# Trustworthy Generative AI for Digital Heritage Preservation: A Multimodal Cultural Authenticity Framework

Rohit Sen
Department of Computer Science, University of New Hampshire, Durham, NH, USA.
contactrohit@unh.edu

**Abstract**

The preservation of digital heritage increasingly relies on generative artificial intelligence to restore, reconstruct, and reinterpret cultural artefacts that have been damaged, lost, or dispersed. While generative models offer unprecedented capacity to synthesize high-fidelity images, texts, and multimodal representations, they also introduce profound risks to cultural authenticity, historical accuracy, and community trust. This paper proposes a multimodal cultural authenticity framework designed to embed trustworthiness into every layer of generative AI systems deployed for digital heritage preservation. The framework integrates three interdependent pillars: provenance-aware data governance, culturally grounded evaluation metrics, and human-in-the-loop validation mechanisms that incorporate domain expertise from heritage professionals and source communities. We examine structural trade-offs between generative fidelity and cultural sensitivity, between scalability and contextual specificity, and between automation and interpretability. Through a systems-level analysis, we argue that current approaches to evaluating generative models—based predominantly on photographic realism or lexical similarity—fail to capture the epistemic and affective dimensions of cultural heritage. The paper further addresses governance architectures that support equitable representation, algorithmic fairness across diverse cultural traditions, and long-term sustainability of preservation infrastructures. Deployment considerations include computational resource disparities, institutional capacity, and the risk of reinforcing colonial epistemologies through unreflective AI mediation. By synthesizing insights from large-scale systems engineering, socio-technical infrastructure studies, and critical heritage studies, we propose a holistic pathway toward generative AI that is not only technically robust but also culturally accountable. The framework aims to serve as a reference for researchers, policymakers, and cultural institutions seeking to deploy generative AI in heritage contexts without compromising the integrity of the cultures they seek to preserve.

**Keywords**

digital heritage, generative AI, cultural authenticity, multimodal framework, trustworthiness, algorithmic fairness, heritage governance.

## 1. Introduction

The rapid advancement of generative artificial intelligence has opened transformative possibilities for digital heritage preservation. Text-to-image synthesis, multimodal generation, and neural style transfer now allow institutions to reconstruct damaged frescoes, visualize lost architectural elements, and generate synthetic datasets that fill gaps in archival records [1,2]. However, the very capabilities that make these models powerful also render them sources of epistemic risk. Generative models are trained on vast corpora of predominantly Western,

contemporary, and commercially oriented data, which embed cultural biases that can distort or erase non-dominant heritage traditions [3,4]. When such models are deployed in heritage contexts without careful oversight, they risk producing outputs that are visually convincing yet culturally inaccurate, thereby undermining the authenticity that preservation efforts seek to protect.

The concept of cultural authenticity in digital heritage is not a static property but a relational construct shaped by community consent, historical continuity, and interpretive authority [5]. Authenticity cannot be reduced to pixel-level fidelity or surface-level resemblance to a reference image. Instead, it involves the preservation of symbolic meanings, material practices, and the social contexts in which artefacts originally functioned. Generative AI, by its nature, extrapolates from statistical patterns and often lacks the contextual understanding required to respect these deeper dimensions. As a result, the deployment of generative models in heritage settings demands a systematic framework that embeds trustworthiness at multiple scales—from data curation and model training to output validation and long-term stewardship.

This paper addresses the gap between the technical capabilities of generative AI and the socio-cultural requirements of heritage preservation by proposing a multimodal cultural authenticity framework. The framework is designed to operate across the entire lifecycle of generative AI systems, incorporating provenance-aware data governance, culturally grounded evaluation metrics, and human-in-the-loop mechanisms that center the voices of source communities. We examine the structural trade-offs that emerge when attempting to balance generative flexibility with cultural accountability, and we draw on case illustrations from diverse heritage domains—including Indigenous oral traditions, East Asian manuscript restoration, and Afro-Caribbean architectural reconstruction—to highlight both opportunities and pitfalls.

## 2. Background and Related Work

Digital heritage preservation has long relied on computational methods such as 3D scanning, photogrammetry, and digital archiving to document and disseminate cultural artefacts [6]. These methods, while valuable, typically operate under a paradigm of faithful reproduction: the goal is to capture a static representation of an object or site as it exists at a given moment. Generative AI introduces a paradigm shift by enabling the synthesis of missing or degraded content, thereby moving from documentation to reconstruction and, in some cases, creative reinterpretation [7]. This shift raises fundamental questions about the locus of authority and the boundaries of acceptable intervention.

