heritage

MDPI

Article

Generalising HARF Beyond Bamiyan: A Bounded Cross-Site Evaluation of a Generative-AI Heritage Reconstruction Framework at the Temple of Bel, Palmyra

Kawsar Arzomand * and Tatiana Kalganova

Department of Engineering, College of Engineering, Design and Physical Sciences, Brunel University, London UB8 3PH, UK; tatiana.kalganova@brunel.ac.uk

* Correspondence: kawsar.arzomand@brunel.ac.uk

Abstract

The Heritage-Aligned Reconstruction Framework (HARF) was developed and initially evaluated using the Western Buddha of Bamiyan to structure evidence-constrained generative-AI heritage reconstruction. This article examines its bounded generalisation to the Temple of Bel at Palmyra, a materially, formally and iconographically contrasting heritage monument. The Dynamic Prompt Blueprint was re-instantiated using documentary evidence concerning the temple's dimensions, limestone construction, Corinthian order, architectural layout, inscriptions and decorative programme, while retaining the framework's top-level structure. A fourteen-item Palmyra questionnaire block was completed by 32 members of the interdisciplinary panel who also completed the Bamiyan evaluation. Within the Palmyra evaluation, Corinthian-column representation received the most favourable rating, whereas restored relief, inscription and surface detail received the least favourable rating. Eighteen respondents (56.2%) selected "none of the iterations are historically accurate" when assessing relief iconography. The paired difference in global prompt-sufficiency ratings did not reach the conventional significance threshold ($p = 0.061$), although ratings were positively associated across the two sites (Spearman $\rho = 0.664$, $p < 0.001$). The seven-domain keyword-coverage profiles were also positively rank-correlated ($\rho = 0.771$, $p = 0.043$). These findings support the bounded re-instantiation of HARF's prompt schema and diagnostic evaluation logic in a contrasting heritage context. They do not establish equivalence between the two implementations, archaeological validity of the generated Temple of Bel images or validation of the complete HARF workflow.

Keywords: cultural heritage reconstruction; generative artificial intelligence; HARF; Temple of Bel; Palmyra; bounded cross-site evaluation; prompt sufficiency; interdisciplinary evaluation; digital heritage

Academic Editors: Min Fan, Aynur Kadir and Özge Nilay Yalçın

Received: 18 May 2026
Revised: 31 July 2026
Accepted: 20 August 2026
Published: 21 August 2026

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 [1,2]. Among the most consequential losses of the twenty-first century are the sixth-to-seventh-century CE Buddhas of Bamiyan, destroyed in March 2001 [3], and the Temple of Bel at Palmyra, intentionally destroyed with explosives in August 2015 [4]. Digital reconstruction methods have been applied to Palmyra through visual reconstruction, photogrammetry, crowdsourced imagery, and immersive three-dimensional

environments [5,6]. Generative-AI reconstruction poses a distinct evaluation problem: its outputs may appear visually coherent while remaining archaeologically unsupported. The methodological issue is therefore not only how to generate plausible images, but how to distinguish visual plausibility from evidential plausibility [7–9].

The Heritage-Aligned Reconstruction Framework (HARF) was introduced to address the evidential organisation and evaluation of generative-AI heritage reconstruction [10]. HARF integrates a Dynamic Prompt Blueprint, a Prompt Sufficiency Index, computational pre-evaluation, and expert-guided assessment within an evidentially constrained workflow. Unlike model-centred approaches concerned principally with fine-tuning, structural conditioning or generative performance, HARF structures the documentary basis of prompt development and enables comparison between computational and human evaluations. A companion empirical study applied HARF to the Western Buddha of Bamiyan and reported a divergence between algorithmic similarity scoring and expert archaeological judgement, with the principal perceived weaknesses concentrated in drapery articulation, cranial morphology, and hand gesture [11]. The Bamiyan evaluation nevertheless leaves one substantive question open. A framework developed and evaluated through a figural Gandhāran monument may require different evidential descriptors and assessment priorities when applied to monumental architecture. That cross-site question has not previously been examined empirically within HARF.

Generative-AI Heritage Reconstruction and the Positioning of HARF

Generative-AI research in cultural heritage has moved beyond unconstrained text-to-image production towards workflows combining documentary evidence, geometric recording and controlled image generation. Photogrammetric models and field photographs have been used as inputs for diffusion-based inpainting and generative fill, allowing damaged or missing architectural elements to be visualised in relation to surviving fabric. Such outputs remain probabilistic representations rather than evidentially supported restitutions unless they are corroborated through archaeological traces, archival documentation and established reconstruction principles [7].

Other approaches have introduced explicit structural and stylistic controls into the generation process. Stable Diffusion has been combined with ControlNet to preserve spatial organisation, LoRA to encode domain-specific visual characteristics, and culturally informed prompt vocabularies to constrain symbolic and stylistic content. The resulting outputs have been assessed through multidimensional expert frameworks covering structural integrity, stylistic fidelity, semantic accuracy, visual coherence and cultural appropriateness. This work demonstrates that generative heritage reconstruction increasingly depends on the combined control of structure, style and cultural meaning rather than on visual plausibility alone [8].

Architectural-heritage workflows have similarly integrated laser scanning, photogrammetry, expert-led feature annotation, semantic filtering, model fine-tuning and combined computational and specialist evaluation. These approaches seek to encode the morphological and stylistic grammar of architectural traditions within controlled generative pipelines. For the present comparison, their relevant contribution lies in domain-specific model adaptation and the quantitative regulation of structural and stylistic fidelity during generation [9].

At the three-dimensional scene level, multimodal frameworks have combined textual and semantic information with geometric constraints, domain adaptation and spatial-pose control. These methods address the conflict between generative variability and reconstruction precision by constraining entity geometry and spatial relationships throughout generation and scene integration [12].

