Prompting the Past: Linguistic Transformations and Cultural Accuracy in AI-Generated Image Reconstructions for Multivocal Cultural Heritage

Ravini Wimalasuriya[1] **Lea Krause**[2] **Gert-Jan Burgers**[1]
[1]Faculty of Humanities [2]Faculty of Science
Vrije Universiteit Amsterdam, The Netherlands
raviniwim@gmail.com {l.krause, g.l.m.burgers}@vu.nl

Abstract

This research explores the intersection of cultural heritage and Generative AI (Gen-AI), examining AI-generated historical image reconstructions as a potential tool for visualising multiple perspectives in heritage interpretation. In critical heritage studies, the concept of multivocality or polyvocality advocates for representing diverse, often underrepresented, perspectives in how heritage is understood and communicated. We evaluated three prominent AI image generation models across three heritage test cases. A total of 13 user prompts generated 39 images, which underwent both linguistic analysis of intermediate prompt transformations and systematic visual assessment by heritage experts for historical accuracy and cultural sensitivity. The findings revealed both strengths and limitations of the models. While the models produced visually compelling outputs and, in some cases, meaningfully distinct depictions across perspectives, they also exhibited representation imbalances, neutralisation and amplification tendencies, inconsistencies in human portrayal, and misinterpretations introduced during the linguistic transformation of user inputs. Based on these findings, we propose initial guidelines for structured prompt construction that target the specific failure patterns identified. The research suggests that generative AI could serve as a supplementary tool, not a definitive historical source, for exploring multivocal heritage interpretation, particularly in museum and visitor engagement contexts, provided it is used critically and in conjunction with expert input.

1 Introduction

The emergence of critical heritage studies in the late 20th century challenged traditional heritage discourses that predominantly reflected state and institutional authority. This shift opened pathways for focusing on the diversity lens of the factual past, the diverse perspectives, which were often

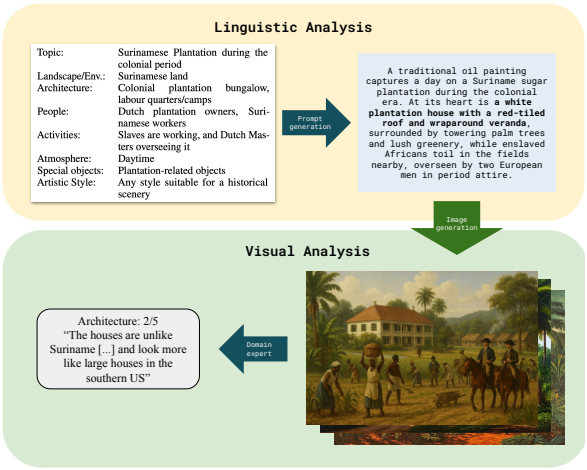

Figure 1: Overview of the research workflow. User-provided keywords across eight cultural criteria (left) are transformed by DALL-E 3 into an expanded model-generated prompt (top right), which is then used to generate an image (right). The resulting image undergoes two parallel analyses: a linguistic analysis examining how the intermediate prompt transformation altered the user's input, and a visual analysis in which a heritage domain expert rates the image's historical accuracy across the same eight criteria on a 1–5 scale. The example shown depicts the Dutch visitor's perspective for the Surinamese plantation test case.

marginalised or silenced due to power dynamics, and the contemporary socio-economic forces that shaped the appropriation of heritage and its interpretation. This context supported the emergence of the concept 'Multivocality' or 'Polyvocality', allowing previously oppressed or overlooked voices to be heard in heritage interpretation.

The existence of multiple voices for heritage sites, objects, and histories is often shaped by nationality, ethnicity, class, gender, education, culture and other forms of human identity (Derrida, 1994; Kojan, 2008; Deumert, 2018; Smith, 2020). Accordingly, when knowledge is co-created with stakeholders and communities, rather than solely with heritage experts, it fosters dialogue and un-

Proceedings of the 6th International Conference on Natural Language Processing for the Digital Humanities, pages 83–95

derstanding among those with different relationships to the past (Franke et al., 2024). Moreover, truthful heritage interpretation requires critically acknowledging these diverse and at times conflicting perspectives. Hence, heritage professionals and scholars increasingly explore innovative methods and tools to pursue more inclusive, diverse, and reflective approaches to facilitate multivocality.

This research investigates the potential of generative AI to produce historical image reconstructions that represent multiple cultural perspectives in heritage contexts (see Figure 1). The study makes three contributions: *(1)* It provides a linguistic analysis of the intermediate prompt transformation process, i.e. how AI models rewrite user-provided inputs before generating images, and identifies how these transformations affect culturally specific content. *(2)* It presents a systematic, expert-assessed comparison of three leading image generation models across three historically complex and contested heritage test cases. *(3)* Drawing on the findings, it proposes practical guidelines for structured prompt construction aimed at improving the historical accuracy and cultural sensitivity of AI-generated heritage imagery.

2 Literature Review and Related Work

2.1 Multivocality / Polyvocality

Heritage institutions are increasingly incorporating previously silenced or marginalised voices into historical narratives. Examples include London's Migration Museum, The Hague's Humanity House (refugee stories), and the Smithsonian's Slavery and Freedom exhibition. Common strategies include participatory or co-curative exhibitions, oral history initiatives, and collaborative research with underrepresented communities.

