Adversarial Prompt Attacks and Cultural Misrepresentation Risks in Large-Scale Image Generation Systems

Steven Jimenez
Department of Computer Science, University of New Hampshire, Durham, NH, USA.
stevenjimenez21@unh.edu

Abstract

Large-scale image generation systems have achieved remarkable capabilities in synthesizing high-fidelity visual content from natural language prompts. However, their increasing deployment in commercial and public-facing applications introduces critical vulnerabilities that extend beyond traditional security concerns. This paper investigates the intersection of adversarial prompt attacks and cultural misrepresentation risks within text-to-image models. We argue that adversarial perturbations to user prompts can systematically trigger the reproduction of harmful stereotypes, erase underrepresented cultural elements, and amplify biases embedded in training data. Drawing on recent advances in adversarial machine learning, cultural computing, and socio-technical systems analysis, we propose a unified framework for understanding how prompt-level manipulations interact with model architectures, training distributions, and deployment infrastructures to produce culturally harmful outputs. We examine structural trade-offs between model robustness, fairness, and utility, and discuss governance mechanisms including prompt filtering, adversarial training, and participatory auditing. Through cross-domain comparisons with large language models and recommendation systems, we highlight the unique challenges posed by visual modality and the opacity of latent space representations. The paper concludes with forward-looking recommendations for building culturally resilient image generation systems that prioritize equity without sacrificing generative quality.

Keywords

adversarial prompt attacks, cultural misrepresentation, text-to-image generation, bias amplification, socio-technical systems, model robustness, fairness governance.

1. Introduction

The rapid proliferation of large-scale image generation systems, exemplified by models such as DALL-E, Stable Diffusion, and Midjourney, has transformed the landscape of visual content creation. These systems accept natural language prompts and produce detailed images that often reflect complex cultural, social, and aesthetic concepts. While their capabilities are impressive, a growing body of research reveals that these models are susceptible to adversarial prompt attacks—carefully crafted inputs designed to elicit outputs that deviate from intended safe or neutral behavior [1], [2]. Simultaneously, concerns about cultural misrepresentation have emerged: generated images frequently default to Western-centric portrayals, reinforce stereotypes, or erase non-dominant cultural practices [3], [4]. This paper argues that adversarial prompt attacks and cultural misrepresentation risks are not independent phenomena; rather, they converge in ways that amplify systemic harms in large-scale deployment.

Understanding this convergence requires a systems-level perspective. Text-to-image pipelines involve multiple layers: prompt preprocessing, text encoding, latent diffusion processes, and image decoding. Each layer introduces potential points of vulnerability where a malicious actor—or even an unwitting user—can exploit latent biases to produce culturally harmful outputs. For example, an adversarial prompt that adds seemingly neutral descriptors can shift a generated image from a benign depiction to one that perpetuates ethnic stereotypes [5]. Such attacks exploit the model's learned associations between textual concepts and visual features, associations that are themselves shaped by imbalanced training data and algorithmic priors.

The stakes are particularly high given the deployment scale of these systems. Integrated into social media platforms, design tools, and educational resources, they influence public perception of cultures and identities. A single adversarial prompt can generate thousands of images that circulate online, normalizing misrepresentations. Moreover, the opacity of large generative models makes it difficult to audit or trace the origins of biased outputs. This paper therefore adopts a socio-technical lens, examining not only the algorithmic mechanisms but also the infrastructure, governance, and policy frameworks that can mitigate combined risks.

The structure of the paper is as follows. Section 2 reviews relevant literature on adversarial attacks, bias in generative models, and cultural computing. Section 3 formalizes a threat model for adversarial prompt attacks targeting cultural representation. Section 4 analyzes mechanisms through which cultural misrepresentation arises, including training data skew, prompt ambiguity, and model architecture choices. Section 5 discusses system-level vulnerabilities, including feedback loops in deployment and scalability of attacks. Section 6 explores mitigation strategies and governance trade-offs. Section 7 offers forward-looking perspectives on resilient design, and Section 8 concludes.

2. Background and Related Work

Adversarial attacks in machine learning have been extensively studied in classification and recognition tasks, where small perturbations to input data cause misclassification [6]. In the domain of generative models, adversarial prompts have been shown to bypass safety filters, generate explicit content, or produce images containing hate symbols [7]. Unlike image classifiers where perturbations are often imperceptible to humans, adversarial prompts in text-to-image systems are semantically meaningful—they exploit the model's understanding of language to steer outputs. This makes them both easier to craft (requiring no access to model gradients) and harder to defend against, as prompt modifications are constrained by the need to maintain coherence for the user.