Taken together, these studies show that expert involvement, domain-specific conditioning, geometric constraint and mixed evaluation are established components of generative-AI heritage research. HARF addresses a narrower methodological question: whether a previously specified, evidence-constrained prompt-and-evaluation architecture can be re-instantiated from one destroyed heritage monument to a materially, formally and iconographically contrasting monument. The present study therefore evaluates the portability of HARF's Dynamic Prompt Blueprint, prompt-sufficiency assessment and diagnostic evaluation logic, rather than proposing a new generative model, fine-tuning method or geometrically validated reconstruction pipeline.

This article reports a bounded cross-site evaluation of HARF at the Temple of Bel in Palmyra. The contrast with Bamiyan is deliberate. The Western Buddha is a figural monument carved into a sandstone cliff, finished with mud–straw plaster and governed by Gandhāran iconographic conventions; the Temple of Bel is a Graeco-Roman temple built with locally quarried limestone and governed by classical orders, recorded proportions, inscriptional evidence and a relief-bearing architrave [13,14]. The two cases therefore place different demands on an evidence-constrained reconstruction framework. The present questionnaire-based study examines whether HARF's prompt schema and evaluative logic can be re-instantiated in the Palmyra case; it does not determine whether the generated images constitute archaeologically validated reconstructions of the temple. The study does not adjudicate specialist debates concerning Palmyrene archaeology or post-destruction reconstruction, and Syrian and Palmyrene community perspectives were not examined empirically [4,15,16].

In this article, generalisation is used in a bounded methodological sense. It refers to the re-instantiation of HARF's prompt schema and evaluative logic in a materially, formally and iconographically contrasting heritage case without altering the framework's top-level architecture. It does not denote statistical generalisation across heritage sites or validation of the complete HARF workflow.

This article makes three contributions. First, it provides an empirical test of whether HARF's Dynamic Prompt Blueprint and associated prompt-sufficiency assessment can be re-instantiated from the Western Buddha of Bamiyan to the Temple of Bel while retaining the framework's top-level categories. Secondly, it provides paired cross-site evidence on global prompt sufficiency and domain-level keyword coverage, together with a domain-resolved Palmyra evaluation showing more favourable ratings for Corinthian-column representation than for relief, inscription and surface detail. Thirdly, it distinguishes HARF's function from other generative reconstruction pipelines: its contribution lies in structuring evidential inputs and diagnosing where generated outputs remain weakly supported by the available evidence, rather than in model fine-tuning, geometric reconstruction or output-level archaeological validation.

2. Materials and Methods

2.1. Re-Instantiating the Dynamic Prompt Blueprint for Palmyra

The Dynamic Prompt Blueprint is a tiered slot schema comprising structural anchors, contextual descriptors and technical control parameters [10]. Its top-level categories are intended to remain stable, while the evidential content assigned to each category changes according to the heritage object under examination. For the Temple of Bel, the categories were repopulated using the principal architectural study of the temple [13], the UNESCO World Heritage record [17] and specialist literature concerning the interpretation of its roofline [18].

Bamiyan-specific descriptors were replaced by Temple of Bel-specific evidence concerning limestone construction, recorded cella dimensions, the Corinthian order, the colonnade and podium, interior thalamoi, relief and inscriptional content, environmental context

and uncertainty surrounding roofline treatment. Negative constraints were reformulated to exclude unsupported façades, materials and decorative features. The test therefore concerns whether the same prompt architecture can organise a different evidential corpus; it does not test whether a fixed Bamiyan prompt can be reused or whether the complete HARF workflow has been validated for Palmyra.

Table 1 reports the resulting cross-site instantiation. The top-level categories and tier structure were retained. Paradata-logging conventions are treated as retained only to the extent that the corresponding generation records are available.

Table 1. 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	First century AD; dedicated in 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 diameter
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 constraints	Exclude unattested anatomical, iconographic and decorative features, including anachronistic elements	Exclude Petra-like façades, Roman-imperial conflations and undocumented decorative elements; treat uncertain roofline features as interpretive
Technical	Paradata fields	Platform and model version; prompt record; available generation settings; refinement rationale; seed where exposed	The same fields, recorded where available; completeness varied according to the platform and retained generation records

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 [6,19]. 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 [20,21].

The Palmyra instrument therefore examines whether HARF's prompt architecture and questionnaire-based evaluation can identify the domains in which generative outputs remain weakly supported by the available evidence; it does not establish final reconstruction validity.

A fourteen-item Palmyra block was embedded within the Phase I questionnaire used for the Bamiyan study [10,11] and administered through Google Forms (Google LLC, Mountain View, CA, USA). The block followed a structured monument briefing presenting the temple's principal dimensions, architectural form, material characteristics and cultural significance, so that respondents could assess the images against a defined evidentiary baseline rather than against prior personal knowledge alone [17]. Six items elicited ordered five-category ratings of structural fidelity, the proportional accuracy of the podium, staircase and colonnade, Corinthian-column representation, the fidelity of the remaining columns, walls and roofline against the photographic record, the realism of restored relief, inscription and texture detail, and overall keyword sufficiency. Accuracy items were coded from 5 ("Highly Accurate") to 1 ("Not Accurate"), while prompt sufficiency was coded from 5 ("Highly Sufficient") to 1 ("Insufficient").

Five multi-select items elicited the architectural features judged missing or inaccurate, the candidate images regarded as most accurate, the aspects perceived to improve across iterations, the elements appearing most divergent from the documented ruins, and the descriptive domains requiring additional keyword coverage. Two single-choice items recorded the iteration judged most accurate for relief iconography and the iteration considered most convincing in its representation of the symbolic objects held by figures within the relief. One open-ended item invited suggestions for prompt refinement.

Question 30 permitted multiple candidate selections and therefore produced selection frequencies rather than an exclusive candidate preference. The 32 respondents returned 54 selections in total. The candidates were also displayed at unequal sizes, with Image 4 occupying substantially greater display area than the remaining candidates. Question 30 is therefore reported descriptively and is not used to support the study's substantive conclusions. Respondents evaluated the candidates against the image plates reproduced in Figures 1–3.