However, implementing multivocality is complex. Smith (2020) notes that even well-intentioned multivocal projects can yield mixed outcomes, as parties must accept coexisting, at times contradictory truths. Zheng (2023) highlights that amplifying marginalised voices remains harder in practice than in theory, partly because audiences are conditioned to accept singular narratives (Atalay, 2008; Barnabas, 2016). Authorised Heritage Discourse (AHD), highlighted by Smith (2006), further points out the resistance to these alternative viewpoints.

However, at the same time, scholars also underline the need to critically examine multivocality, as not all voices carry equal validity. While welcoming diverse perspectives, scholars advocate for the critical evaluation of each voice to ensure truthful heritage interpretation and to prevent misuse, whether for malicious intentions or unintended misinterpretations (Atalay, 2008; Kojan, 2008; Wylie, 2008).

2.2 Natural Language Processing

Research in NLP and computer vision increasingly examines how textual prompts shape the outputs of text-to-image (T2I) models. Oppenlaender (2024) established that prompt structure significantly affects generation quality. More broadly, using one language model to rewrite or optimise prompts for another model, or for a downstream generation model, has become common practice in NLP pipelines, from automated prompt engineering (Hao et al., 2023) to chain-of-thought prompt rewriting for improved image-text alignment (Wang et al., 2025). DALL-E 3 implements this approach by default: user prompts are automatically transformed before image generation. Our study examines this intermediate transformation as a site of potential meaning change, analysing the linguistic operations through which culturally specific user inputs are reshaped, an aspect that, to our knowledge, has not been systematically studied for heritage-related content.

Separately, a growing body of work addresses bias in T2I models. Wan et al. (2024) survey bias across gender, skin tone, and geo-cultural dimensions, finding the latter remains under-explored. Zhang et al. (2024) demonstrate that T2I models systematically underrepresent or stereotype cultures from less-represented regions in training data. These findings align with our observations of Western-centric defaults across the three heritage test cases. Additionally, T2I models exhibit well-documented compositional failures, difficulty in correctly binding attributes to objects in complex scenes (Huang et al.; Zarei et al., 2025). Heritage reconstruction prompts are inherently compositional, requiring models to associate specific architectural styles, attire, and activities with distinct cultural groups within a single scene, making this limitation particularly consequential in our domain.

2.3 Generative AI in Cultural Heritage

AI and generative AI are increasingly applied in cultural heritage for restoration, documentation, digital reconstruction, and public engagement. AI

image generation models have been used to reconstruct destroyed or incomplete heritage sites and to support immersive AR/VR storytelling (Altaweel et al., 2024; Arzomand et al., 2024; Moral-Andrés et al., 2024; Kutlu and Şimşek, 2025).

However, gen-AI historical image generation faces notable limitations. Models frequently produce visually plausible but historically inaccurate details, reflecting biases in training data that are disproportionately drawn from Western contemporary sources (Rane, 2023; Foka and Griffin, 2024; Spennemann, 2024; Sukkar et al., 2024; He et al., 2025; Liu et al., 2025). The closed-source nature of leading models further complicates scholarly verification of how inputs are processed and outputs are generated.

While substantial research exists on both multivocality in heritage and AI image generation separately, their intersection—using gen-AI to represent multiple perspectives through historical image reconstructions—remains largely unexplored. He et al. (2025) offer the closest prior work, examining how individuals used Stable Diffusion to create personal visual narratives about cultural heritage sites based on their own memories. While they did not investigate the representation of diverse perspectives attributed to the heritage sites themselves, their findings highlight that users' familiarity with a site significantly affects output accuracy, alongside the impact of cultural biases in model training data.

Our research extends this work in three specific ways. First, it incorporates a linguistic analysis of the intermediate prompt transformation process, how AI models rewrite user inputs before generating images, an aspect that remains under-examined in the literature (see Section 2.2). Second, it introduces expert verification by heritage professionals to assess the historical accuracy of depicted details. Third, it systematically compares model performance across multiple culturally contested heritage scenarios, rather than a single site or tradition.

3 Methodology

Given the interdisciplinary nature of this research, we adopted a Mixed Methods approach with a Deductive orientation, while combining Experimental Strategy with Case Study Strategy. This allowed assessing the behaviour of different generative AI models across varied heritage scenarios.

Accordingly, we selected heritage test cases with multiple, often conflicting perspectives, including religious, colonial, conflict, slavery, and multilayered heritage, to evaluate generative AI's ability to represent diverse perspectives. Furthermore, each case incorporated a random perspective to examine how AI responds to users with limited subject familiarity, informed by He et al.'s (2025) findings on the impact of users' familiarity or prior knowledge of the particular heritage context.

The case studies are:

(i) **A Surinamese colonial plantation** — This case study examines the Surinamese plantations during the Dutch colonial period, a contested history of forced labour and cultural identity (3 perspectives: Random, Dutch, Surinamese).

(ii) **Religious beliefs associated with Sri Pada Mountain in Sri Lanka** — This case study focuses on Sri Pada Mountain in Sri Lanka, a religious heritage site uniquely sacred to Buddhist, Hindu, Islamic, and Christian communities (6 perspectives: Buddhist, Hindu, Islamic, Christian-I, Christian-II, Random).