Research on cultural bias in generative models has documented systematic disparities in how different cultural traditions are represented. For example, text-to-image models tend to generate stereotypical depictions of non-Western cultures and often struggle to produce accurate visual representations of culturally specific artefacts, clothing, or architectural styles [8,9]. A recent study by Shi and colleagues revealed that state-of-the-art text-to-image systems exhibit a measurable cultural gap, where cultural knowledge encoded in the training data is heavily skewed toward Western and mainstream contexts, leaving many global traditions underrepresented or misrepresented [15]. This finding underscores the need for frameworks that explicitly evaluate and correct for cultural authenticity rather than relying on generic fidelity metrics.

Trustworthiness in AI has been extensively studied in domains such as healthcare, finance, and criminal justice, where fairness, accountability, and transparency are paramount [10,11].

However, heritage applications introduce unique challenges because the stakeholders include not only end users and institutions but also living communities whose cultural identities are intimately tied to the artefacts being generated. Existing fairness frameworks often focus on demographic parity or equal opportunity, which are ill-suited to heritage contexts where authenticity is inherently particular and non-comparable across cultures [12]. A culturally grounded approach must instead prioritize consent, provenance, and the right of communities to govern their own heritage representations.

## 3. The Cultural Authenticity Challenge

Cultural authenticity in generative AI is best understood as a multi-layered construct that encompasses material accuracy, symbolic fidelity, and procedural legitimacy. Material accuracy refers to the physical correctness of the generated output—color, texture, shape—relative to available evidence. Symbolic fidelity involves the preservation of meanings, rituals, and social functions that the artefact carried in its original cultural context. Procedural legitimacy concerns the processes by which the output is produced, including whether the generating model was trained with data obtained through free, prior, and informed consent from the relevant communities, and whether those communities have a role in validating the final result [13].

The tension between these layers becomes evident in practice. For instance, a generative model trained on photographs of Mayan stelae may produce highly realistic images of glyphs, but if the model was trained without consulting contemporary Maya communities, the output may inadvertently reproduce colonial narratives by presenting the artefacts as belonging to a "lost" civilization rather than a living culture. Similarly, a model used to restore a damaged Byzantine mosaic might generate patterns that are statistically plausible but historically inaccurate if the training data primarily includes mosaics from a different region or period.

Current evaluation metrics for generative models—such as Fréchet Inception Distance, Inception Score, or CLIP-based similarity—are designed to measure distributional overlap with a reference dataset. These metrics are ill-suited for heritage applications because they reward outputs that match the statistical properties of the training data, which itself may be biased or incomplete. A model that generates a generic "Asian temple" when prompted with a specific temple name may score well on standard metrics if the training data contains many similar images, even though the output is culturally nonspecific. Addressing this challenge requires the development of evaluation protocols that incorporate expert human judgment, community-defined criteria, and provenance tracking.

## 4. Framework Architecture

The proposed multimodal cultural authenticity framework is structured around three interdependent modules that operate across the generative AI pipeline: data governance, model evaluation, and output validation. Each module is designed to be modular and adaptable to different heritage contexts, but together they form a coherent system for embedding trustworthiness.

The data governance module addresses the sourcing, curation, and documentation of training data. It mandates that all heritage data used for generative modeling must include provenance metadata that records the source community, conditions of collection, and any restrictions on use. This module also implements a tiered consent framework that distinguishes between public domain artefacts, those with community stewardship agreements, and those that are culturally sensitive or secret. The governance module further includes mechanisms for data

augmentation that are culturally aware, ensuring that synthetic data generated to address training imbalances respects the symbolic constraints of the original traditions.

The model evaluation module introduces a suite of culturally grounded metrics that go beyond standard distributional comparisons. These metrics include community-based validation scores, where representatives from the source culture assess the fidelity and appropriateness of generated outputs; symbolic coherence measures that compare generated motifs against known cultural patterns; and provenance alignment scores that track whether the training data's provenance is reflected in the model's generative tendencies. Evaluation is not a one-time step but is integrated into iterative training cycles, allowing models to be refined based on feedback from domain experts.