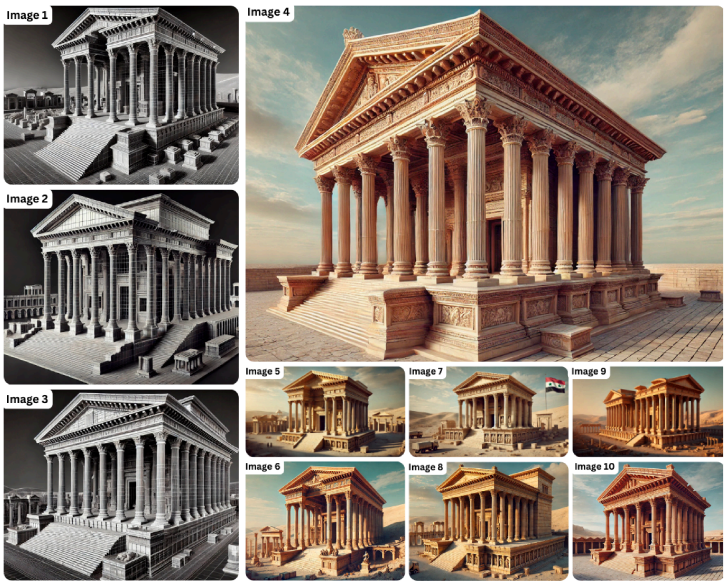

Figure 1. Architectural and structural-accuracy stimuli: the ten AI-generated Temple of Bel reconstruction candidates presented to the panel. The images were evaluation stimuli and were not presented as final reconstructions.

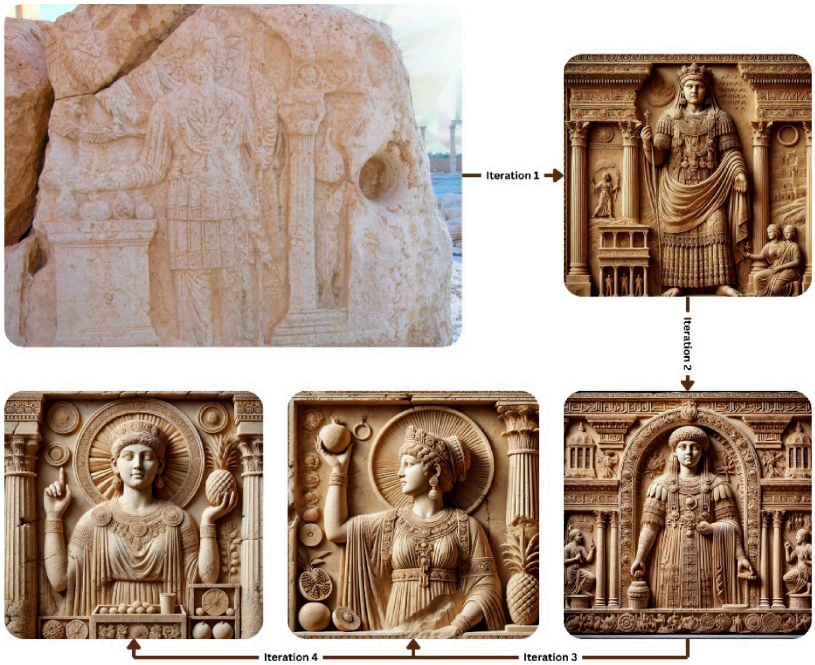

Figure 2. Iterative-refinement stimuli for Temple of Bel relief iconography. Top left: reference photograph of a Palmyra relief (Francesco Bandarin/UNESCO, 2007); remaining panels: AI-generated iterations used as evaluation stimuli.

Figure 3. 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.

2.3. Global Prompt Sufficiency and Domain-Level Keyword Coverage

Prompt adequacy was examined through two questionnaire-derived measures elicited for both sites within the same instrument. First, conceptually parallel global items asked

respondents to rate the sufficiency of the Bamiyan and Palmyra keyword sets using five ordered response categories. Secondly, multi-select items asked which of seven descriptive domains required additional keyword coverage.

The domain-level responses are analysed as keyword-coverage profiles. They are not presented as a recalculation of the construction-stage weighted Prompt Sufficiency Index described in the framework article, because the present instrument did not collect the domain-specific importance weights and sufficiency ratings required by that index [10].

For each domain i, the coverage value was calculated as

$$C_i = 1 - p_i$$

where p_i is the proportion of paired respondents indicating that domain i required additional keyword coverage. Values are reported separately for each domain and as an unweighted mean across the seven domains, calculated identically for both sites.

The Bamiyan global sufficiency item was configured as a single-choice question. The corresponding Palmyra item was configured as a checkbox question, although all 32 respondents selected only one response category. The Palmyra responses were therefore treated as ordinal for exploratory paired analysis. The difference in response control is retained as a methodological limitation.

2.4. Panel, Recruitment and Analytical Strategy

The Palmyra block was administered to the Phase I interdisciplinary panel recruited through academic networks, professional associations and institutional channels, as described in the framework study [10]. Thirty-three respondents completed the Bamiyan section and 32 completed the Palmyra block. Respondents represented 12 countries and professional backgrounds spanning digital heritage, artificial intelligence and machine learning, heritage conservation, cultural and heritage studies, archaeology, architecture, and construction and engineering. Four respondents identified archaeology as their primary professional background. Accordingly, the findings are interpreted as interdisciplinary panel assessments rather than specialist archaeological validation. Familiarity with the Temple of Bel was elicited through a four-category item. Of the 32 Palmyra completers, 10 selected "no familiarity", 10 "slight familiarity", 6 "moderate familiarity" and 6 "high familiarity".

Throughout this paper, 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 and paired cross-site analyses. 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.

