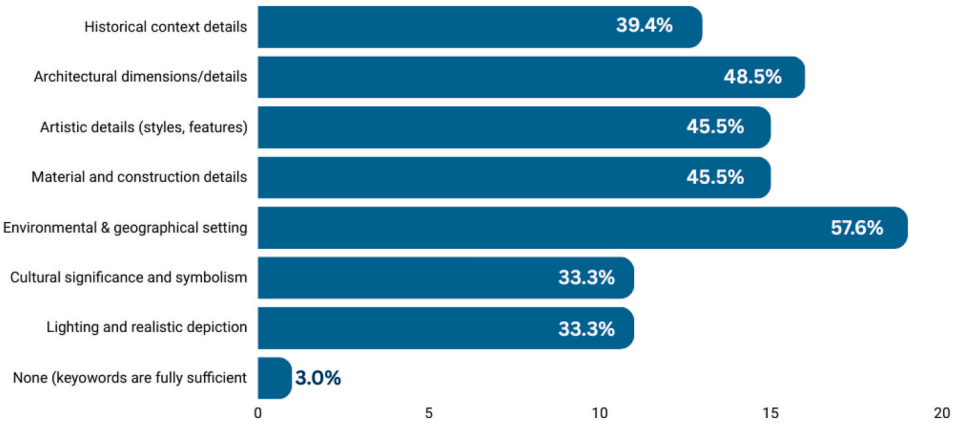

Fig. 6. Domains prioritised by experts for prompt enrichment (n = 33). The distribution visualises which categories were most frequently prioritised for refinement in AI-guided heritage visualisation.

Table 3
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/version drift can undermine reproducibility.	Version pinning; timestamped prompts; saved seeds/metadata; full provenance for each generation.	High
Expert fatigue/ forced response	53-item load may reduce response quality for less-familiar participants.	Familiarity-based segmentation; "uncertain" option; analysis by familiarity bands; monitor completion times.	Moderate
Automated pre-selection conflicts	Algorithmic scores can mis-rank culturally correct images.	Discrepancy logs; expert primacy in final selection; conflicts trigger re-generation, not manual fixes.	High
Cultural representation bias	Western-centric frames may skew evaluations.	Diverse international panel; explicit inclusion of Central Asian & Buddhist specialists; lens-stratified analysis.	Critical
Temporal-context challenge	Reconstructing 6th-century forms with 21st-century AI trained on contemporary imagery.	Strong negative prompting; systematic anachronism suppression; conservative treatment of speculative elements with clear disclaimers.	High

Using seven domains; environmental context, architectural dimension, artistic detailing, material realism, historical markers, cultural symbolism, and lighting, the Bamiyan instantiation produced 97/ 7 ≈ 13.86, which normalises to 3.6/5. The lowest ratings were concentrated in environmental description, material treatment, and symbolic content, indicating the areas in which descriptive coverage was least complete. Re-scoring after each refinement cycle showed measurable increases in domains where additional detail had been incorporated, demonstrating that the PSI provided a practical means for tracking incremental improvements in prompt specification.

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, e.g. stepped Gandhāran pleats consistent with wet-fold drapery or stratified sandstone with monastic cells. This refinement strengthened the clarity of key descriptive elements and reduced ambiguity in categories associated with lower PSI scores.

Table 4
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.'"
3D realism 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."

3.3. Cross-site evaluation

The DPB was designed to remain structurally constant even when applied to different cultural or archaeological contexts. Slots populated with Bamiyan-specific descriptors were substituted with equivalents from neighbouring or parallel traditions without altering the hierarchy of evidential tiers. Table 5 presents a condensed example in which monument identifiers, material references, environmental descriptors, and symbolic markers were replaced with variables appropriate to sites

Table 5
Cross-site substitution of DPB structural tiers.