(iii) **1640 Dutch siege of the Portuguese-held Galle Fort in Sri Lanka** — This case study addresses the Dutch VOC siege of Portuguese-held Galle Fort in Sri Lanka in 1640, a turning point in colonial history, not only for Sri Lanka but also for the Indian Ocean region, changing colonial rule from the Portuguese to Dutch (4 perspectives: Dutch, Portuguese, Sri Lankan, Random).

3.1 Image Generation Set-up

Using zero-shot text-to-image generation, we prompted the following three GenAI models with a total of 13 perspectives (from the three case studies), resulting in 39 images.

- DALL-E 3 (Betker et al., 2023)
- Stable Diffusion V3.5 (Esser et al., 2024)
- Midjourney V.6 (Midjourney, Inc., 2024)

Informed by Oppenlaender (2024), we constructed the 13 user inputs for the following eight keyword categories which we derived from the prompt engineering guidelines of the selected models:

(i) Topic
(ii) Landscape/Environment
(iii) Architecture
(iv) People (including attire)
(v) Activities
(vi) Atmosphere
(vii) Special Objects (if any)
(viii) Artistic Style

These identified keyword categories serve dual purposes: (i) guiding effective prompt construction, and (ii) functioning as cultural parameters/criteria for evaluating the historical accuracy of the generated images.

To identify relevant keywords for each perspective, we conducted semi-structured discussions with 7 members of the general public, each lasting approximately 20–40 minutes. We selected participants through purposive sampling to represent the cultural perspectives relevant to each test case: individuals with Dutch and Surinamese backgrounds for Test Case 1, practitioners of the relevant religious traditions for Test Case 2, and individuals with Dutch, Portuguese, and Sri Lankan heritage connections for Test Case 3. When a participant lacked knowledge for a given keyword category, we recorded the phrase "I don't know" as the keyword input. This deliberate inclusion simulates visitors with limited prior knowledge, informed by He et al. (2025), who found that user familiarity with a heritage site significantly affects output accuracy. Each random perspective was obtained from a participant with no prior familiarity with the specific heritage site or its cultural context, simulating a general visitor with limited subject knowledge. Throughout this study, the term 'visitor' refers to an individual representing a particular cultural perspective, simulating the range of people who might engage with a heritage site, whether as members of an associated community, a historically connected nationality, or as a general member of the public.

The user keywords were first input into DALL-E 3, which automatically rewrites user inputs into expanded prompts before generating images (see Table 2). As discussed in Section 2.3, this intermediate prompt rewriting is common practice in current T2I pipelines. It also represents a key site of potential meaning change and constitutes one of the central objects of analysis in this study. Section 3.2 examines how this transformation reshapes user intent across all 13 perspectives.

To compare image generation performance on an equal basis, we used the DALL-E-expanded prompts as identical inputs for Stable Diffusion V3.5 and Midjourney V.6. This two-stage design serves two purposes. First, it enables analysis of how LLM-based prompt rewriting affects culturally specific inputs. Second, it allows a controlled comparison of three image generation models interpreting the same enriched prompt. This means Stable Diffusion and Midjourney were not tested on raw user keywords. Their scores reflect image generation capability given an already-transformed prompt.

3.2 Analytical Procedures

This user prompt-driven AI image generation was followed by analytical and evaluation procedures that combined qualitative and quantitative techniques, analysing two key aspects:

1. **Linguistic Analysis** — examining the intermediate conversion of user-given keywords into model-generated prompts
2. **Visual Analysis** — evaluating the historical accuracy of the AI-generated images

We identified the following eleven linguistic operations through iterative analysis of the 13 keyword-prompt generation pairs, drawing on established categories from paraphrase typology (Vila et al., 2014) and adapted to the specific context of LLM-based prompt rewriting for image generation:

(i) Synthesis/Generation
(ii) Specification
(iii) Expansion
(iv) Enrichment
(v) Clarification/Disambiguation
(vi) Merging
(vii) Substitution
(viii) Omission
(ix) Rephrase
(x) Visual mapping
(xi) Style-context matching

To assess the accuracy of historical details depicted in the generated images, three heritage experts evaluated the 39 image outputs using a custom-designed evaluation form. One domain-specific expert with relevant specialisation (in history, art history, or heritage interpretation respec-

tively) assessed each test case. The evaluation combined quantitative and qualitative assessments: each image was rated across the 8 cultural criteria (see Section 3), using a scale of 1 (no resemblance to the historically expected depiction) to 5 (highly accurate and culturally appropriate representation), with space provided for textual remarks. The qualitative feedback was particularly valuable for interpreting lower scores. It is important to note that the experts were not asked to evaluate the perspective conveyed by the images, but solely to assess the plausibility of visual details across the 8 cultural criteria. As each test case was evaluated by a single expert, inter-rater reliability could not be calculated; this limitation is discussed in Section 6.

Following expert feedback, we cross-analysed all 39 images using three methods:

(a) Single perspective across models

(b) Multiple perspectives within one model

(c) Multiple test cases within one model

4 Data Analysis & Results

The data analysis across three test cases suggested notable variations in how AI models performed depending on the heritage contexts.

4.1 Intermediate Linguistic Transformations between User-Given Keywords and Model-Generated Prompts

In Test Case 1 (Surinamese Plantations), linguistic operations such as *Specification*, *Enrichment*, and *Expansion* have been consistently applied for six keyword categories, except in *Atmosphere* and *Artistic Style*, across Random, Dutch, and Surinamese visitor perspectives, to make prompts more comprehensive for supporting image generation (see Table 1). It is noteworthy that requested architectural references have been omitted or reinterpreted depending on the visitor's standpoint (i.e. accommodating 'Dutch colonial architecture' in all prompts, and omitting 'Surinamese architectural elements', except in the Surinamese prompt). Notably, the model retained the input 'Surinamese workers' only in the Surinamese visitor prompt, but substituted it with 'Enslaved Africans, labourers' for the other two perspectives.