Ordinal responses are reported primarily through medians and interquartile ranges, with means and standard deviations retained secondarily for comparability with the framework study. Categorical proportions are reported with Wilson 95% confidence intervals [22]. Differences across the five accuracy items were tested using the Friedman test with Kendall's W, followed by pairwise Wilcoxon signed-rank tests with matched-pairs rank-biserial correlation as the effect size. One-sample Wilcoxon signed-rank tests were used to assess whether each accuracy-item distribution differed from the scale midpoint of 3. Paired cross-site differences were examined using the Wilcoxon signed-rank test, and associations between cross-site responses were examined using Spearman's rank correlation. Familiarity-stratified comparisons used Kruskal–Wallis tests with epsilon-squared, supplemented by exploratory Mann–Whitney comparisons of low-familiarity and moderate-to-high-familiarity respondents. All families of multiple comparisons were adjusted using the Benjamini–Hochberg procedure [23]. Because the two upper familiarity strata contained six respondents each, the subgroup analyses were underpowered, and non-

significant findings were not interpreted as evidence of equivalence or the absence of an effect. Statistical analyses were conducted using Python 3.13 (Python Software Foundation, Wilmington, DE, USA), with the pandas, NumPy and SciPy libraries.

2.5. Scope of Quantitative Analysis

Quantitative analysis in this study concerns ordered differences among panel ratings, paired cross-site prompt-sufficiency judgements, associations between cross-site responses and familiarity-stratified sensitivity. It does not constitute output-level geometric validation. The computational pre-evaluation pipeline used in the framework study—structural similarity, ORB feature matching with RANSAC homography estimation and proportional tolerance screening—was not repeated for Palmyra [10,11].

A defensible post hoc replication of that pipeline was not possible using the present materials. The surviving photographs and generated candidates were not produced as viewpoint-matched pairs; similarity and feature-matching measures across unmatched viewpoints, lighting conditions and ruin states would primarily reflect composition and framing. The questionnaire also did not produce a valid per-candidate expert score against which a computational ranking could be evaluated. Question 30 was multi-select and was additionally affected by unequal display size. Claims in this article are therefore restricted to the analyses supported by the questionnaire design.

3. Results

3.1. Panel Ratings of the Temple of Bel Reconstructions

Corinthian-column representation received the highest rating (median 3.5, IQR 3–4), with 16 of 32 respondents rating it "Mostly" or "Highly Accurate". Restored relief, inscription and surface detail received the lowest rating (median 2.0, IQR 2–3), with 6 of 32 respondents rating it "Mostly" or "Highly Accurate". The full distribution of panel ratings across the five assessed dimensions is reported in Table 2.

Table 2. Interdisciplinary-panel ratings of the AI-generated Temple of Bel images (n = 32). Values represent panel-rated perceived fidelity rather than measurements of archaeological accuracy. Square brackets report Wilson 95% confidence intervals for the proportion rating each item 4 or 5.

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

Ratings differed across the five accuracy items (Friedman $\chi^2 = 36.46$, df = 4, $p < 0.001$, Kendall's W = 0.285). Mean ranks ordered the items as Corinthian-column representation (3.64), overall structure (3.39), proportions (3.22), remaining columns, walls and roofline (2.78), and restored relief, inscription and surface detail (1.97).

All four contrasts involving restored relief, inscription and surface detail remained significant after Benjamini–Hochberg correction, with matched-pairs rank-biserial correlations

ranging from 0.686 to 1.00. The largest contrast was with Corinthian-column representation: every respondent who differentiated between the two rated the columns more favourably (W = 0, adjusted $p < 0.001$, rank-biserial r = 1.00). Corinthian-column representation was also rated above the remaining columns, walls and roofline item (W = 24.0, adjusted $p = 0.014$, r = 0.719).

After Benjamini–Hochberg correction, none of the five accuracy-item distributions differed significantly from the scale midpoint of 3 (all adjusted $p \geq 0.054$). The analysis therefore supports a relative ordering of the assessed domains rather than a claim that any individual domain was rated above or below moderate accuracy.

Respondents most frequently identified column placement and height (40.6%), interior thalamoi (37.5%), roofline merlons (34.4%), ceiling zodiac reliefs (31.2%) and architrave relief carving (28.1%) as missing or inaccurate features.

The candidate-selection item permitted multiple selections. Image 4 was included among the selections of 16 respondents, with 54 selections returned across all ten candidates. Because Image 4 was displayed at substantially greater size than the remaining candidates, its selection frequency cannot be separated from visual salience and is not used analytically.

3.2. Iterative Refinement

Iterative refinement did not resolve the panel's concerns regarding relief iconography. As shown in Figure 4, eighteen of the 32 respondents (56.2%; 95% CI 39.3–71.8%) selected "none of the iterations are historically accurate" when evaluating the relief images against the supplied reference material. Responses for the remaining four iterations were dispersed, with no individual iteration selected by more than five respondents. These findings indicate that successive refinement did not produce a clear panel preference for an iconographically credible reconstruction.

