Evaluating Cultural Hallucinations in Multimodal Generative AI: A Transparency and Trustworthiness Perspective

Brnav Mathur
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
helloarnav@uc.edu

Keling Fu
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
kemail@unr.edu

Lingxuan Xu
Department of Computer Science, George Mason University, Fairfax, VA, USA.
hellolingxuan@gmu.edu

Shane D. May
Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA.
shane135@missouri.edu

Abstract

The rapid advancement of multimodal generative artificial intelligence has introduced capabilities for synthesizing images, text, and video with unprecedented fidelity. While substantial effort has focused on factual hallucinations in language models and object-level hallucinations in vision-language systems, a more insidious class of failure remains underexamined: cultural hallucination, defined as the systematic distortion, stereotyping, or erasure of cultural symbols, practices, and visual identities. This paper provides a system-level analysis of cultural hallucinations through the intertwined lenses of transparency and trustworthiness. We argue that cultural hallucinations arise not merely from insufficient data representation but from a confluence of architectural design choices, opaque curation pipelines, and evaluation protocols that prioritize object-centric accuracy over culturally grounded semantic fidelity. Drawing on interdisciplinary literature spanning machine learning, science and technology studies, and algorithmic auditing, we examine how contrastive pre-training objectives, dataset homogenization, and prompt-to-image decoding amplify hegemonic visual narratives. We further assess the limitations of existing transparency instruments, including model cards and datasheets, in capturing culturally specific failure modes, and propose extensions that incorporate culture-aware auditing, participatory red-teaming, and interpretability probes sensitive to symbolic meaning. Deployment-level consequences are analyzed, highlighting how cultural hallucinations erode public trust, perpetuate symbolic marginalization, and constrain the safe integration of generative models into education, media, and cross-cultural communication. We conclude by outlining a governance and research infrastructure framework that treats cultural veracity as a first-class dimension of model trustworthiness, calling for multi-stakeholder observatories, culturally

stratified evaluation benchmarks, and continuous monitoring mechanisms that transcend static accuracy metrics.

Keywords

cultural hallucination, multimodal generative AI, transparency, trustworthiness, cultural representation, text-to-image generation, algorithmic auditing, socio-technical systems.

1. Introduction

The scaling of generative artificial intelligence models has transformed the landscape of synthetic media, with text-to-image systems such as Stable Diffusion, DALL·E, and Midjourney now capable of rendering high-resolution visual content from natural language prompts. These systems rely on large-scale multimodal architectures that align textual and visual representations through contrastive objectives, typically pre-trained on billions of image-text pairs harvested from the web. As these models enter consumer applications, concerns about their reliability and social impact have intensified, prompting a robust literature on hallucinations, biases, and accountability. Much of the early discourse around hallucination in multimodal AI has centered on factual inaccuracies, where generated images contain objects that contradict physical reality or textual prompts, or on demographic stereotyping that links certain identities to professions or attributes in reductive ways [1], [2]. Model evaluation frameworks and transparency measures, such as model cards and datasheets, were initially developed to surface these statistical disparities and to document intended use cases and limitations [3], [4]. Yet these frameworks remain largely inadequate for diagnosing and mitigating a deeper category of failure that we term cultural hallucination.

Cultural hallucinations refer to the systematic misrepresentation or erasure of culturally situated visual elements—traditional attire, architectural styles, rituals, culinary practices, and symbolic iconography—when generative models are queried with prompts referencing specific cultures or geographies. Unlike demographic stereotyping, which often reduces individuals to a handful of attributes, cultural hallucination involves a flattening of the rich visual vocabularies that constitute collective identity, substituting them with generic, Western-centric, or touristic clichés. This phenomenon has profound implications for the transparency and trustworthiness of multimodal AI, as it undermines the ability of diverse communities to see themselves faithfully represented in synthetic media and erodes confidence in the cultural competence of deployed systems. The present paper undertakes a system-level analysis of cultural hallucinations, positioning them not as isolated annotation errors but as emergent properties of the entire socio-technical pipeline—from data collection and pre-training objectives to model architecture and evaluation methodologies. Through this lens, we examine how architectural trade-offs, governance gaps, and infrastructural choices collectively produce cultural misrepresentation and what transformations in transparency mechanisms, auditing practices, and policy infrastructures are required to move toward culturally veracious generative AI.