The linguistic operations *Enrichment*, *Expansion*, and *Specification* have been used across all eight criteria in all six perspectives for Test Case 2 (Sri Pada Mountain religious heritage site), producing poetic language and descriptively detailed descriptions. However, except in the Buddhist and Random perspectives, inaccuracies have emerged in all prompts, as the model has incorrectly implied the pre-existence of a footprint or introduced architectural elements, supposedly based on documented history rather than the user's keyword input.

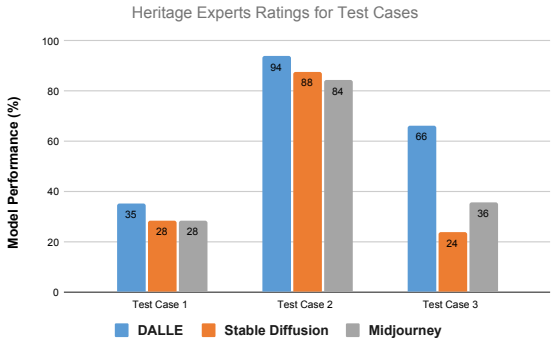

Figure 2: Cumulative heritage expert ratings (and percentages of maximum possible scores) for AI-generated images across three test cases and three models. Scores aggregate one expert's ratings per test case across all perspectives and eight cultural criteria, each rated on a 1–5 Likert scale.

Test Case 3 (Dutch siege of Galle Fort) with four perspectives (Dutch, Portuguese, Sri Lankan, and Random), had an inverse relationship between input details and output quality. While the model extensively expanded the Random visitor's brief input into the most detailed and contextually accurate model prompt, the Dutch and Portuguese visitors' comprehensive inputs were simplified, with commander names omitted and specific historical details substituted by vague generalisations.

Across all three test cases, these intermediate linguistic transformations have had both positive and negative impacts. Positive effects are refining vague inputs into coherent and visually descriptive prompts, while negative effects include inaccuracies, cultural misinterpretations, and the modification of user intents, which directly impacted the resulting historical image reconstructions.

4.2 Comparative Image Analysis

Heritage expert ratings are recorded as cumulative scores for the AI-generated image depictions of each test case (see Figure 2). DALL-E consistently outperformed Stable Diffusion and Midjourney, with the strongest results in Test Case 2 and

Dutch Visitor's Perspective			
Criterion	**User Input Keyword(s)**	**Prompt Phrase(s)/ Expansions**	**Linguistic Operation(s) Observed by the Author**
1) Topic	Surinamese Plantation during the colonial period	a day on a Suriname sugar plantation during the colonial era	Specification
2) Landscape/ Environment	Surinamese land	surrounded by towering palm trees and lush greenery	Expansion Enrichment
3) Architecture	Colonial plantation bungalow, labour quarters/ camps	a white plantation house with a red-tiled roof and wraparound veranda	Specification Enrichment Omission
4) People	Dutch plantation owners, Surinamese workers	enslaved Africans… two European men in period attire	Substitution Specification
5) Activities	Slaves are working, and Dutch Masters overseeing it	toil in the fields… overseen by…	Rephrase
6) Atmosphere	Daytime	…captures a day	*(no significant observation)*
7) Special Objects (if any)	Plantation-related objects	baskets, wheelbarrow, agricultural tools	Generation
8) Artistic Style	Any style suitable for a historical scenery	A traditional oil painting	Style-context matching

Table 1: Linguistic transformations of the Dutch visitor's perspective in Case Study 1 (Surinamese Plantations), showing user-provided keywords, DALL-E 3's rewritten prompt phrases, and the linguistic operations identified for each criterion. Notable transformations include the substitution of 'Surinamese workers' with 'enslaved Africans,' the omission of user-requested labour quarters, and the generation of specific objects not present in the original input.

the weakest across all models in Test Case 1.

In Test Case 1, involving Surinamese colonial plantations, the cumulative scores of all perspectives given by heritage experts for each model were: DALL-E 42, Stable Diffusion 34, and Midjourney 34, out of a possible 120. The score differences were minimal, and no model scored higher than 3 on any individual cultural criterion. These widespread inaccuracies across all visitor perspectives reveal that all three models struggled in representation (see Figure 3).

Test Case 2, which focused on the plurality of religious beliefs of the Sri Pada Mountain heritage site, surprisingly revealed strong results. Despite initial expectations that South Asian religious settings might be challenging to represent for predominantly Western-trained models, all three performed reasonably well in representing diverse pilgrim perspectives. DALL-E led with 225 points, followed by Stable Diffusion with 210, and Midjourney with 202, out of 240. However, some Stable Diffusion outputs resembled Far East Asian rather than South Asian landscapes.

Test Case 3, depicting the Dutch capturing Portuguese-held Galle Fort in 1640, revealed significant performance gaps. DALL-E scored 106, Midjourney 57, and Stable Diffusion 38, out of a possible 160, with each model scoring roughly half that of the preceding one. While DALL-E produced the most historically plausible visualisations, all models still displayed weaknesses in historical accuracy across the four visitor perspectives.