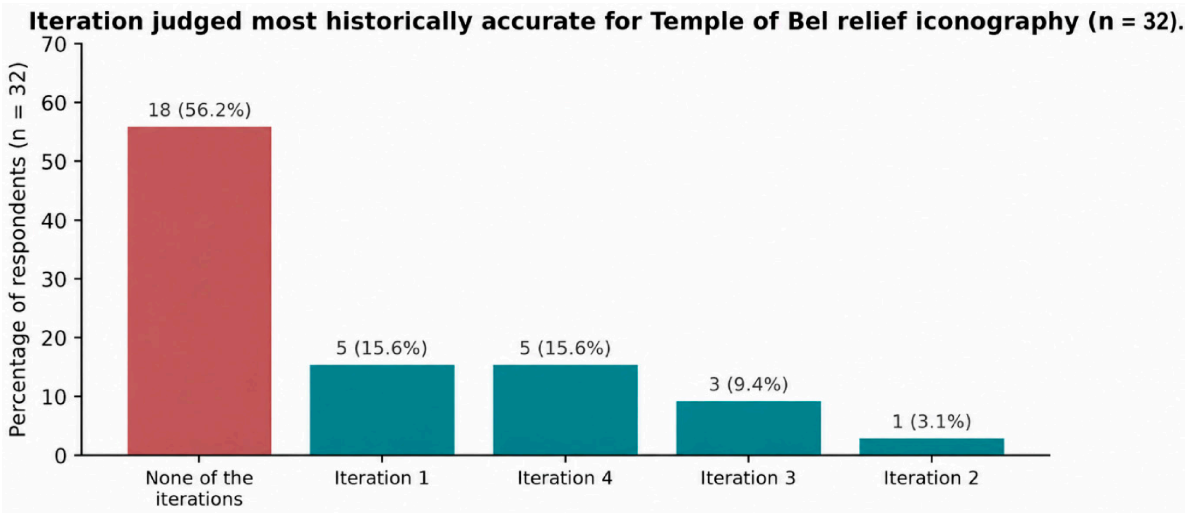

Figure 4. Iteration judged most historically accurate for Temple of Bel relief iconography. The modal response was "none of the iterations are historically accurate" (56.2%; 95% CI 39.3–71.8%).

When asked which iteration most convincingly represented the symbolic objects held by figures within the relief, respondents were divided equally between "none" and "Iteration 4" (10/32 each; 31.2%). Respondents identified perceived improvements in background architectural elements (46.9%) and decorative jewellery (43.8%) more frequently than in inscriptions and iconography (34.4%) or facial expressions and proportions (31.2%). The pattern suggests that perceived visual refinement was more evident in general architectural and decorative features than in historically and iconographically specific content.

3.3. Feature-Level Contrast Against the Surviving Record

Ratings on the items referring to the surviving ruin and decorative detail were lower than those on the preceding macro-architectural items. The remaining columns, walls and roofline item returned a mean of 2.88, while restored relief, inscription and surface detail returned a mean of 2.47, compared with means of 3.28, 3.22 and 3.47 for overall structure, proportions and Corinthian-column representation, respectively.

The two groups of items differ both in evidentiary reference and in the feature being assessed. The contrast therefore cannot isolate the independent effect of comparison with photographs, and no causal interpretation is made. After Benjamini–Hochberg correction, Corinthian-column representation was rated above the remaining columns, walls and roofline item, whereas the comparisons between that item and overall structure or proportions did not reach significance. Restored relief, inscription and surface detail was rated below each of the other four accuracy items.

As shown in Figure 5, respondents most frequently identified relief sculptures and wall carvings as divergent from the supplied photographic record (62.5%; 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 (31.2%). The supported finding is therefore narrower: panel-rated representation of the standing columnar order was more favourable than panel-rated restoration of relief, inscription and surface detail.

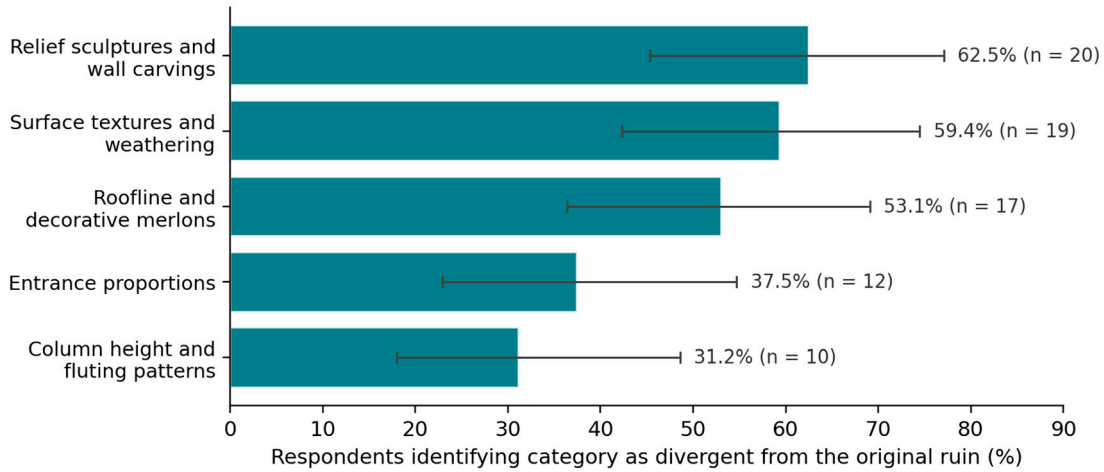

Figure 5. Features identified as divergent from the supplied photographic record of the Temple of Bel. Bars report the percentage of respondents selecting each feature; error bars denote Wilson 95% confidence intervals.

3.4. Prompt Sufficiency and Cross-Site Keyword Coverage

The global prompt-sufficiency rating for the Temple of Bel returned a median of 3.5 (IQR 3–4; mean 3.41, SD 1.10). Sixteen of the 32 respondents selected one of the two highest response categories, five selected the highest category and two selected the lowest category. Among the same 32 respondents, the corresponding Bamiyan item returned a median of 4.0 (IQR 3–4; mean 3.69, SD 0.97), with 20 respondents selecting one of the two highest categories.

The paired difference did not reach the conventional significance threshold (Wilcoxon $W = 39.5$, $p = 0.061$, rank-biserial $r = 0.484$): 13 respondents rated Bamiyan higher, 4 rated Palmyra higher and 15 gave identical ratings. The result does not establish equivalence, and the response distribution directionally favoured Bamiyan. Global prompt-sufficiency ratings were positively associated across the two sites (Spearman $\rho = 0.664$, $p < 0.001$).

Domain-level keyword coverage is reported in Table 3. The two seven-domain profiles were positively rank-correlated (Spearman ρ = 0.771, $p = 0.043$, k = 7). Environmental context was the least sufficiently specified domain in both cases, followed by architectural dimensions and layout. Unweighted mean coverage was 0.567 for Bamiyan and 0.652 for Palmyra. Cultural symbolism was the only domain with a descriptively lower Palmyra value, although the difference was small. Because this analysis rests on seven aggregated domains, it is treated as exploratory evidence of cross-case similarity rather than proof of framework transfer.