2. Characterizing Cultural Hallucinations

To adequately address cultural hallucinations, it is first necessary to delineate the phenomenon from related but distinct categories of model failure. Factual hallucination in text generation, such as inventing historical events or fabricating citations, involves deviation from verifiable world knowledge. In vision-language models, object hallucination describes the generation of visual elements that are not mentioned in the prompt or that violate commonsense scene composition [11], [12]. While cultural hallucinations may involve factual

errors, their core harm lies in semiotic distortion: misrepresenting the meaning, context, and valence of cultural signifiers. For instance, when a model generates an image for "traditional wedding ceremony" and defaults to a Western white dress and church backdrop regardless of the prompt's cultural qualifier, it is not merely making an object-level error but erasing entire systems of ritual and symbolism. Such outputs reflect a tendency to reproduce what has been called markedness in visual semantic AI, where the unmarked default is implicitly coded as Western, male, and white, while other identities are rendered as deviations that must be explicitly invoked [8]. The consequence is a visual monoculture that systematically marginalizes non-hegemonic visual traditions.

Prior research has documented the embedding of stereotypes in multimodal datasets and the resulting amplification in generative outputs. Investigations of datasets such as LAION-400M have revealed substantial overrepresentation of Western imagery and the presence of toxic content linked to identity groups [5], [6]. Complementary evaluations of DALL·E and related models have demonstrated that prompts referencing professions or social roles produce outputs skewed along gender and racial lines, and that subtle linguistic cues can activate cultural biases [7], [9], [10]. However, the cultural dimension extends beyond demographic attributes. Cultural hallucination concerns the integrity of visual language itself: the patterns, color palettes, spatial arrangements, and material textures that convey cultural meaning. A model that generates "a street market in Lagos" with architectural elements indistinguishable from a generic European bazaar is performing a form of cultural hallucination, even if the demographic appearance of the depicted people is locally plausible. This form of error is particularly challenging to detect because it cannot be reduced to a mismatch between a discrete set of attributes and a ground-truth label; it requires evaluation of the semantic fidelity of the entire visual composition with respect to culturally specific knowledge.

The opacity of current evaluation protocols exacerbates the problem. Benchmarks for text-to-image models primarily measure alignment with textual descriptions through metrics such as CLIP score or human preference rankings that are themselves shaped by raters drawn from narrow demographic pools. These protocols are ill-equipped to surface culturally specific knowledge gaps because they lack cultural contextualization. A generated image may receive high alignment scores while remaining deeply incongruent from the perspective of a cultural insider. Thus, cultural hallucinations persist invisibly, eroding trust among communities that never see their visual worlds accurately rendered while reinforcing a false sense of model competence among developers and regulators.

3. Architectural and Data Provenance of Cultural Misrepresentation

Cultural hallucinations are not superficial artifacts that can be fixed with targeted data augmentation; they are structurally embedded in the dominant pipeline for multimodal generative AI. This pipeline comprises three interdependent stages: large-scale dataset construction, contrastive vision-language pre-training, and prompt-conditioned generation. At each stage, design decisions that optimize for scale, efficiency, and generic cross-modal alignment systematically suppress cultural specificity.

The dataset construction phase begins with web-scale crawling that inherently amplifies the content and aesthetic norms of digitally dominant populations. The resulting corpora, while vast, are deeply unrepresentative of global visual culture. Even when geographic diversity is targeted, the images indexed under a country label often reflect the perspectives of tourists, international media agencies, or commercial stock photography rather than the self-representation of local communities [15]. Cultural artifacts are further abstracted through

captioning processes that describe objects in generic terms, stripping away the cultural context that gives them meaning. A handwoven textile from Oaxaca may be labeled simply as "a colorful blanket," losing the specific patterns, techniques, and ceremonial significance that distinguish it. This semantic erasure is then frozen into the model's representational space during contrastive pre-training, where the objective rewards alignment based on co-occurrence statistics that favor frequently depicted, globally recognizable concepts.

The contrastive training paradigm, exemplified by CLIP, constructs a shared embedding space where images and texts that co-occur in the dataset are pulled together while unrelated pairs are pushed apart. While this objective has proven remarkably effective for zero-shot transfer, it operates without any grounding in cultural semantics. Concepts that are visually similar but culturally distinct—such as different types of religious head coverings or ceremonial masks—may be collapsed into nearby regions of the embedding space, making it difficult for the downstream generative model to disambiguate them. The consequence is that prompts seeking culturally specific outputs frequently retrieve a blend of superficial visual features that correspond to the nearest statistical neighbors in the training distribution, resulting in hybridized and contextually inaccurate imagery.