Category	Bamiyan Instantiation	Cross-Site Substitution
Monument	Western Buddha (Salsal)	Temple of Bel, Palmyra
Material	Mud-straw plaster	Roman concrete with limestone veneer
Figural Context	Standing Buddha, Abhaya mudrā	Corinthian columns, Roman temple façade
Environment	Sandstone cliff with 750 monastic caves	Hindu temple complex with pillared mandapas
Symbolism	Gandhāran Indo-Hellenistic influences	Khmer cosmological motifs (Preah Vihear)

such as Palmyra or Preah Vihear. The substitution exercise showed that the DPB could accommodate these site-appropriate terms while preserving the underlying tier logic. Reviewers noted that the schema remained readable and transparent across cases, and that descriptive decisions were traceable when populated with locally verified terminology. This confirmed that the framework could support comparative reconstruction without requiring structural modification.

The ability to replace Bamiyan-specific clauses with site-appropriate alternatives while maintaining the same structural tiers demonstrates the DPB's modular transferability. It supports the use of a consistent prompt architecture across culturally distinct sites when grounded in locally verified evidence.

3.4. 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 agreed that the DPB's structured recording of evidential anchors, contextual assumptions, and negative prompts improved the traceability of reconstruction decisions. This assessment was consistent with the principles of the London Charter (Denard, 2009) and subsequent guidance on documentation and accountability in virtual archaeology (López-Menchero Bendicho and Grande, 2011).

Participants highlighted three aspects of the workflow that supported ethical clarity:

1. **Controlled exclusion:** Systematic omission of undocumented features, such as unverified decorative elements (haloes, unrecorded wall paintings, or uncertain façade symmetries) are excluded from text prompt formulation. The framework also employs negative prompting protocols to prevent speculative inflation of the imagery and reduced the risk of culturally inconsistent representations.
2. **Documentation of inference:** Interpretive steps are explicitly annotated with their rationale, notes on comparative reasoning and confidence levels. This documentation enables clear differentiation between verifiable evidence and scholarly interpretation within the prompt structure, aligning with best practice for paradata documentation (Bentkowska-Kafel and Denard, 2012)
3. **Transparent output metadata**: Each completed reconstruction is accompanied by a metadata statement outlining its evidential scope, temporal framing, and uncertainties was regarded as helpful for scholarly interpretation and for communicating limitations to non-expert audiences (Richards-Rissetto and von Schwerin, 2017).

Across these areas, experts characterised the DPB's transparency protocols as a practical means of maintaining cultural accountability and interpretive clarity. The evaluation indicated that the framework enables reconstructions to remain consistent with available evidence while clearly signalling the limits of what can be reliably represented.

3.5. 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 quantitative validation mechanism. The PSI enables descriptive completeness to be assessed systematically, allowing targeted refinement of underspecified domains and offering a transparent record of how expert guidance shapes each iteration. Expert review forms the fourth component, ensuring that prompt revision remains aligned with established archaeological and art-historical canons rather than with platform-specific tendencies. The final contribution concerns ethical clarity: the DPB embeds procedures for identifying uncertainty, recording interpretive rationale, and excluding undocumented or speculative features, consistent with current expectations of cultural responsibility in digital reconstruction (Brosché et al., 2017; Kila, 2019; Sloggett and Scott, 2022; Veysel, 2020).

These mechanisms reposition the prompt as a scholarly tool that mediates between empirical evidence and computational generation. The DPB provides a means of producing visualisations whose descriptive content, interpretive reasoning, and procedural constraints are explicitly documented and therefore open to verification. The framework is not limited to the Bamiyan case. Its tier-based structure allows evidential descriptors to be substituted with site-specific terminology while preserving the underlying logic and validation procedure. This adaptability supports comparative studies and enables reconstructions across different cultural, architectural, and chronological contexts to be produced within a coherent methodological standard.

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

4. Limitations and future work

The framework presented here demonstrates that prompt design for heritage reconstruction can be rendered evidential, auditable, and reproducible. 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-Menchero Bendicho and Grande, 2011, 2017; Bendicho & Grande, 2011; 2017) First, the generative platforms remain volatile and evolving. First, the reproducibility of generative outputs remains partly dependent on the volatility of commercial models. Version updates, opaque adjustments to training data, and non-deterministic sampling mean that prompt-level traceability does not guarantee identical results across time or across platforms, even under version pinning and full metadata capture. HARF mitigates this instability through the DPB and PSI, which hard-link each prompt clause to evidential sources and record interpretive choices as paradata, yet complete temporal reproducibility is not currently achievable within existing model-governance structures.