Table 3. Domain-level keyword-coverage profiles for the 32 respondents who completed both questionnaire sections. The *n*/*N* columns report respondents indicating that additional keyword coverage was required relative to the number of valid respondents. Coverage was calculated as one minus that proportion; higher values indicate greater perceived sufficiency.

Domain	Bamiyan *n*/*N*	Palmyra *n*/*N*	Bamiyan Coverage	Palmyra Coverage
Environmental context	19/32	17/32	0.406	0.469
Architectural dimensions and layout	16/32	15/32	0.500	0.531
Artistic detail	14/32	12/32	0.563	0.625
Material detail	14/32	10/32	0.563	0.688
Historical and period detail	12/32	7/32	0.625	0.781
Cultural symbolism	11/32	12/32	0.656	0.625
Lighting	11/32	5/32	0.656	0.844
Unweighted mean	-	-	0.567	0.652

Nine respondents supplied substantive prompt-refinement comments. These concerned the cella footprint, replacement of evocative language with concrete descriptors, weathering and patina, stepped merlons and the organisation of keywords across refinement stages.

3.5. Sensitivity to Site Familiarity

Ratings were disaggregated across the four Temple of Bel familiarity strata: no familiarity ($n = 10$), slight familiarity ($n = 10$), moderate familiarity ($n = 6$) and high familiarity ($n = 6$). No statistically detectable differences were observed across the four strata (Kruskal–Wallis H values 0.56–2.37; all Benjamini–Hochberg-adjusted $p = 0.907$). The corresponding epsilon-squared values were approximately zero. No statistically detectable differences were observed between the 20 low-familiarity and 12 moderate-to-high-familiarity respondents (all adjusted $p \geq 0.826$).

These non-significant findings do not establish that familiarity had no effect. The two upper strata contained only six respondents each and the analysis was underpowered. Descriptively, the moderate-to-high-familiarity group rated relief, inscription and surface detail lower than the low-familiarity group (means 2.08 and 2.70, respectively; rank-biserial r = −0.31), while rating keyword sufficiency somewhat higher (means 3.67 and 3.25, respectively; r = 0.18).

The principal ordering was also present within the moderate-to-high-familiarity subgroup: 7 of 12 respondents rated Corinthian-column representation 4 or 5, compared with 1 of 12 for restored relief, inscription and surface detail. This subgroup finding is exploratory.

4. Discussion

The Palmyra evaluation supports a bounded interpretation of the cross-site portability of selected HARF components. The principal finding is not that the Temple of Bel outputs were historically accurate; they were not consistently judged as such. Rather, the DPB's top-level structure could be re-instantiated for a monumental Graeco-Roman architectural case, and the adapted questionnaire enabled the interdisciplinary panel to identify domains of perceived evidential weakness. Corinthian-column representation was assessed more favourably than relief, inscription and surface 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 contrasting heritage case and identify domains in which generated outputs remain weakly supported.

Recent generative-heritage approaches address complementary but distinct methodological problems. Diffusion-based reconstruction and model-adaptation workflows seek to improve visual, structural or stylistic fidelity through inpainting, geometric conditioning, fine-tuning and expert feature encoding [7–9], while multimodal scene-reconstruction methods constrain spatial relationships at the three-dimensional level [12]. HARF does not replace these approaches or claim superior reconstruction performance. Its narrower contribution is to organise documentary evidence and examine whether a predefined prompt-and-evaluation architecture remains diagnostically usable when the heritage case changes.

The cross-site prompt analyses also qualify this conclusion. The paired global prompt-sufficiency difference did not reach the conventional significance threshold and directionally favoured Bamiyan; equivalence between the two implementations therefore cannot be inferred. However, the positive association between respondents' global ratings and the exploratory correlation between the seven-domain coverage profiles indicate that the same evaluative structure elicited partly comparable response patterns across the two cases. These findings support bounded portability rather than identical framework performance.

This finding is consistent with a provisional cross-case instability pattern. At Bamiyan, this instability concentrated in figural micro-grammar: drapery articulation, cranial morphology, and hand gesture [11]. At Palmyra, the least favourable assessments concerned relief, inscription and surface detail, while roofline treatment and interior cult-architectural features were also identified as areas of concern. The top-level prompt architecture was retained, but the distribution of perceived weakness differed. This matters because framework transferability does not imply output reliability. Framework portability should therefore be assessed partly by whether the same evaluative structure can identify case-specific domains of weakness, rather than by visual coherence alone. Table 4 summarises this bounded cross-site comparison.

The resulting portability claim is restricted to the HARF components re-instantiated in the Palmyra case. The DPB schema was repopulated with Palmyra-specific architectural, decorative, material and inscriptional evidence. The questionnaire-based evaluation was adapted to the Temple of Bel. The construction-stage PSI 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 complete paradata record were not repeated at Palmyra. The Palmyra case therefore examines schema-level re-instantiation and questionnaire-based evaluation, not full replication of the Bamiyan workflow.

Table 4. Three-layer model of bounded HARF portability across Bamiyan and Palmyra.

Layer	Role in the Evaluation	Bamiyan	Palmyra	Implication
Stable framework layer	Top-level prompt structure and questionnaire-based evaluation	Original implementation	Re-instantiated implementation	Selected components can be retained across cases
Site-specific evidence layer	Material, geometry, iconography and environmental context	Gandhāran figural monument, mud–straw plaster, drapery and mudrā	Limestone temple, Corinthian colonnade, thalamoi and inscriptions	Evidential content must be specified for each case
Case-specific instability layer	Features receiving the least favourable assessments	Drapery, cranial morphology and hand gesture	Relief, surface detail, inscription and roofline	Instability domains must be identified and evaluated separately

This result has a practical implication for future applications of HARF. Macro-structural recognisability should not be treated as sufficient evidence of evidential or archaeological 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 evidentially sensitive features would need to be defined from the documentary record before evaluation begins.