Generative architectures based on latent diffusion add another layer of distortion. Starting from a compressed latent representation of an image, the denoising process is guided by the textual prompt through cross-attention mechanisms that selectively amplify certain visual concepts. Because the training distribution is skewed toward Western visual culture, the model’s prior over visual concepts is similarly skewed. When a prompt contains a culturally specific term with low prior probability, the denoising process tends to over-rely on high-probability generic features, effectively hallucinating the missing cultural details. This explains why a prompt such as "a family celebrating Lunar New Year" might produce an image with generic festive decorations and a red color palette but fail to include specific ritual elements like ancestral altars, traditional calligraphy couplets, or regionally accurate attire. The model is not randomly hallucinating; it is probabilistically interpolating from a prior that lacks the necessary cultural density.

Addressing these architectural origins requires more than curating balanced datasets. It demands a rethinking of pre-training objectives to incorporate cultural distance metrics, the development of disentangled representations that separate cultural style from generic object categories, and decoding strategies that can dynamically consult external cultural knowledge bases during generation. Transparency about these architectural limitations is a prerequisite for meaningful auditing and trust.

4. Transparency Mechanisms for Cultural Auditing

Transparency in AI systems is widely regarded as a foundational enabler of accountability, yet existing transparency instruments were not designed to surface cultural hallucinations. Model cards document performance characteristics across high-level demographic slices, while datasheets describe dataset composition and collection procedures. These tools have advanced the visibility of aggregate bias but are structurally limited in capturing semantic cultural fidelity [3], [4], [16]. A model card may report equalized generation rates of male and female figures across occupations, but it cannot reveal that all generated depictions of a traditional healer from the Peruvian Amazon are visually indistinguishable from a generic shamanic stereotype derived from popular culture. Datasheets may list the geographic sources of training images without detailing the cultural provenance, framing, or representational intent of those images.

To address these gaps, we propose the conceptual extension of transparency mechanisms into the cultural domain through three interlocking instruments: culture-aware datasheets, cultural interpretability probes, and participatory red-teaming. Culture-aware datasheets would augment standard dataset documentation with fields that capture the cultural lens of the image producer, the contextual meaning of depicted symbols, and the consent framework under which culturally sensitive imagery was collected. Such documentation would enable downstream auditors to distinguish between depictions made by community insiders and those produced by external observers, providing a signal about epistemic authority. Developing these datasheets at scale is challenging but can be piloted in partnership with cultural heritage institutions and indigenous media organizations.

Cultural interpretability probes are diagnostic tools designed to test whether a model's internal representations encode culturally relevant distinctions. Building on techniques that identify which model layers are sensitive to protected attributes, culture-oriented probes would measure the degree to which the embedding space separates visually distinct cultural concepts or collapses them into an undifferentiated "ethnic" category. These probes can be used to identify model components responsible for cultural flattening and to guide targeted fine-tuning or architectural modifications. Importantly, the design of such probes must be co-developed with cultural domain experts to ensure that the distinctions being tested align with meaningful, community-defined categories rather than imposed taxonomies.

Participatory red-teaming extends the practice of adversarial testing by involving cultural practitioners and community members in the deliberate probing of generative models for culturally incongruent outputs. Unlike generic red-teaming that searches for toxic or policy-violating content, culturally situated red-teaming would solicit structured narratives, ritual descriptions, and material culture prompts from diverse communities and systematically document where the model's output deviates from culturally grounded expectations. Recent work that benchmarks cultural gaps in text-to-image generation demonstrates the feasibility of quantifying such deviations, revealing systematic underrepresentation of non-Western cultural elements across popular models [18]. Integrating these findings into public transparency reports would signal to affected communities that their visual cultures are recognized as validity criteria, thereby building trust through demonstrated attentiveness rather than abstract commitments to fairness.

The institutionalization of cultural auditing also requires changes to disclosure norms. Currently, model developers rarely release detailed information about the cultural composition of training data or the performance of their systems on culturally stratified evaluation sets. Regulatory frameworks, such as the European Union's AI Act, are beginning to mandate transparency for high-risk systems, but cultural fidelity is not yet a designated dimension of risk. Advocacy for its inclusion must articulate how cultural misrepresentation in globally deployed models can cause collective harm, including the erosion of intangible cultural heritage and the reinforcement of symbolic hierarchies.