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, PSI scores reflect the completeness of the schema rather than the completeness of the historical record, and descriptive adequacy must be interpreted in relation to the depth and reliability of available sources. This is particularly evident for anatomical and facial regions of the Western Buddha that had deteriorated long before 2001, where the absence of stable visual anchors places principled constraints on what can be reconstructed.

Third, biases embedded within training data continue to affect representational accuracy. Generative systems trained on large-scale, unbalanced, and predominantly Western visual datasets reproduce stylistic conventions that conflict with regional sculptural grammars, especially in non-Western contexts. These biases manifest as symmetrical drapery, cinematic surface treatments, and hybridised morphological features. While constraint-based prompting and expert evaluation attenuate these tendencies, they cannot fully compensate for

structural imbalances in the training data, reinforcing the need for curated, domain-specific datasets.

Fourth, the scope of expert input poses further limitations. Although the validation process achieved disciplinary breadth and methodological diversity, it underrepresented perspectives from local and descendant communities, and religious practitioners. Incorporating these viewpoints is essential for cultural accountability and for broadening interpretive legitimacy; however, doing so raises methodological questions concerning authority, representational plurality, and the integration of community-based knowledge within structured computational workflows.

Finally, this implementation focuses on two-dimensional still imagery. Extending HARF to volumetric or temporal reconstructions, such as 3D models, animations, or immersive environments, introduces distinct challenges related to spatial coherence, lighting stability, uncertainty communication, and the calibration of geometric tolerances in three dimensions.

Future work should refine the PSI weighting scheme across heritage domains and test its sensitivity to incremental prompt adjustments. A dedicated 3D pipeline that integrates photogrammetry, laser scanning, and survey-grade anchors would enable quantitative validation of volumetric fidelity while retaining interpretive flexibility. Collaboration between heritage institutions and AI developers will be essential for building balanced training corpora that mitigate stylistic drift and support culturally attuned generation. Establishing FAIR-compliant repositories for prompts, model versions, and paradata would further strengthen reproducibility as generative platforms evolve. Equally important is the systematic communication of uncertainty through layered visualisations, alternative reconstruction scenarios, and interactive interfaces that distinguish evidence from interpretation. At the policy level, extending the London and Seville frameworks to include AI-specific guidance on disclosure, metadata retention, and community participation would provide a normative foundation for responsible deployment of generative systems in heritage contexts.

5. Conclusion

This study demonstrated that generative AI could support the reconstruction of destroyed heritage when its operation is anchored by evidence and embedded within structured expert oversight. Quantitative analysis showed that the strongest candidates produced within the Heritage-Aligned Reconstruction Framework (HARF) consistently fell within the *minor revision* range (Weighted Scores 0.75–0.84), with none achieving exemplar-level plausibility. These evaluations proved stable across bootstrap resampling and weighting perturbations, confirming that the framework improves historical plausibility while delineating clear limits in current model capability.

Expert evaluation clarified the sources of both progress and persistent failure. Improvements were most evident within strongest dimensional constraints, such as the head–bun separation ($\Delta = +0.13$, mean improvement; Cohen's $d = 0.59$, moderate effect size). By contrast, elements requiring nuanced iconographic understanding, drapery articulation, hair morphology, and aspects of facial modelling, remained inconsistent across platforms. Reviewers consistently identified niche geometry, proportional fidelity, and adherence to Gandhāran sculptural conventions as primary source of trust, while symmetrical artefacts, modernising tendencies, and inconsistent anatomical treatment were recurrent indictors of inauthenticity. These findings demonstrate that computational alignment must operate together with interpretive discipline to produce culturally credible outcomes.

Methodologically, HARF formalises the relationship between evidence, prompt design, and evaluation by linking dimensional anchors, documentary sources, and algorithmic indicators into a transparent and reproducible workflow. Its modular architecture allows adaptation to other heritage contexts without enforcing stylistic uniformity, providing scalable method for sites where documentation is uneven or fragmentary. This positions HARF as a bridge between computational generation and the evidentiary standards of archaeological and art-historical practice.