4.3 Overall Observations

We observed several patterns across the three test cases. The linguistic analysis revealed that when users provided vague and ambiguous text inputs (i.e. "I don't know", "…maybe?"), especially in random visitor perspectives, the model (DALL-E) filled the gaps with information that was at times accurate and sometimes not. In the Surinamese test case, we noticed an imbalance in representation of architectural elements, as the colonial bungalows were consistently and prominently depicted, while the user-requested structures, like labour houses, were largely ignored, appearing only in two of the nine images generated. Further we observed imbalances in the Galle Fort test case, with the depiction of Portuguese and Dutch forces, where images have frequently under-represented the Dutch forces regardless of the visitor perspective, especially in Stable Diffusion and Midjourney. However, linguistic transformations that occurred in prompts (see Section 3.2) appear to have had an impact on

this imbalance.

The analysis of the linguistic transformation process further revealed notable model tendencies: neutralisation and amplification, observed in certain contexts. In the Surinamese test case, the model tended to soften or neutralise references to colonial violence (in user inputs) without explicit instructions, and expert feedback confirmed that the resulting images have often been romanticised, lacking any reflection of the historical hardship inherent to the setting. One possible explanation is that the model's safety or content moderation mechanisms may influence how it handles references to historical violence, though this cannot be confirmed given the closed-source nature of these models.

Conversely, in the Sri Pada Mountain test case, the linguistic transformation has amplified the spiritual and emotional language, deepening the sacred tone of the images. However, this same linguistic transformation process has introduced a critical error in the Hindu, Islamic, Christian-I and Christian-II pilgrim perspectives, by misrepresenting those less-documented religious beliefs (the act of imprinting a footprint as a pre-existing one on the rock summit), leading to inaccurate image outputs.

Furthermore, Test Case 3 revealed a surprising inverse relationship between input details and output quality, where the random visitor's brief input yielded the most detailed and contextually accurate prompt, while the model simplified the comprehensive Dutch and Portuguese visitor inputs into brief, vague, generalised prompts.

Across the three test cases, the models performed better on the spiritual heritage scenario (Test Case 2) than on the historical event reconstructions (Test Cases 1 and 3). However, this likely reflects differences in representational complexity rather than an inherent category distinction—Test Case 2 involved fewer compositional demands per image frame. Furthermore, notable differences in how human figures are portrayed were evident across the models. Stable Diffusion produced vague depictions, Midjourney frequently avoided showing faces, and DALL-E consistently rendered highly detailed, expressive faces in the foreground. Similarly, the models differed in their use of colour and style. Despite prompts requesting traditional oil painting styles and historical scenes, Stable Diffusion images consistently used bright, high contrast colours, which experts remark as unrealistic and less effective for historical scene depictions.

Overall, errors, omissions, and misinterpretations of user inputs during the intermediate linguistic transformation process were observed across all perspectives and test cases. In addition, the varied nature of the heritage test cases required domain-specific experts (one per case, with expertise in domains including history, art history, and heritage interpretation) to validate the historical accuracy of the image depictions. However, it also introduced varying levels of criticality and interpretive subjectivity, which should be kept in mind when interpreting the results.

5 Discussion

5.1 Potential of Gen-AI in Fostering Meaningful Dialogue on Multivocality in Heritage Interpretation

Despite the limitations identified, the findings suggest that Gen-AI may offer a useful starting point for engaging audiences with the complexities of heritage interpretation, particularly when used alongside expert guidance. Across the three test cases, while models often struggled with historical accuracy, they demonstrated an ability to visualise multiple perspectives within a single heritage scenario, providing an initial basis for exploring how diverse perspectives on heritage might be visualised.

For instance, in the Sri Pada Mountain case, AI models captured distinct spiritual atmospheres for Buddhist, Hindu, Islamic, Christian pilgrims (and Random visitors), reflecting the site's layered, overlapping beliefs. Even in the more challenging Surinamese and Galle Fort contexts, generated visuals reflected diverse standpoints, local, colonial, and neutral, illustrating how generative AI can serve as a starting point for critical conversations about how complex histories are viewed and remembered.

The models tended to simplify or neutralise sensitive histories in the Surinamese plantation and Galle Fort cases, while amplifying spiritual and emotional language in the Sri Pada Mountain case. These tendencies likely reflect the composition of training data, which is predominantly drawn from digitised Western sources and dominant historical narratives. In heritage contexts such as the Surinamese and Galle Fort cases, the perspectives of colonised and enslaved populations were historically excluded from official documentation by the very authorities who produced it. The available training data for these contexts is therefore both

scarce and skewed toward the dominant perspective, meaning the models have fewer and less diverse references to draw on. AI-generated heritage imagery consequently risks compounding existing representational gaps, amplifying well-documented perspectives while further marginalising those that were suppressed at the source. Recognising these limitations is essential, as gen-AI alone cannot resolve historical bias or representation gaps.

When such imagery is considered for use in cultural institutions, additional considerations arise beyond historical accuracy. This study consulted individuals with Surinamese heritage to inform prompt construction; however, the deployment of AI-generated imagery depicting enslaved populations in museum or visitor engagement settings would require broader consultation with descendant communities. Not only on what is depicted, but on the conditions under which such imagery is displayed, contextualised, and circulated. As several generated images in this study tended toward romanticised depictions of plantation life, institutions adopting these tools should establish review processes that ensure such imagery is accompanied by appropriate contextualisation and does not inadvertently aestheticise historical suffering.