This comparison 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, selected elements of HARF's evaluative architecture can be retained while identifying a different pattern of perceived weakness. 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 present case indicates that the top-level schema can be retained, while its descriptors and evidentially sensitive domains must be specified from the documentary record of each site.

The rating pattern is also consistent with the possibility that descriptor granularity influenced the evaluation outcomes. Structural and columnar slots incorporated recorded metric values, whereas inscriptional and decorative slots relied more heavily on categorical descriptors. The present design cannot separate descriptor granularity from feature complexity or documentary availability. Nevertheless, subsequent applications should record the evidential resolution of each prompt slot rather than merely whether that slot has been populated.

The Palmyra results also qualify the role of iterative refinement. The iteration sequence did not produce a clear panel preference for an iconographically credible relief reconstruction, 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 recent

generative-heritage research distinguishing visual coherence from structural, semantic and cultural fidelity [8,9].

The study therefore supports bounded framework-level re-instantiation rather than complete validation. The findings show that selected elements of HARF's schema and evaluative logic can be re-instantiated across one substantial contrast. Further Palmyrene-specialist assessment, community-grounded evaluation and appropriately designed computational comparison would be required to examine those dimensions. 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. Selected elements of HARF were portable at the level of prompt architecture and questionnaire-based evaluation; historical plausibility remained dependent on site-specific evidence, specialist judgement, and, where relevant, community interpretation.

5. Limitations

Several limitations qualify these findings. First, the panel was interdisciplinary rather than composed of specialists in Palmyrene archaeology. Twenty of the thirty-two Palmyra completers selected either "no familiarity" or "slight familiarity" with the Temple of Bel. The structured monument briefing provided a shared evidentiary baseline, but it could not substitute for independent specialist knowledge. For those respondents, the ratings therefore partly reflect correspondence between the generated images and the supplied briefing and image materials.

The questionnaire design introduces further constraints. The global prompt-sufficiency items differed in response control: the Bamiyan item was single-choice, whereas the Palmyra item was configured as a checkbox, although all respondents selected only one category. The candidate-selection item was multi-select and was affected by unequal image display sizes; it therefore cannot support an exclusive candidate ranking. Comparisons between the macro-architectural items and the ruin- or detail-oriented items are also confounded by differences in both evidentiary reference and feature content.

No Palmyra-specific specialist evaluation was undertaken. The level of site-specific archaeological scrutiny provided by the specialist phase of the Bamiyan evaluation was therefore not reproduced here [11]. No Syrian or Palmyrene community-evaluation component was undertaken, comparable to the community-grounded evaluation conducted for Bamiyan [24]. 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 [15]. No Palmyra-specific computational similarity pipeline was conducted; the cross-site comparison is therefore restricted to questionnaire-based evaluation and does not establish output-level geometric correspondence. The study also compares only one monument from each of two broad heritage categories, limiting inference beyond the two cases examined. Finally, reproducibility constraints arising from the volatility of generative platforms apply equally here. Available generation records partially address this constraint, but incomplete or platform-dependent paradata limit exact reproducibility [25].

6. Conclusions

A bounded cross-site evaluation indicates that selected elements of HARF's prompt architecture and evaluative logic can be re-instantiated from a figural Gandhāran monument to a monumental Graeco-Roman architectural case while retaining the framework's top-level categories. The re-instantiation was evaluative rather than reconstructive: the Palmyra outputs were not consistently judged historically accurate. The distribution of perceived evidential weakness differed between the two cases: concerns in the Bamiyan

evaluation concentrated on drapery, cranial morphology and hand gesture, whereas the Palmyra evaluation identified relief, inscription and surface detail as the least favourably assessed domains, with roofline and interior cult-architectural features also identified as areas of concern. This difference supports a provisional cross-case instability pattern rather than a general typology-dependent principle.

The broader implication is that frameworks such as HARF are best understood as diagnostic instruments rather than reconstruction engines. Their value lies in structuring evidential inputs, recording prompt logic, eliciting prompt-sufficiency judgements and identifying domains in which generated outputs remain weakly supported. They do not remove the need for specialist judgement, nor do they convert visual plausibility into historical plausibility. Palmyrene-specialist assessment, Syrian or Palmyrene community evaluation and an appropriately designed computational comparison would be required before broader claims concerning archaeological validity, cultural legitimacy or output-level correspondence could be made.

Author Contributions: Conceptualisation, K.A.; methodology, K.A.; investigation, K.A.; formal analysis, K.A.; data curation, K.A.; visualisation, K.A.; project administration, K.A.; writing—original draft preparation, K.A.; writing—review and editing, T.K.; supervision, T.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the British Council Warm Welcome Scholarships scheme, administered by Brunel University London.

Institutional Review Board Statement: This study was approved by the Brunel University of London Research Ethics Committee under reference 50187-LR-Jan/2025-53624-2.

Informed Consent Statement: Informed consent was obtained from all participants involved in the study.

Data Availability Statement: The aggregated data supporting the findings of this study are reported in the article. Raw respondent-level data are not publicly available because of ethical restrictions, respondent confidentiality, and the risk of deductive identification within a small expert panel. Further details may be made available from the corresponding author upon reasonable request, subject to ethical approval and data-protection requirements.

