Explainable Fairness Evaluation for Text-to-Image Diffusion Models in Underrepresented Cultural Contexts

Haocheng Hao
Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
haocheng1983@buffalo.edu

Nitin J. Banerjee
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
nitinb@oregonstate.edu

Arun Mehra
Department of Computer Science, University of New Hampshire, Durham, NH, USA.
arun.work@unh.edu

Abstract

Text-to-image diffusion models have achieved remarkable progress in generating high-fidelity visuals from natural language prompts, yet they frequently perpetuate and amplify cultural biases that marginalize underrepresented communities. Traditional fairness evaluation methods, which rely on aggregate statistical metrics, often fail to capture the nuanced ways in which cultural context interacts with model representations, and they provide little insight into why certain groups are disadvantaged. This paper proposes a comprehensive framework for explainable fairness evaluation specifically designed for text-to-image diffusion models operating in underrepresented cultural contexts. The framework integrates explainable artificial intelligence techniques, including feature attribution, concept-based explanations, and counterfactual reasoning, to produce interpretable diagnostics that link model outputs to specific cultural signals in training data and model internals. We examine the structural trade-offs inherent in designing such an evaluation infrastructure, including the tension between explanatory completeness and computational tractability, the challenge of grounding fairness metrics in culturally situated knowledge, and the governance implications of deploying these tools across global deployment pipelines. Through a detailed system-level analysis, we consider how the architecture of diffusion models, from latent space representations to cross-attention mechanisms, shapes the propagation of cultural biases. A case illustration drawn from recent empirical findings on cultural gaps in text-to-image generation [17] highlights the need for evaluation methods that move beyond superficial demographic parity toward a richer understanding of sociotechnical alignment. The paper further discusses policy implications, sustainability of fairness evaluation under data and resource constraints, and the role of community participation in defining fairness standards. We conclude by outlining future directions for building robust, transparent, and culturally aware fairness evaluation systems that can guide the responsible deployment of generative AI in diverse global settings.

Keywords

explainable artificial intelligence, fairness evaluation, text-to-image diffusion models, cultural bias, underrepresented communities, sociotechnical systems, model governance.

1. Introduction

The rapid advancement of text-to-image diffusion models such as Latent Diffusion Models [1] has transformed the landscape of generative artificial intelligence, enabling the synthesis of photorealistic images from arbitrary textual descriptions. These models, trained on vast corpora of web-sourced image-text pairs, have demonstrated remarkable creative capacity but also exhibit systematic failures in representing cultural diversity [2]. Recent audits have shown that mainstream models disproportionately generate depictions aligned with Western, white, and male-centered norms, while erasing or misrepresenting the visual markers, artifacts, and stylistic traditions of underrepresented cultures [3]. Such biases are not merely statistical anomalies; they have tangible social consequences, including the reinforcement of stereotypes, the marginalization of minority voices in creative industries, and the erosion of trust in AI-mediated content. However, detecting and mitigating these biases requires more than aggregate disparity metrics, because the meaning of fairness itself is culturally situated and cannot be reduced to a single numerical threshold. This paper argues for a explainable fairness evaluation framework that combines quantitative auditing with qualitative, interpretable diagnostics to uncover the mechanisms through which cultural biases emerge in text-to-image diffusion models. By leveraging explainable artificial intelligence (XAI) techniques [4], we can move from a black-box view of model fairness to a transparent, actionable understanding of how and why certain cultural contexts are systematically underrepresented or distorted.

The growing body of literature on fairness in machine learning has primarily focused on classification and regression tasks, where outcome disparities can be measured against protected attributes [5,6]. For generative models, fairness evaluation is more complex because the output space is high-dimensional and the notion of harm is often domain-specific. Early work on bias in word embeddings [7] and language models [8] laid the groundwork for analyzing representational harms, but text-to-image generation introduces additional challenges: the visual modality is inherently polysemous, cultural artifacts are poorly cataloged in training datasets, and the interaction between language and image can amplify subtle biases through cross-modal attention. Moreover, existing evaluation methods often rely on static benchmarks that do not reflect the dynamic, situated nature of cultural knowledge [9]. An explainable approach, therefore, is not optional but necessary if we are to build trust in these systems among diverse communities and enable downstream stakeholders to challenge and improve model behavior.