The output validation module provides a human-in-the-loop mechanism that requires generated heritage content to be reviewed by a panel of stakeholders before it can be disseminated or archived. This panel includes heritage professionals, community elders, and AI ethicists. The validation process uses a structured rubric that separates material accuracy from symbolic fidelity, and it includes an appeals process for contested outputs. This module also generates a "cultural authenticity score" that is attached to each output as machine-readable metadata, enabling downstream users to assess the confidence level of the generated content.

## 5. Multimodal Integration and Trade-offs

Generative AI in heritage preservation often involves multiple modalities—text, image, audio, and 3D geometry—that must be integrated coherently. For example, reconstructing a historical manuscript may require generating both the visual appearance of the text and the linguistic content, while also preserving the tactile qualities of the original parchment. Each modality introduces its own set of authenticity challenges, and the integration process amplifies the potential for emergent biases.

A central trade-off exists between generative flexibility and cultural specificity. High-fidelity models that are capable of producing diverse outputs tend to rely on large, heterogeneous training datasets that dilute cultural particularities. Conversely, models trained exclusively on a single culture's data may achieve high symbolic fidelity but lack the ability to generalize across related traditions or to fill gaps in the record. The framework addresses this by employing a federated approach to model training, where separate models are fine-tuned for specific cultural domains and then combined through a multimodal fusion layer that respects domain boundaries.

Another critical trade-off is between automation and interpretability. Fully automated generation pipelines can achieve high throughput but produce outputs that are opaque to human reviewers. Interpretable models, on the other hand, often sacrifice generative quality. The framework promotes a hybrid architecture in which generative processes are accompanied by explanatory modules that highlight the provenance of key features in the output, allowing reviewers to trace why a particular pattern or color was generated. This interpretability is essential for building trust among communities who may be skeptical of algorithmic decision-making.

## 6. Governance and Policy Implications

The deployment of generative AI for digital heritage preservation cannot be divorced from broader governance questions about who decides what counts as authentic, who benefits from

preservation efforts, and who bears the risks of misrepresentation. Current intellectual property frameworks are ill-equipped to handle the collective and intergenerational nature of cultural heritage [14]. A governance architecture for trustworthy generative AI must therefore go beyond individual consent and embrace community-based data sovereignty models.

One promising approach is the adoption of cultural data trusts, in which legal and technical structures are established to give communities ongoing control over how their heritage data is used. These trusts would define usage licenses, audit mechanisms, and revenue-sharing arrangements for any commercial applications. The framework aligns with the CARE Principles for Indigenous Data Governance (Collective Benefit, Authority to Control, Responsibility, Ethics), which have been influential in other digital heritage initiatives [16]. Implementing such governance structures requires collaboration between heritage institutions, technology developers, and policymakers to create regulatory environments that recognize the unique status of cultural data.

Policy implications also extend to funding and infrastructure. Many heritage institutions, particularly in the Global South, lack the computational resources and technical expertise needed to deploy large generative models. Without targeted investment, the benefits of generative AI will accrue primarily to well-resourced institutions in wealthy countries, potentially exacerbating existing inequalities in heritage representation. The framework advocates for a distributed infrastructure model that leverages federated learning and edge computing to reduce reliance on centralized cloud resources, while also supporting capacity-building programs that train local practitioners in AI ethics and data governance.

## 7. Deployment and Sustainability

Deploying a trustworthy generative AI system for digital heritage requires careful attention to operational sustainability. Models must be maintained over decades, not months, because heritage preservation is a long-term commitment. This raises challenges related to model drift, data refreshment, and the preservation of versioned outputs. The framework recommends the creation of a “heritage model lifetime plan” that includes regular audits by community panels, updates to training data as new archival material becomes available, and migration strategies to new model architectures without losing accumulated knowledge.

Sustainability also encompasses environmental considerations. Large generative models consume significant energy, and their continued operation contributes to carbon emissions that disproportionately affect the same communities whose heritage is being preserved. The framework encourages the use of model compression techniques, efficient inference hardware, and renewable energy sources where possible. Additionally, the framework promotes a “minimum viable model” philosophy, where only the smallest model capable of achieving acceptable cultural authenticity is deployed, rather than defaulting to the largest available model.

Another deployment challenge is the risk of misuse. Generative outputs that are convincingly authentic could be used to fabricate historical narratives or to create deepfakes of heritage artefacts. The framework includes cryptographic watermarking that embeds provenance information directly into generated outputs, making it possible to verify their origin and the authenticity score assigned during validation. This watermarking also serves as a deterrent against unauthorized commercial exploitation.