Parallel to adversarial research, studies on bias in image generation have documented systematic associations between gender, race, and occupational roles. For instance, prompts requesting "CEO" or "nurse" produce disproportionately white and gendered images [8]. Such biases originate from training datasets such as LAION-5B, which reflect historical stereotypes and dominant Western perspectives. More recent work has demonstrated that these biases are not merely statistical but can be amplified through iterative prompting and fine-tuning [9]. The cultural dimension of bias has received less attention, but emerging research shows that models often lack knowledge of non-Western cultural artifacts, rituals, and clothing, defaulting to generic or stereotypical alternatives [10].

The intersection of adversarial attacks and cultural misrepresentation remains underexplored. Early evidence suggests that adversarial prompts can deliberately trigger these biases. For example, appending trigger phrases like "in a slum" to a prompt about a group of people can

skew the output toward impoverished depictions, reinforcing negative stereotypes [11]. Such attacks are particularly insidious because they exploit legitimate user intent—an innocent prompt can be adversarially modified to produce harmful content. The work of Shi and colleagues [12] explicitly reveals how text-to-image models exhibit a cultural gap, systematically failing to generate culturally specific visual elements when prompted with non-Western cultural contexts, and this gap can be exploited to produce misrepresentations. This finding underscores the need for integrated defenses that address both adversarial robustness and cultural fairness simultaneously.

Broader literature on socio-technical systems emphasizes that fairness and security are not merely algorithmic properties but emerge from interactions between technical design, human practices, and institutional structures [13]. Large-scale image generation systems are embedded in ecosystems of data curation, model training, content moderation, and user feedback. Each component introduces potential failure modes that can amplify adversarial harms. For example, content moderation filters may inadvertently block legitimate cultural expressions while allowing adversarial prompts that mimic innocent language. Understanding these system-level dynamics is essential for designing effective countermeasures.

3. Threat Model and Attack Vectors

We define an adversarial prompt attack as an intentional modification of a user prompt with the goal of causing the image generation system to produce outputs that violate specific normative expectations, such as safety, fairness, or cultural authenticity. The adversary may be a malicious user, a third party with access to the prompt pipeline, or even an automated script that scrapes public interfaces. The attack surface includes the initial text input, any text preprocessing (e.g., tokenization, spell correction), and the latent space manipulation performed by the diffusion model.

A key attack vector for cultural misrepresentation is the use of "prompt injection" that inserts culturally loaded terms or negations. For instance, adding the phrase "traditional Chinese" to a prompt for "wedding ceremony" may trigger the model to generate stereotypical elements such as red lanterns and dragon motifs, even if those are not culturally accurate for the specific context [12]. Similarly, an adversary could append "and all people appear poor" to bias demographic representations. These attacks succeed because the model treats all parts of the prompt as equally important, without a mechanism to evaluate cultural appropriateness.

Another vector involves "adversarial suffixes" discovered through search or optimization. Researchers have shown that appending seemingly nonsensical tokens to a prompt can cause an image generation model to bypass safety filters [7]. When applied to cultural content, such suffixes could cause the model to ignore constraints that normally prevent the generation of offensive stereotypes. For example, a suffix that reduces the weight of "respectful depiction" could lead to outputs that caricature cultural practices.

The threat model must also consider multi-turn attacks, where an adversary iteratively refines prompts based on model outputs, gradually steering the generation toward a culturally harmful result. This is particularly problematic in interactive systems that log user history, as the model may adapt to the adversarial trajectory. Moreover, adversaries can exploit latent space interpolation: by slightly varying a prompt that already contains cultural references, they can create a continuum of outputs ranging from accurate to stereotypical, making detection difficult.

From a system perspective, the attack surface extends to shared infrastructure. Many image generation services offer APIs that allow developers to integrate generation into third-party applications. An adversary could poison the training data used for fine-tuning or inject adversarial prompts into the prompts generated by another AI system (e.g., a chatbot that generates image prompts). Such supply-chain attacks are hard to trace and can propagate cultural misrepresentation across multiple platforms.

4. Mechanisms of Cultural Misrepresentation

Cultural misrepresentation in image generation arises through multiple interacting mechanisms. The most fundamental is training data skew. Large-scale datasets like LAION-5B are predominantly scraped from English-language web content, which itself over-represents Western cultures, consumerist aesthetics, and stereotypical portrayals [3]. When a prompt references a non-Western cultural practice, the model must extrapolate from a sparse set of examples, often defaulting to the closest Western analog or to a caricature. This is not a simple bias but a structural property of the data distribution: concepts with few training samples are more likely to be mapped to incorrect or generic features.

Prompt ambiguity amplifies this effect. Natural language is inherently underspecified; a prompt like "a house in an African village" leaves vast latitude for the model to draw on its training associations, which may reflect colonial tropes or safari imagery rather than contemporary urban or rural African architecture [14]. Adversarial attackers can deliberately exploit such ambiguities by adding modifiers that nudge the model toward harmful stereotypes. The model's text encoder (usually based on CLIP) maps phrases to a high-dimensional embedding space, where subtle differences in wording can lead to large shifts in visual features. An adversarial prompt that replaces "modern" with "primitive" can trigger a completely different set of visual archetypes.