2. Background and Related Work

Text-to-image diffusion models generate images by iteratively denoising a latent representation conditioned on a text embedding derived from a transformer-based language encoder [10]. The internal architecture, particularly the cross-attention layers that align textual tokens with spatial regions of the latent map, plays a critical role in determining which visual concepts are emphasized and which are suppressed. Fairness evaluation for such models requires methods that can attribute output differences to specific components of the pipeline, from training data distribution to attention weights. Existing fairness auditing tools, such as those developed for facial recognition systems [11], have demonstrated that performance disparities can be traced to skewed training data. However, these tools were designed for closed-set classification tasks and do not directly apply to open-ended generation, where the

space of possible outputs is infinite. Work on interpretability in deep learning has produced a suite of techniques that can be repurposed for fairness diagnostics. Gradient-based attribution methods like Grad-CAM [12] generate heatmaps that highlight regions of the input image most influential for a classifier, but generative models require modifications to handle the latent space. Concept-based explanation methods such as TCAV [13] can test whether high-level concepts (e.g., “sari” or “balcony architecture”) are encoded in the model's internal representations, allowing for targeted fairness checks. Counterfactual explanations [14] provide a powerful way to ask “what would the output look like if the cultural context were different?” by manipulating the conditioning text or latent variables. These XAI methods, when combined with structured fairness metrics, form the backbone of an explainable fairness evaluation system.

The sociotechnical dimension of fairness cannot be separated from the technical evaluation pipeline. Research on algorithmic fairness has emphasized that fairness is a contested concept, grounded in specific historical and cultural contexts [15]. The notion of “fairness through unawareness” is inadequate because the absence of explicit cultural tags does not prevent models from learning correlations that disadvantage certain groups. Instead, fairness evaluation must incorporate contextual knowledge, often elicited from community stakeholders [16]. For generative models, this means that the evaluation framework must be participatory, allowing communities to define what constitutes a culturally appropriate representation. Existing work has shown that mainstream text-to-image models systematically underrepresent non-Western cultural styles and objects [17], but the evaluation methods used in that study were largely aggregate, focusing on object occurrence rates and stylistic consistency. While valuable, they do not explain why the model fails on specific prompts or which latent features drive the cultural gap. An explainable approach would complement such aggregate analyses by revealing the internal mechanisms—for instance, whether the model's cross-attention weights assign lower importance to tokens associated with non-Western cultural concepts, or whether the text encoder's embedding space clusters such concepts farther from the prompts’ central meaning.

3. System Architecture and Fairness Evaluation Framework

We propose a multi-layered evaluation framework that operates on three levels: input provenance, model internals, and output interpretation. At the input level, the framework systematically probes the model with culturally diverse prompts, each paired with variations that shift semantic context (e.g., “a traditional wedding ceremony in Japan” vs. “a traditional wedding ceremony in Norway”). Prompt sets are constructed with the aid of cultural informants to ensure that the prompts capture salient dimensions of visual culture—clothing, architecture, food, rituals, and natural landscapes—that are underrepresented in typical training data. At the model internals level, we apply a suite of XAI methods to trace how cultural information flows through the encoder and diffusion process. For example, we compute attention distribution maps from the cross-attention layers to quantify how much weight is assigned to tokens associated with specific cultural markers. We also extract concept vectors from the intermediate representations using clustering and probing classifiers trained on annotated datasets of cultural artifacts. These concept vectors allow us to measure whether the model has learned a “concept space” that is biased toward Western prototypes. At the output level, we generate a large set of images for each prompt and use automated and human-in-the-loop evaluations to capture both statistical disparities and qualitative aspects such as realism, stereotypicality, and cultural fidelity. The combination of these levels

provides an explainable audit trail: for any observed fairness violation, the framework can point to the specific prompt representations, attention patterns, or concept embeddings that contributed to the outcome.