Limitations remain where the surviving record is incomplete or degraded. Archival gaps, particularly in facial and anatomical details that had eroded before 2001, constrained achievable fidelity. Model biases associated with globally imbalanced training data persist in surface-level distortions, and platform variability also affected consistency despite controlled settings. Addressing these constraints will require domain-adapted training datasets, cross-regional validation, and deeper integration of community-based knowledge where archival gaps can be improved through cultural memory.

Taken together, the results indicate that credible reconstruction is possible only when generative models operate within explicit evidentiary and interpretive boundaries. Future reliability of AI-mediated heritage reconstruction will depend less on advances in visual realism than on sustained methodological rigour, transparency, and cultural accountability.

Ethical approval

This study received ethical approval from the **Brunel University London College of Engineering, Design and Physical Sciences Research Ethics Committee** (Reference: **50187-LR-Jan/2025–53624–2**).

The expert-evaluation survey included an information and consent statement outlining voluntary participation, confidentiality, anonymised data handling, and secure storage of responses.

Funding statement

This research was supported through the British Council's Warm Welcome Scholarships scheme. The funding body had no role in the design of the study; in the collection, analysis, or interpretation of data; or in the writing of the manuscript.

CRediT authorship contribution statement

Kawsar Arzomand: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Tatiana Kalganova:** Project administration, Supervision. **Michael Rustell:** Supervision.

Declaration of competing interest

The authors declare that they have **no known competing financial interests or personal relationships** that could have appeared to influence the work reported in this paper.

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

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 → 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 ≈ *[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 a companion empirical study; in the present article it is included to define the evaluation framework rather than to reproduce 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.

References

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.

Andrew, Bevan, Mark, Lake, 2016. In: Bevan, A., Lake, M. (Eds.), Computational Approaches to Archaeological Spaces. Routledge. https://doi.org/10.4324/9781315431932.

Anil, R., others, 2023. PaLM 2 Technical Report. *Google Research Report*.

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.

Arzomand, K., Kalganova, T., 2026. Evaluating generative AI reconstructions of the bamiyan buddhas: computational similarity and expert archaeological assessment [Preprint]. Zenodo. https://doi.org/10.5281/zenodo.20435265.

Avrahami, O., others, 2023. Blended latent diffusion. ACM SIGGRAPH 2023.

Bau, D., Liu, S., Wang, T., Zhu, J.-Y., Torralba, A., 2020. Rewriting a deep generative model. Computer Vision - ECCV 351–369. https://doi.org/10.1007/978-3-030-58452-8_21.

Bentkowska-Kafel, A., Denard, H., 2012. In: Bentkowska-Kafel, A., Hugh, Denard (Eds.), Paradata and Transparency in Virtual Heritage, first ed. Ashgate https://www.researchgate.net/publication/287270234_Paradata_and_transparency_in_virtual_heritage.

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.

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. Int. J. Herit. Stud. 23 (3), 248–260. https://doi.org/10.1080/13527258.2016.1261918.

Cetinic, E., 2023. AI-generated imagery and cultural heritage: a critical perspective. Heritage Science 11 (1), 72. https://doi.org/10.1186/s40494-023-00841-2.

Chang, K.K., Cramer, M., Soni, S., Bamman, D., 2023. Speak, Memory: an Archaeology of Books Known to ChatGPT/GPT-4.

Conte, C., others, 2023. AI in cultural heritage: Bias and evaluation. Heritage Science 11 (1), 103.

Denard, H., 2009. The London Charter for the Computer-based Visualisation of Cultural Heritage. King's College London. https://www.londoncharter.org/.

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.

Flood, F.B., 2002. Between cult and culture: bamiyan, Islamic iconoclasm, and the Museum. Art Bull. 84 (4), 641–659. https://doi.org/10.1080/00043079.2002.10787045.

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 Soc. https://doi.org/10.1007/s00146-025-02349-z.

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.

Gal, R., others, 2023. Textual inversion: generating personalized images. CVPR 2023.

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.

Google DeepMind, 2024. Gemini 1.5 Model Report.

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.

Hou, J., others, 2023. Cross-model evaluation of diffusion systems. CVPR Workshops.

Huvila, I., 2022. Transparency and ethics in information work. J. Doc. 78 (7), 1492–1507.