Acknowledgments: The authors thank the Phase I expert respondents for their participation in the study.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial intelligence
CI	Confidence interval
DPB	Dynamic Prompt Blueprint
GenAI	Generative artificial intelligence
HARF	Heritage-Aligned Reconstruction Framework
IQR	Interquartile range
ORB	Oriented FAST and Rotated BRIEF
PSI	Prompt Sufficiency Index
RANSAC	Random Sample Consensus
SD	Standard deviation
SSIM	Structural Similarity Index Measure
UNESCO	United Nations Educational, Scientific and Cultural Organization

References

1. Brosché, J.; Legnér, M.; Kreutz, J.; Ijla, A. Heritage under attack: Motives for targeting cultural property during armed conflict. *Int. J. Herit. Stud.* **2017**, *23*, 248–260. [CrossRef]
2. Cunliffe, E.; Muhesen, N.; Lostal, M. The Destruction of Cultural Property in the Syrian Conflict: Legal Implications and Obligations. *Int. J. Cult. Prop.* **2016**, *23*, 1–31. [CrossRef]
3. Flood, F.B. Between Cult and Culture: Bamiyan, Islamic Iconoclasm, and the Museum. *Art. Bull.* **2002**, *84*, 641–659.
4. Schmidt-Colinet, A. No temple in Palmyra! Opposing the reconstruction of the Temple of Bel. *Syr. Stud.* **2019**, *11*, 63–85.
5. Silver, M.; Fangi, G.; Denker, A. *Reviving Palmyra in Multiple Dimensions: Images, Ruins and Cultural Memory*; Whittles Publishing Ltd.: Scotland, UK, 2018.
6. Rihani, N. Interactive Immersive Experience: Digital Technologies for Reconstruction and Experiencing Temple of Bel Using Crowdsourced Images and 3D Photogrammetric Processes. *Int. J. Archit. Comput.* **2023**, *21*, 730–756. [CrossRef]
7. Kutlu, İ.; Şimşek, D. Artificial intelligence (AI) assisted digital reconstruction of historical buildings. In *Proceedings of the 4th International Civil Engineering & Architecture Conference Volume 2: Architecture, Seoul, South Korea, 15–17 March 2024*; Springer: Singapore, 2025. [CrossRef]
8. Liu, Y.; Li, Y.; Li, Q.; Wang, S. Application and Evaluation of Stable Diffusion-Based Generative AI in the Digital Reconstruction of Cultural Heritage Patterns. *J. Comput. Cult. Herit.* **2026**. [CrossRef]
9. Yang, Q.; Li, E.; Zhang, Y. Generative AI for architectural heritage conservation: Expert-guided feature encoding, stylistic regeneration and typical application. *Int. J. Digit. Earth* **2026**, *19*, 2681387. [CrossRef]
10. Arzomand, K.; Kalganova, T.; Rustell, M. HARF: A human–AI collaborative framework for cultural heritage reconstruction with expert-guided multi-platform generative AI and systematic prompt engineering. *Digit. Appl. Archaeol. Cult. Herit.* **2026**, *42*, e00554. [CrossRef]
11. Arzomand, K.; Kalganova, T. Evaluating Generative AI Reconstructions of the Bamiyan Buddhas: Computational Similarity and Expert Archaeological Assessment. *Zenodo* **2026**, *Preprint*. [CrossRef]
12. Dang, P.; Zhu, J.; Rao, Y.; Li, W.; Dang, C. A multimodal generative AI-driven 3D geographic scene reconstruction method. *Int. J. Geogr. Inf. Sci.* **2026**, *40*, 1431–1455. [CrossRef]
13. Seyrig, H.; Amy, R.; Will, E. Le temple de Bel à Palmyre. Vol. I: Texte et planches; Vol. II: Album. In *Revue Archéologique*; (Number 1); Librairie Orientaliste Paul Geuthner: Paris, France, 1975. Available online: https://www.google.co.uk/books/edition/Le_Temple_de_B%C3%AAl_a_Palmyre/0JtbtgEACAAJ?hl=en (accessed on 20 July 2026).
14. Arzomand, K.; Rustell, M.; Kalganova, T. From Ruins to Reconstruction: Harnessing Text-to-Image AI for Restoring Historical Architectures. *Chall. J. Struct. Mech.* **2024**, *10*, 69. [CrossRef]
15. Abdulkarim, M.; Seigne, J. The Future of the Temple of Bel in Palmyra after Its Destruction. *Bull. Am. Soc. Overseas Res.* **2024**, *391*, 93–106. [CrossRef] [PubMed]
16. Munawar, N.A. Reconstructing Cultural Heritage in Conflict Zones: Should Palmyra be Rebuilt? *Ex. Novo J. Archaeol.* **2017**, *2*, 33–48. [CrossRef]
17. UNESCO. *Site of Palmyra*; UNESCO World Heritage Centre: Paris, France, 2026. Available online: https://whc.unesco.org/en/list/23 (accessed on 20 July 2026).
18. Schmidt-Colinet, A.; Seigne, J. The battlements of temples in Hellenistic-Roman Syria. The crenellations of the Sanctuary of Bel in Palmyra reconsidered. *Rev. Archéologique* **2022**, *73*, 57–77. [CrossRef]
19. Wahbeh, W.; Nebiker, S.; Fangi, G. Combining Public Domain and Professional Panoramic Imagery for the Accurate and Dense 3d Reconstruction of the Destroyed Bel Temple in Palmyra. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2016**, *III–5*, 81–88. [CrossRef]
20. Denard, H. *The London Charter for the Computer-Based Visualisation of Cultural Heritage*; King's College London: London, UK, 2009. Available online: https://www.londoncharter.org/ (accessed on 20 July 2026).
21. López-Menchero Bendicho, V.M.; Grande, A. The Seville Principles: International Principles of Virtual Archaeology (2011). *International Forum of Virtual Archaeology (IFVA)*. 2011. Available online: https://smartheritage.com/seville-principles/ (accessed on 20 July 2026).
22. Wilson, E.B. Probable Inference, the Law of Succession, and Statistical Inference. *J. Am. Stat. Assoc.* **1927**, *22*, 209–212. [CrossRef]
23. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1995**, *57*, 289–300. [CrossRef]
24. Arzomand, K.; Kalganova, T. Community-Grounded Authenticity at Bamiyan: Afghan Perspectives on AI-mediated Heritage Reconstruction. *Zenodo* **2026**, *Preprint*. [CrossRef]
25. Huvila, I. Improving the usefulness of research data with better paradata. *Open Inf. Sci.* **2022**, *6*, 28–48. [CrossRef]