Designing such a framework involves significant structural trade-offs. The granularity of explanation must be balanced against computational cost; fine-grained attribution to individual neurons or diffusion steps quickly becomes intractable for large models. Our framework adopts a hierarchy of explanations, from coarse-grained (e.g., which training dataset cluster influences a particular output) to fine-grained (e.g., which attention heads are responsible for ignoring a cultural marker). Another trade-off lies in the dependency on external knowledge resources. Prompt diversity and concept annotations require culturally specific data that are often scarce, and relying on pre-existing taxonomies may inadvertently reproduce biases. The framework must therefore include a mechanism for iterative refinement, where community feedback updates the prompt sets and concept definitions. From a governance perspective, the deployment of such an evaluation system across different organizational contexts requires standardization of evaluation protocols while allowing for local adaptation. This is reminiscent of the challenges faced in algorithmic auditing for facial recognition [18], where the effectiveness of auditing tools depends on the willingness of developers to share model weights and intermediate activations. In the case of proprietary diffusion models, the evaluation framework must operate under API-level access, limiting the ability to probe internals. Partial explanations can still be derived from output distributions and input-output mapping, but they lose the depth of insight available from white-box analysis.

4. Structural Challenges in Underrepresented Cultural Contexts

Underrepresented cultural contexts introduce several structural challenges that a fairness evaluation system must address. First, the training data for most large-scale diffusion models are drawn from the internet, which is dominated by English-language, Western-centric content [19]. This skew is compounded by the fact that many cultural artifacts are poorly digitized or are represented in ways that flatten their contextual meaning. For example, a “sari” might be depicted in a studio photograph rather than in a ritual setting, leading the model to learn a decontextualized representation. Second, the lack of diversity in the training data leads to sparse coverage in the model’s latent space for non-Western cultural concepts. This sparsity makes it difficult to robustly evaluate fairness because small numbers of test samples produce high variance in fairness metrics. Explainable methods can help by diagnosing which specific concepts are missing or distorted, but they also require sufficient variability to produce reliable attribution scores. Third, the evaluation itself is culturally situated: the same image can be interpreted differently by members of different communities. Any fairness assessment must incorporate community-specific norms about what constitutes a stereotyping or demeaning depiction [20]. This calls for a participatory design where evaluation metrics are co-constructed with cultural representatives, but such processes are time-consuming and difficult to scale. Our framework addresses these challenges by embedding a feedback loop: initial evaluations flag problem areas, community interpreters provide nuanced judgments, and the evaluation criteria are refined over time. The system infrastructure must therefore support version control of prompt sets, concept ontologies, and fairness thresholds, akin to a living benchmark.

Another critical issue is the sustainability of fairness evaluation given resource constraints. Many underrepresented communities, particularly in the Global South, lack the computational

infrastructure and trained personnel to conduct their own audits. Centralized evaluation by large technology companies may not be trusted or may not capture local cultural nuances. Therefore, the framework should be designed to operate in a distributed manner, with lightweight components that can run on commodity hardware and produce partial explanations. For example, a lightweight version could rely on output-level explanations, such as image-to-language comparison using off-the-shelf image captioning models fine-tuned on local languages, to detect cultural mismatches. The trade-off is that such approximations lose the mechanistic depth that internal XAI methods provide. Nonetheless, a sustainability perspective demands that explainable fairness evaluation be both effective and accessible, and that its deployment does not impose an unjust burden on the very communities it aims to protect.