## 8. Conclusion

Trustworthy generative AI for digital heritage preservation demands a paradigm shift from models that optimize for visual realism to systems that are accountable to the cultural communities whose heritage they process. The multimodal cultural authenticity framework presented in this paper provides a structural blueprint for embedding trustworthiness at every stage of the generative pipeline, from data governance through model evaluation to output validation. By centering community consent, cultural specificity, and procedural legitimacy, the framework offers a path toward generative AI that enhances rather than undermines the authenticity of digital heritage.

The trade-offs inherent in this approach—between flexibility and specificity, automation and interpretability, scalability and sustainability—are not obstacles to be eliminated but tensions to be managed through deliberate design and inclusive governance. Future work should focus on operationalizing the framework in diverse institutional contexts, developing community-friendly validation tools, and refining evaluation metrics that capture symbolic fidelity. Ultimately, the goal is not to replace human judgment with algorithmic authority but to augment the capabilities of heritage professionals and communities with AI tools that are transparent, fair, and respectful of the deep cultural significance embodied in every artefact.

**References**


1. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623). https://doi.org/10.1145/3442188.3445922

2. Birhane, A., Prabhu, V. U., & Kahembwe, E. (2021). Multimodal datasets: Misogyny, pornography, and malignant stereotypes. arXiv preprint arXiv:2110.01963.

3. Crawford, K., & Paglen, T. (2019). Excavating AI: The politics of images in machine learning training sets. AI Now Institute.

4. Denton, E., Hanna, A., Amironesei, R., Smart, A., & Wood, S. (2021). On the genealogy of machine learning datasets: A critical history of ImageNet. Big Data & Society, 8(2), 1–15. https://doi.org/10.1177/20539517211035955

5. Dutta, S., & Ghosh, A. (2023). Cultural authenticity in the age of generative AI: A framework for heritage reconstruction. Digital Scholarship in the Humanities, 38(4), 1567–1582. https://doi.org/10.1093/llc/fqad038

6. Economou, M. (2020). Heritage in the digital age: A critical review of the literature. International Journal of Heritage Studies, 26(4), 345–361. https://doi.org/10.1080/13527258.2019.1650640

7. Elgammal, A., & Saleh, B. (2015). Quantifying creativity in art networks. In Proceedings of the 6th International Conference on Computational Creativity (pp. 102–109).

8. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86–92. https://doi.org/10.1145/3458723

9. Hardesty, L. (2022). Generative models and cultural representation: The case of text-to-image synthesis. Journal of Cultural Analytics, 7(3), 1–22. https://doi.org/10.22148/001c.38652

10. Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudík, M., & Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1–16). https://doi.org/10.1145/3290605.3300830

11. Jagadish, H. V., Stoyanovich, J., & Howe, B. (2021). The pursuit of fairness in data systems. Proceedings of the VLDB Endowment, 14(12), 2901–2914. https://doi.org/10.14778/3476311.3476378

12. Kasy, M., & Abebe, R. (2021). Fairness, equality, and power in algorithmic decision-making. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 576–586). https://doi.org/10.1145/3442188.3445919

13. Kukutai, T., & Taylor, J. (Eds.). (2016). Indigenous data sovereignty: Toward an agenda. ANU Press.

14. Morin, J. F., & Saab, S. (2022). Intellectual property and cultural heritage in the digital age. Journal of World Intellectual Property, 25(3), 567–585. https://doi.org/10.1111/jwip.12245

15. Shi, C., Li, S., Guo, S., Xie, S., Wu, W., Dou, J., ... & Chua, T. S. (2025). Where Culture Fades: Revealing the Cultural Gap in Text-to-Image Generation. arXiv preprint arXiv:2511.17282.

16. Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., ... & Hudson, M. (2020). The CARE principles for indigenous data governance. Data Science Journal, 19(1), 43. https://doi.org/10.5334/dsj-2020-043

17. Srinivasan, R. (2017). Whose global village? Rethinking how technology shapes our world. MIT Press.

18. Such, J. M., & Criado, N. (2023). Responsible AI in cultural heritage: A review of ethical frameworks. AI & Society, 38(2), 873–888. https://doi.org/10.1007/s00146-022-01589-x

19. Tylor, E. B. (1871). Primitive culture: Researches into the development of mythology, philosophy, religion, art, and custom. John Murray.

20. UNESCO. (2021). Ethics of artificial intelligence in heritage: A policy guide. UNESCO Publishing.