5. Trustworthiness and Socio-Technical Deployment

Trustworthiness in AI is a multi-dimensional construct encompassing reliability, safety, fairness, and explainability, but its cultural dimension remains undertheorized. A system that consistently misrepresents the visual culture of a community cannot be considered trustworthy from that community's standpoint, regardless of its performance on generic benchmarks. This insight shifts trustworthiness from a purely technical attribute to a relational property negotiated between the system, its developers, and the diverse publics that interact

with its outputs. The deployment of generative AI in educational tools, cross-cultural media platforms, and digital heritage preservation therefore demands an expanded conception of trust that includes cultural veracity as a core component.

In educational contexts, generative AI is increasingly used to create illustrative content for history, geography, and social studies curricula. A model that generates historically garbled or culturally flattened imagery can inadvertently teach students inaccurate visual narratives, reinforcing orientalist or colonial framings [20]. When a student searches for "traditional clothing of the Maasai" and receives an image that blends elements from multiple unrelated East African groups into a single stereotypical composite, the educational harm is both factual and epistemic: it undermines the student's ability to distinguish between distinct cultures and perpetuates the notion that non-Western peoples are interchangeable. The deployment of such systems in formal education thus constitutes a high-stakes context where cultural hallucination poses direct risks to the integrity of knowledge transmission.

In media and creative industries, cultural hallucinations can lead to public controversies that erode brand trust and inflame cultural tensions. Instances where advertising campaigns or film pre-production materials generated by AI inadvertently caricature cultural symbols can trigger accusations of cultural appropriation and insensitivity, even when no malicious intent is present. The difficulty of attributing responsibility across diffuse supply chains—from dataset curators to model developers, fine-tuners, and end-user prompt engineers—complicates remediation and highlights the need for governance structures that assign clear accountability for cultural harms. Drawing on ethical frameworks for AI that emphasize pluralism and human dignity can help orient these governance discussions toward substantive cultural respect rather than procedural compliance [19].

The erosion of trust is not limited to external stakeholders; it affects developer communities as well. When model builders lack systematic feedback loops to detect cultural hallucinations, they may overestimate the global competence of their systems and make deployment decisions that trigger backlash. Establishing continuous post-deployment monitoring infrastructures that aggregate culturally stratified user reports and automatically flag anomalous output patterns can serve as an early warning system, enabling iterative refinement. Such infrastructures must be designed with privacy-preserving and consent-based mechanisms, particularly when collecting data from marginalized communities who have historically been subjected to extractive research practices.

6. Governance, Policy, and Infrastructure for Cultural Veracity

The governance of cultural hallucinations demands a multi-level approach that spans internal organizational practices, industry standards, and public regulation. At the organizational level, AI development entities should establish cultural integrity review boards that include anthropologists, cultural practitioners, and civil society representatives. These boards would be empowered to audit training datasets, evaluate generative outputs before release, and issue binding recommendations for mitigation. Such structures move beyond ad-hoc fairness checklists to continuous oversight grounded in domain expertise.

Industry-wide standards can be advanced through multi-stakeholder bodies that develop culturally stratified benchmark suites. These benchmarks must go beyond demographic parity metrics to include qualitative assessments of cultural plausibility, judged by panels of community reviewers using detailed rubrics that capture the nuances of material culture, symbolic meaning, and spatial practice. The computational infrastructure to support such

evaluation—distributed annotation platforms, secure community data enclaves, and compensation models for cultural labor—represents a significant engineering and governance challenge but is essential for making cultural veracity a tractable engineering target.

Regulatory frameworks can accelerate the adoption of these practices by mandating cultural impact assessments for high-risk generative AI applications, analogous to environmental impact assessments. Such assessments would require model developers to disclose the cultural composition of training data, to report performance on culturally stratified test suites, and to outline mitigation plans for identified gaps. International coordination through bodies such as UNESCO, which already stewards conventions on intangible cultural heritage, could provide normative guidance and facilitate the sharing of best practices across jurisdictions. The goal is to embed cultural considerations into the lifecycle of AI systems not as an afterthought but as a pillar of responsible innovation.

Finally, the sustainability of these governance interventions depends on building a research and education infrastructure that cultivates interdisciplinary expertise at the intersection of AI and cultural studies. Graduate curricula must integrate cultural theory into machine learning programs, and funding agencies should prioritize projects that develop open-source tooling for cultural auditing. By treating cultural veracity as a legitimate and rigorous research domain, the field can attract diverse talent and ensure that the next generation of multimodal systems is designed from the outset to honor the plurality of human visual expression.