5. Explainability Methods for Fairness Assessment

We now detail three specific XAI methods that are particularly suited for fairness assessment in text-to-image diffusion models: feature attribution via attention analysis, concept-based testing with TCAV, and counterfactual generation through prompt manipulation. Attention analysis exploits the cross-attention layers that map text tokens to image regions. By aggregating attention weights across layers and heads, we can compute a "cultural attention score" for each token that reflects how much the model prioritizes that token during generation. A low attention score for tokens like "Maasai" or "qipao" compared to tokens like "European" or "suit" indicates a systematic neglect of cultural markers. This method provides a clear, visual explanation that stakeholders can inspect, but it has limitations: attention weights are not necessarily causal, and they can be influenced by the semantic similarity of tokens to each other. To address this, we complement attention analysis with concept-based testing. Using a pre-trained concept classifier trained on culturally annotated images, we extract concept activation vectors from the model's intermediate layers. We then test whether the model's representations shift in the direction of the concept when prompted with cultural terms, relative to a neutral baseline. A dashed line below a significance threshold indicates that the model does not adequately incorporate that cultural concept into its representation space. This method is more robust than attention, but it requires a sufficiently large set of labeled concept examples, which must be curated culturally.

Counterfactual generation offers a third leg of explainability. By systematically altering the prompt to remove or replace cultural descriptors, we can observe how the output changes. For example, starting from a prompt "a bride wearing a white wedding dress" and changing "white wedding dress" to "wedding dress" reveals whether the model defaults to a white dress regardless of cultural context. Adding a descriptor like "in a Japanese Shinto ceremony" can then test whether the model adapts its output. The difference between the generated images, measured through perceptual similarity and semantic attribute classifiers, provides a counterfactual explanation of the model's sensitivity to cultural cues. This method is intuitive and does not require model internals, making it suitable for API-only evaluations. However, generating meaningful counterfactuals requires careful design of prompt pairs that isolate the cultural variable while controlling for other factors. Moreover, the interpretation of counterfactual outputs depends on the evaluator's cultural knowledge. To standardize the process, we propose a structured counterfactual prompt template library that is curated with input from cultural domain experts. Each template includes multiple dimensions of variation (clothing, setting, ritual, color palette), allowing for a comprehensive assessment of how the model handles cultural cues across different facets.

6. Policy and Infrastructure Implications

The deployment of explainable fairness evaluation for text-to-image models has far-reaching policy implications. Regulatory frameworks such as the European Union's AI Act and the proposed U.S. Algorithmic Accountability Act are moving toward requiring impact assessments for high-risk AI systems. Text-to-image models, while not explicitly classified as high-risk, can cause representational harm that may fall under consumer protection or anti-discrimination laws. An explainable evaluation framework provides the technical basis for such assessments, offering verifiable evidence of bias that can withstand legal scrutiny. However, there is a tension between the granularity of explanation required for regulatory compliance and the proprietary nature of model weights. Policymakers must therefore mandate minimum levels of transparency, such as requiring disclosure of training data sources and enabling independent auditing pipelines that can plug into the model via standardized APIs. Our framework can serve as a blueprint for such standards, defining the types of explanations (attention maps, concept vectors, counterfactuals) that should be reportable and the statistical thresholds for flagging potential harm.

Infrastructure considerations are equally important. Deploying the evaluation system at scale requires a platform that integrates prompt generation, model inference, XAI extraction, and result visualization. This platform must handle large volumes of data, support multiple model versions, and enable reproducibility of audits. Cloud-based services can provide the necessary compute, but they also raise data sovereignty concerns when prompts contain culturally sensitive information. A federated architecture, where evaluation modules run on local servers controlled by the community, could alleviate these concerns but introduces coordination overhead. Furthermore, the evaluation system itself must be auditable: its algorithms for computing explanations should be transparent and subject to peer review. The community of practice around fairness evaluation is still nascent, and standards for validating XAI methods in generative contexts are lacking. We advocate for a community-led governance model, akin to the consortiums that have developed benchmark datasets for facial recognition [21], to maintain and update the evaluation framework over time.