Nevertheless, its capacity to produce multivocal visual narratives offers heritage professionals, educators, and communities a participatory tool that, used critically and alongside expert input, can foster richer and more inclusive dialogue around complex and contested heritage. To support this critical use, we propose practical guidelines derived from the failure patterns documented above.

5.2 Guidelines for Structured Prompt Construction in Heritage Image Reconstruction

The analysis across three test cases and three models revealed recurring failure modes in AI-generated heritage imagery: omission of culturally specific elements, substitution of user terminology with generalised alternatives, romanticisation of sensitive historical contexts, and imbalanced representation of different groups within a scene. Based on these observations, we propose the following guidelines for constructing prompts that mitigate these specific issues.

We recommend describing the intended scene using structured keywords across eight dimensions, rather than in paragraph form, as this reduces the likelihood of elements being merged or omitted during the intermediate prompt transformation:

(a) **Topic** — the event or scenario, including explicit time period and geographic specificity, to prevent the model defaulting to generic historical settings.

(b) **Landscape/Environment** — geography, lighting, season, flora and fauna, structured as foreground, middle-ground, and background where possible, to reduce the model's reliance on stereotypical landscape compositions.

(c) **Architecture** — structures, materials, and culturally distinctive designs. Our analysis showed that colonial-dominant structures were consistently prioritised while user-requested alternatives (e.g. labour quarters) were omitted. Listing all intended structures with equal specificity can counteract this tendency.

(d) **People (including attire)** — clothing, ethnicity, and social roles, using precise terminology. The substitution of 'Surinamese workers' with 'enslaved Africans' in some perspectives but not others demonstrates that user-provided terms are not always preserved. Explicit, consistent terminology across this dimension is critical.

(e) **Activities** — specific actions and interactions, described with sufficient detail to prevent neutralisation. Vague descriptors like 'working in the fields' were consistently softened; specifying the nature and conditions of labour can resist romanticisation.

(f) **Atmosphere/Ambience** — mood and emotional tone. This dimension proved particularly important for counteracting the model's tendency to aestheticise difficult histories. Explicit descriptors such as 'reflecting hardship and forced labour' rather than neutral terms like 'daytime' can guide the model away from picturesque defaults.

(g) **Special Objects** — tools, machines, furnishings, transportation. When left unspecified, the model generated plausible but historically inaccurate objects. Providing specific items grounded in the historical context reduces this risk.

(h) **Artistic Style** — preferred style aligned with the historical context. Our findings showed that style-context matching was one of the more reliable transformations, but specifying style explicitly remains important to avoid anachronistic visual treatments.

We further recommend the following process:

1. Define the image's intended purpose and the perspective it should represent.

2. Construct keywords across all eight dimensions, ensuring equal specificity for all cultural groups and elements in the scene.

3. Generate the prompt using a suitable LLM, then critically review the expanded prompt before image generation — checking specifically for omissions, substitutions, and neutralisation of sensitive content.

4. Generate images across multiple models to compare outputs, as our results showed significant variation in how models interpreted identical prompts.

5. Seek expert review, particularly for professional or educational use in cultural institutions.

These guidelines do not guarantee historically accurate outputs, but they address the specific failure patterns documented in this study and provide a structured basis for iterative improvement.

6 Conclusion and Future Work

This research investigated whether generative AI image models can produce historical reconstructions that reflect multiple cultural perspectives. It examined both the linguistic transformations applied to user prompts and the visual accuracy of the resulting outputs across three heritage test cases and three T2I models. Three findings emerge from this exploratory study. First, intermediate prompt rewriting, a largely unstudied stage in T2I pipelines, substantially affects output accuracy, introducing both helpful enrichments and culturally inaccurate modifications. Second, model performance varied across heritage contexts: all three models performed reasonably on the spiritual heritage scenario but struggled with the representational complexity of colonial and military scenarios. Third, recurring patterns of representation imbalance, content neutralisation, and Western-centric defaults appeared across test cases, consistent with the broader literature on cultural bias in T2I systems. Based on these findings, we proposed practical guidelines for structured prompt construction in cultural heritage contexts. Future work should expand to additional models and heritage contexts, include multiple expert evaluators per test case to enable inter-rater reliability assessment, and test whether structured prompts measurably improve accuracy through iterative application. Generative AI shows early promise as a supplementary tool for exploring multivocal heritage interpretation, but its use requires critical evaluation and expert oversight.

Limitations

The scope of this research was limited to three AI models and three heritage test cases; broadening to a wider range of heritage scenarios and models would yield deeper insights into model behaviour. Each test case was evaluated by a single heritage expert, precluding inter-rater reliability measurement, and engaging multiple experts per case is recommended to reduce the impact of varying interpretive thresholds. The experimental design used DALL-E 3's expanded prompts as inputs for Stable Diffusion and Midjourney, meaning those models were not tested under their own pipeline conditions, which limits direct comparability. The semi-structured discussions involved seven purposively sampled participants, a small sample that may not fully represent the breadth of cultural perspectives relevant to each case. Furthermore, the study assessed historical accuracy of visual details but did not evaluate whether end-users perceived the images as representing meaningfully distinct perspectives. While individuals with Surinamese heritage were consulted for keyword generation, this study did not assess the broader ethical considerations that would arise from deploying AI-generated imagery of enslaved populations in cultural institutions. Future work should explore how museums and heritage sites can establish appropriate ethical frameworks, including descendant community consultation, contextualisation protocols, and display guidelines, before integrating such imagery into public-facing interpretation. Finally, the closed-source nature of the models tested limits our ability to make causal claims about why specific transformations or outputs occurred.