7. Conclusion

Cultural hallucinations in multimodal generative AI represent a pressing but underappreciated failure mode that undermines the transparency and trustworthiness of systems increasingly embedded in global communication, education, and creative production. This paper has argued that cultural hallucinations are not mere data artifacts; they are systemic outcomes of an infrastructure that privileges scale and generic alignment over cultural fidelity. Through detailed analysis of dataset construction, contrastive pre-training, and diffusion-based generation, we have illuminated how architectural choices and evaluation conventions conspire to produce a visual monoculture that erases the richness of the world's symbolic traditions. Extending transparency mechanisms into the cultural domain through culture-aware datasheets, interpretability probes, and participatory red-teaming offers a path toward more accountable systems, while expanding the definition of trustworthiness to include cultural veracity provides a normative foundation for deployment decisions. The governance and policy recommendations outlined here call for institutional innovations—cultural integrity boards, mandatory impact assessments, and community-based evaluation infrastructures—that can hold the entire AI lifecycle accountable to the diverse publics it claims to serve. Addressing cultural hallucination is not merely a technical correction but a structural commitment to epistemic pluralism, without which generative AI will continue to amplify the narrow visual imagination of a digitally dominant minority at the expense of global cultural heritage. The path forward lies in building systems that do not simply reflect the world as it has been unevenly documented but honor the world as it is lived, imagined, and maintained by countless communities.

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 (FAccT ’21) (pp. 610–623). ACM. https://doi.org/10.1145/3442188.3445922

2. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., ... & Gabriel, I. (2021). Ethical and social risks of harm from language models. arXiv preprint arXiv:2112.04359.

3. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT ’19) (pp. 220–229). ACM. https://doi.org/10.1145/3287560.3287596

4. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT ’20) (pp. 33–44). ACM. https://doi.org/10.1145/3351095.3372873

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

6. Luccioni, A. S., & Viviano, J. D. (2021). What’s in the box? An analysis of the LAION-400M dataset. arXiv preprint arXiv:2111.02114.

7. Cho, J., Zala, A., & Bansal, M. (2022). DALL-Eval: Probing the reasoning skills and social biases of text-to-image generative transformers. arXiv preprint arXiv:2202.04053.

8. Wolfe, R., & Caliskan, A. (2022). Markedness in visual semantic AI. In 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’22) (pp. 1269–1279). ACM. https://doi.org/10.1145/3531146.3533174

9. Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., ... & Hashimoto, T. (2023). Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’23) (pp. 1493–1504). ACM. https://doi.org/10.1145/3593013.3594095

10. Struppek, L., Hintersdorf, D., Friedrich, F., Brack, M., Schramowski, P., & Kersting, K. (2022). Exploiting cultural biases via homoglyphs in text-to-image synthesis. arXiv preprint arXiv:2209.08891.

11. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., & Wen, J.-R. (2023). Evaluating object hallucination in large vision-language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 2925–2941). Association for Computational Linguistics.

12. Dai, W., Liu, Z., Ji, Z., Su, D., & Fung, P. (2023). HallusionBench: You see what you think? Or you think what you see? An advanced diagnostic benchmark for multimodal hallucination. arXiv preprint arXiv:2310.14566.

13. Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T., & Saenko, K. (2018). Object hallucination in image captioning. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 4035–4045). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1437

14. Biten, A. F., Gomez, L., & Karatzas, D. (2022). Let there be a clock on the beach: Reducing object hallucination in vision-language models. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (pp. 1381–1390). IEEE.

15. Prabhu, V. U., & Birhane, A. (2020). Large image datasets: A pyrrhic win for computer vision? arXiv preprint arXiv:2006.16923.

16. Hutchinson, B., Smart, A., Hanna, A., Denton, E., Greer, C., Kjartansson, O., ... & Mitchell, M. (2021). Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21) (pp. 560–575). ACM. https://doi.org/10.1145/3442188.3445918

17. Raji, I. D., & Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19) (pp. 429–435). ACM. https://doi.org/10.1145/3306618.3314244

18. C. Shi, S. Li, S. Guo, S. Xie, W. Wu, J. Dou, C. Wu, C. Xiao, C. Wang, Z. Cheng, et al. (2025)Where culture fades: revealing the cultural gap in text-to-image generation.arXiv preprint arXiv:2511.17282.

19. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Schafer, B. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

20. Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. New York University Press.