Kila, J.D., 2019. Iconoclasm and cultural heritage destruction during contemporary armed conflicts. In: The Palgrave Handbook on Art Crime. Palgrave Macmillan, UK, pp. 653–683. https://doi.org/10.1057/978-1-137-54405-6_30.

Klimburg-Salter, D., 1989. The Kingdom of Bamiyan : Buddhist Art and Culture of the Hindu Kush. Instituto Universitario Orientale/Istituto Italiano per il Medio ed Estremo Oriente. https://books.google.co.uk/books?id=XRZuOQAACAAJ.

Liu, Vivian, Chilton, Lydia B, 2022. Design Guidelines for Prompt Engineering Text-to-Image Generative Models. CHI Conference on Human Factors in Computing Systems 1–23. https://doi.org/10.1145/3491102.3501825, ACM, New York, NY, USA. https://dl.acm.org/doi/10.1145/3491102.3501825. (Accessed 29 April 2022).

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/.

Lugmayr, A., others, 2022. RePaint: inpainting using diffusion models. CVPR 2022.

Marwick, B., 2022. Open science and reproducibility in archaeology. Adv. Archaeol. Pract. 10 (1), 1–13.

Morgan, C., Eve, S., 2012. DIY and digital archaeology: what are you doing to participate? World Archaeol. 44 (4), 521–537. https://doi.org/10.1080/00438243.2012.741810.

Muldoon, J., Wu, B.A., 2023. Artificial intelligence in the colonial matrix of power. Philos. Technol. 36 (4), 80. https://doi.org/10.1007/s13347-023-00687-8.

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, 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.

OECD, E. U. E. C. J. R. C., 2008. Handbook on Constructing Composite Indicators: Methodology and User Guide. OECD. https://doi.org/10.1787/9789264043466-en.

OpenAI, 2024. OpenAI Sora System Card.

Oppenheim, A.N., 2000. Questionnaire Design, Interviewing and Attitude Measurement, second ed. Bloomsbury Academic/Continuum.

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).

Pietroni, E., 2021. The role of evaluation in virtual museums. Heritage 4 (3), 1825–1845. https://doi.org/10.3390/heritage4030100.

Richards-Rissetto, H., von Schwerin, J., 2017. Transparency and accountability in digital archaeology. J. Archaeol. Method Theor 24, 1–24.

Rienjang, W., Stewart, P., 2023. In: Rienjang, W., Stewart, P. (Eds.), Gandhāran Art in its Buddhist Context. Archaeopress Archaeology. https://doi.org/10.32028/9781803274737.

Rombach, R., Blattmann, A., Ommer, B., 2022. Text-Guided Synthesis of Artistic Images with Retrieval-Augmented Diffusion Models.

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.

Shih, N.-J., 2025. AI-assisted 3D Preservation and Reconstruction of Temple Arts.
Sloggett, R., Scott, M., 2022. Climatic and Environmental Threats to Cultural Heritage. Routledge. https://doi.org/10.4324/9781003163312.
Sun, T., Zhang, Y., Fang, L., 2023. Diffusion models for image generation: a survey. ACM Comput. Surv. 55 (12), 1–38. https://doi.org/10.1145/3562082.
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.
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 World Heritage Centre, 2011. World Heritage List: Cultural Landscape and Archaeological Remains of the Bamiyan Valley.
Veysel, Apaydin, 2020. In: Apaydin, V. (Ed.), Critical Perspectives on Cultural Memory and Heritage. UCL Press. https://doi.org/10.14324/111.9781787354845.
W3C, 2020. JSON-LD 1.1: a JSON-based Serialization for Linked Data.
Wilkinson, M.D., Dumontier, M., Aalbersberg, Ij 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., et al., 2016. The FAIR guiding principles for scientific data management and stewardship. Sci. Data 3 (1), 160018. https://doi.org/10.1038/sdata.2016.18.
Zhang, C., Zhang, C., Zhang, M., Kweon, I.-S., 2023. Text-to-image diffusion models in Generative AI: a survey. arXiv.org. https://doi.org/10.48550/ARXIV.2303.07909.
Zhang, L., Agrawala, M., 2023. Adding Conditional Control to text-to-image Diffusion Models.