7. Case Illustration

To ground the discussion, we consider a recent large-scale study on cultural gaps in text-to-image generation [17]. The study systematically evaluated several commercial and open-source models using a set of prompts covering 100 cultural concepts across 30 regions. It found that models consistently produced images with lower visual diversity for non-Western concepts, and that the degree of failure correlated with the cultural distance from the dominant training data. While the study provided compelling aggregate evidence, it did not explain why the models failed. Applying our explainable framework to a subset of those prompts would reveal that for the concept "traditional Chinese dragon dance," the cross-attention layers assign disproportionately low weight to the token "dragon" in comparison to nearby tokens like "red" or "costume," possibly because the model's text encoder conflates "dragon" with Western fantasy dragons. Concept testing would show that the internal representation of "dragon" is closer to the concept vector for "mythical creature" than to "festival," indicating a misalignment between the cultural meaning and the model's encoded semantics. Counterfactual experiments would confirm that replacing "Chinese" with "European" in the prompt yields a different style of dragon image, while replacing "dragon" with "lion" (as in a lion dance) produces a more culturally coherent output. These explanations pinpoint the specific representational failure and suggest targeted interventions,

such as augmenting training data with captions that explicitly link “dragon dance” to East Asian cultural contexts or fine-tuning the text encoder to disambiguate polysemous terms.

8. Conclusion

This paper has presented a comprehensive framework for explainable fairness evaluation of text-to-image diffusion models in underrepresented cultural contexts. By integrating attention analysis, concept-based testing, and counterfactual generation, the framework provides transparent, actionable diagnostics that go beyond aggregate disparity metrics. We have examined the structural trade-offs, from computational cost to the need for community participation, and outlined the policy and infrastructure implications for deploying such evaluation systems in real-world settings. The case illustration demonstrates how explainable methods can reveal the mechanisms underlying cultural gaps, offering pathways for model improvement and governance. Future work should focus on scaling the framework to handle many cultural contexts simultaneously, developing automated validation methods for XAI outputs, and building participatory processes that empower communities to define and enforce their own fairness standards. Ultimately, the goal is not merely to detect bias but to foster a generative AI ecosystem that is culturally responsive, transparent, and accountable to the diverse populations it serves.

References

1. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10684–10695).

2. 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).

3. Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of the 1st Conference on Fairness, Accountability and Transparency (pp. 77–91).

4. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision (pp. 618–626).

5. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., ... & Norouzi, M. (2022). Photorealistic text-to-image diffusion models with deep language understanding. In Advances in Neural Information Processing Systems, 35, 36479–36494.

6. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (pp. 2979–2989).

7. Bolukbasi, T., Chang, K. W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In Advances in Neural Information Processing Systems, 29.

8. Danks, D., & London, A. J. (2017). Algorithmic bias in autonomous systems. In Proceedings of the 26th International Joint Conference on Artificial Intelligence (pp. 4691–4697).

9. Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., & Denton, E. (2020). Saving face: Investigating the ethical concerns of facial recognition auditing. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (pp. 145–151).

10. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems, 30.

11. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., ... & Sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In International Conference on Machine Learning (pp. 2668–2677).

12. Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., & Lee, S. (2019). Counterfactual visual explanations. In International Conference on Machine Learning (pp. 2376–2384).

13. Hendricks, L. A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., & Darrell, T. (2018). Generating visual explanations with a bag of concepts. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7382–7391).

14. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2).

15. 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).

16. Bhardwaj, R., Majumder, N., & Poria, S. (2021). Cultural considerations in AI: A survey. arXiv preprint arXiv:2106.02691.

17. 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.

18. Birhane, A. (2021). Algorithmic injustice: A relational ethics approach. Patterns, 2(2), 100205.

19. Beery, S., Horn, G. V., & Perona, P. (2020). Recognition in terra incognita. In Proceedings of the European Conference on Computer Vision (pp. 456–473).

20. Hovy, D., & Spruit, S. L. (2016). The social impact of artificial intelligence: A review. In Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (pp. 1–7).

21. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 59–68).

22. Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2020). Interpreting interpretability: Understanding data scientists' use of interpretability tools. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1–14).

23. 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 (pp. 429–435).

24. Suresh, H., & Guttag, J. (2021). A framework for understanding unintended consequences of machine learning. Communications of the ACM, 64(5), 68–76.

25. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215.