Acknowledgments

We would like to thank the individuals representing the general public who participated in brief, informal discussions, which helped us gather keywords for the user prompts. We are also grateful to the heritage experts who generously shared their knowledge and perspectives through the expert evaluation. This work was partially funded by the Netherlands Organization for Scientific Research (NWO NWA # 1518.22.105, HAICu).

References

Mark Altaweel, Adel Khelifi, and Mohammad Hashir Zafar. 2024. Using Generative AI for Reconstructing Cultural Artifacts: Examples Using Roman Coins. *Journal of Computer Applications in Archaeology*, 7(1):301–315.

Kawsar Arzomand, Michael Rustell, and Tatiana Kalganova. 2024. From ruins to reconstruction: Harnessing text-to-image AI for restoring historical architectures. *Challenge Journal of Structural Mechanics*, 10(2):69.

Sonya Atalay. 2008. *Multivocality and Indigenous Archaeologies*, pages 29–44. Springer New York, New York, NY.

Shanade Bianca Barnabas. 2016. Heritage-making and the dilemma of multivocality in south africa: a case of wildebeest kuil. *International Journal of Heritage Studies*, 22(9):690–701.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, and 1 others. 2023. Improving image generation with better captions. Technical report. Available at `https://cdn.openai.com/papers/dall-e-3.pdf`.

Jacques Derrida. 1994. *Specters of Marx: The State of the Debt, the Work of Mourning, and the New International*, 1st edition. Routledge, New York.

Ana Deumert. 2018. The multivocality of heritage – moments, encounters and mobilities. In Angela Creese and Adrian Blackledge, editors, *The Routledge Handbook of Language and Superdiversity*, 1 edition, pages 149–164. Routledge.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning (ICML)*.

Anna Foka and Gabriele Griffin. 2024. AI, Cultural Heritage, and Bias: Some Key Queries That Arise from the Use of GenAI. *Heritage*, 7(11):6125–6136.

Isabel F. Franke, Stefania D. Conte, Claudia A. Libbi, Victor De Boer, and Tilo Hartmann. 2024. A Polyvocal Approach to Virtual Heritage: An Immersive Case Study. *Journal of Computing and Cultural Heritage*, 17(4):1–23.

Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. 2023. Optimizing Prompts for Text-to-Image Generation. *arXiv preprint*.

Zhiting He, Jiayi Su, Li Chen, Tianqi Wang, and Ray LC. 2025. "I Recall the Past": Exploring How People Collaborate with Generative AI to Create Cultural Heritage Narratives.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench: A Comprehensive Benchmark for Open-world Compositional Text-to-image Generation. *NeurIPS*.

David Kojan. 2008. Paths of Power and Politics: Historical Narratives at the Bolivian Site of Tiwanaku. In Junko Habu, Clare Fawcett, and John M. Matsunaga, editors, *Evaluating Multiple Narratives: Beyond Nationalist, Colonialist, Imperialist Archaeologies*, pages 69–85. Springer New York, New York, NY.

İzzettin Kutlu and Deryanur Şimşek. 2025. Artificial intelligence (AI) assisted digital reconstruction of historical buildings. In *4th International Civil Engineering & Architecture Conference Volume 2: Architecture*. Golden Light Publishing.

Shudong Liu, Yiqiao Jin, Cheng Li, Derek F. Wong, Qingsong Wen, Lichao Sun, Haipeng Chen, Xing Xie, and Jindong Wang. 2025. CultureVLM: Characterizing and Improving Cultural Understanding of Vision-Language Models for over 100 Countries. *arXiv preprint*.

Midjourney, Inc. 2024. Midjourney. `https://www.midjourney.com`. Version 6. Accessed: [insert date].

Fernando Moral-Andrés, Elena Merino-Gómez, Pedro Reviriego, and Fabrizio Lombardi. 2024. Can Artificial Intelligence Reconstruct Ancient Mosaics? *Studies in Conservation*, 69(5):313–326.

Jonas Oppenlaender. 2024. A taxonomy of prompt modifiers for text-to-image generation. *Behaviour & Information Technology*, 43(15):3763–3776.

Nitin Rane. 2023. Role and Challenges of ChatGPT and Similar Generative Artificial Intelligence in Arts and Humanities. *SSRN Electronic Journal*.

Claire Smith, editor. 2020. *Encyclopedia of Global Archaeology*. Springer International Publishing, Cham.

Laurajane Smith. 2006. *Uses of Heritage*, 1st edition. Routledge: Taylor & Francis Group, London.

Dirk H. R. Spennemann. 2024. Generative Artificial Intelligence, Human Agency and the Future of Cultural Heritage. *Heritage*, 7(7):3597–3609.

Ahmad W. Sukkar, Mohamed W. Fareed, Moohammed Wasim Yahia, Salem Buhashima Abdalla, Iman Ibrahim, and Khaldoun Abdul Karim Senjab. 2024. Analytical evaluation of midjourney architectural virtual lab: Defining major current limits in ai-generated representations of islamic architectural heritage. *Buildings*, 14(3).

Marta Vila, M. Antònia Martí, and Horacio Rodríguez. 2014. Is this a paraphrase? what kind? paraphrase boundaries and typology. *Open Journal of Modern Linguistics*, 4(1):205–218.

Yixin Wan, Arjun Subramonian, Anaelia Ovalle, Zongyu Lin, Ashima Suvarna, Christina Chance, Hritik Bansal, Rebecca Pattichis, and Kai-Wei Chang. 2024. Survey of Bias in Text-to-Image Generation: Definition, Evaluation, and Mitigation. *arXiv preprint*.

Linqing Wang, Ximing Xing, Yiji Cheng, Zhiyuan Zhao, Donghao Li, Tiankai Hang, Zhenxi Li, Jiale Tao, QiXun Wang, Ruihuang Li, Comi Chen, Xin Li, Mingrui Wu, Xinchi Deng, Shuyang Gu, Chunyu Wang, and Qinglin Lu. 2025. Promptenhancer: A simple approach to enhance text-to-image models via chain-of-thought prompt rewriting. *arXiv preprint arXiv:2509.04545*.

Alison Wylie. 2008. *The Integrity of Narratives: Deliberative Practice, Pluralism, and Multivocality*, pages 201–212. Springer New York, New York, NY.

Arman Zarei, Keivan Rezaei, Samyadeep Basu, Mehrdad Saberi, Mazda Moayeri, Priyatham Kattakinda, and Soheil Feizi. 2025. Improving Compositional Attribute Binding in Text-to-Image Generative Models via Enhanced Text Embeddings. *arXiv preprint*.

Lili Zhang, Xi Liao, Zaijia Yang, Baihang Gao, Chunjie Wang, Qiuling Yang, and Deshun Li. 2024. Partiality and Misconception: Investigating Cultural Representativeness in Text-to-Image Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–25, Honolulu, HI, USA. ACM.

Nana Zheng. 2023. *Coming to Grips with Diverse Voices in Participatory Heritage Initiatives*. PhD thesis, Vrije Universiteit Amsterdam.

A Appendix

Table 2: Case Study 1 – Surinamese Plantations: User input keywords and AI-generated prompts across three visitor perspectives.

	User Input Keywords		AI-generated Prompt (DALL-E 3)
Random Visitor	Topic:	Surinamese Plantation during the colonial period	A colonial-era oil painting showcases a day on a Surinamese plantation, with enslaved African workers toiling under the watch of Dutch overseers. In the foreground, laborers engage in agricultural tasks, while a grand plantation house with Dutch architectural elements dominates the background, surrounded by lush tropical vegetation. The piece employs earthy tones, rich greens, and subtle splashes of color to create a balanced composition, focusing on the contrast between the workers and their overseers.
	Landscape/Env.:	Surinamese land, maybe flat land I guess	
	Architecture:	Colonial architecture, Surinamese architecture	
	People:	Surinamese people, Dutch people	
	Activities:	Surinamese people working in the land, Dutch overseeing it	
	Atmosphere:	Daytime	
	Special objects:	I don't know	
	Artistic Style:	Any style suitable for a historical scenery	
Dutch Visitor	Topic:	Surinamese Plantation during the colonial period	A traditional oil painting captures a day on a Suriname sugar plantation during the colonial era. At its heart is a white plantation house with a red-tiled roof and wraparound veranda, surrounded by towering palm trees and lush greenery, while enslaved Africans toil in the fields nearby, overseen by two European men in period attire.
	Landscape/Env.:	Surinamese land	
	Architecture:	Colonial plantation bungalow, labour quarters/camps	
	People:	Dutch plantation owners, Surinamese workers	
	Activities:	Slaves are working, and Dutch Masters overseeing it	
	Atmosphere:	Daytime	
	Special objects:	Plantation-related objects	
	Artistic Style:	Any style suitable for a historical scenery	
Surinamese Visitor	Topic:	Surinamese Plantation during the colonial period	A historical 2D reconstruction of a Surinamese plantation during the colonial period, set in the lush tropical Surinamese landscape. In the scene, a colonial-style plantation bungalow with Surinamese architectural elements stands prominently. Surinamese workers are seen laboring in the fields under the hot daytime sun, engaged in plantation activities such as harvesting and carrying loads. Dutch plantation owners in colonial attire are sternly overseeing the work. The environment features typical plantation crops, tools, and equipment of the time. The artistic style reflects a realistic yet painterly tone suitable for a historical scene, emphasizing the atmosphere and social dynamics of the colonial era.
	Landscape/Env.:	Surinamese land	
	Architecture:	Colonial plantation bungalow, Surinamese architecture	
	People:	Surinamese workers, Dutch Plantation owners	
	Activities:	My people (Surinamese) working hard, and Colonial Dutch plantation owners are tightly overseeing	
	Atmosphere:	Daytime	
	Special objects:	Plantation-related objects	
	Artistic Style:	Any style suitable for a historical scenery	

Random Visitor's Perspective

DALL-E

Stable Diffusion

Midjourney

Dutch Visitor's Perspective

DALL-E

Stable Diffusion

Midjourney

Surinamese Visitor's Perspective

DALL-E

Stable Diffusion

Midjourney

Figure 3: Case Study 1 – Images generated of Surinamese Plantations during the colonial period for random, Dutch, and Surinamese perspective.