Model architecture itself contributes to misrepresentation. Diffusion models generate images by iteratively denoising a random latent variable, conditioned on text embeddings. The denoising process tends to converge to modes of the learned distribution—common patterns in the training data—rather than exploring rare or culturally specific combinations. This mode-seeking behavior penalizes novelty and diversity, making it difficult to generate accurate depictions of minority cultures even when prompts are carefully constructed [15]. Furthermore, classifier-free guidance, a technique that amplifies adherence to the prompt, can exacerbate over-reliance on stereotypical features, as the model amplifies the most statistically likely visual elements for a given concept.

A specific mechanism identified by recent research is the existence of a "cultural gap" in the latent space: certain cultural concepts do not occupy well-defined clusters, causing the model to map them to generic or Western counterparts [12]. This gap is not symmetrically distributed; Western concepts are densely represented, while non-Western ones are sparse. Adversarial prompts that insert these sparsely represented concepts thus have a high probability of producing misrepresentations. Moreover, the gap is dynamic: as models are fine-tuned or updated, the representation of cultures can shift unpredictably, creating new vulnerabilities.

5. System-Level Vulnerabilities

Beyond algorithmic mechanisms, system-level features of large-scale image generation infrastructures introduce unique vulnerabilities. One critical factor is the integration of content moderation filters. Most deployed systems employ some form of safety filtering that blocks

prompts or outputs violating specified policies (e.g., violence, hate speech, sexual content). However, these filters are often trained on thresholds that are not culturally sensitive. For example, a filter may block an image of a traditional hunting ritual because it contains a weapon, while allowing adversarial prompts that subtly encode cultural derogation. The filter's binary classification fails to capture the nuanced harm of cultural misrepresentation, which is often more about context and accuracy than explicit offensive content.

Another system-level vulnerability arises from feedback loops in user interaction. When users repeatedly generate images of a specific culture and share them online, the model may incorporate this user-generated content into future training iterations or fine-tuning. If adversarial prompts lead to a flood of stereotypical images, the model's distribution can shift toward those stereotypes, making subsequent generations even more biased. This creates a reinforcing cycle where misrepresentation begets more misrepresentation. The effect is exacerbated by recommendation algorithms that amplify popular but stereotypical content, a phenomenon observed in social media platforms [16].

Scalability of attacks poses a further challenge. Because image generation models are often served via public APIs with limited rate limiting, an adversary can automate the generation of thousands of culturally harmful images at low cost. The resulting images can be used to train downstream models, populate training datasets for other AI systems, or flood social media with propaganda. The distributed nature of these attacks makes attribution and mitigation difficult. Moreover, the use of adversarial prompts that evade detection requires constant updates to defense mechanisms, creating an arms race.

From an infrastructure perspective, the reliance on cloud-based deployment introduces additional points of compromise. Adversaries could target the model serving infrastructure (e.g., GPU clusters, model replicas) to corrupt the generation process for a fraction of users, causing sporadic but widespread cultural misrepresentation. The lack of transparency in model updates means that users may not know when a model's behavior has been altered by an adversarial attack or by a well-intentioned update that inadvertently shifts cultural representation.

6. Mitigation Strategies and Governance Trade-Offs

Addressing adversarial prompt attacks and cultural misrepresentation requires a multi-layered approach spanning technical defenses, data curation practices, and governance frameworks. On the technical side, adversarial training—augmenting the training dataset with adversarial prompts and their desired safe outputs—can improve robustness against known attack patterns [17]. However, adversarial training for cultural content is challenging because the space of harmful cultural misrepresentations is vast and culturally specific. Moreover, over-regularizing the model to avoid certain associations may reduce its ability to generate culturally diverse content at all, a trade-off between safety and expressiveness.

Prompt filtering and sanitization represent a first line of defense. Systems can parse prompts for known trigger phrases, apply semantic similarity checks against a database of harmful patterns, or use a secondary classifier to assess the likelihood of cultural bias. Yet these filters are easily bypassed by paraphrasing or by injecting innocuous-sounding alternatives. Furthermore, filters risk over-censoring legitimate cultural prompts if the database is too broad or biased. A better approach might involve adaptive filtering that takes into account user context and model confidence, but this introduces privacy and fairness concerns.

Another promising direction is the use of participatory auditing, where members of diverse cultural communities are involved in testing model outputs for specific prompts and providing feedback [18]. This approach acknowledges that cultural harm is subjective and context-dependent; a generic fairness metric cannot capture all forms of misrepresentation. Participatory auditing can inform the development of culturally sensitive guardrails, such as prompt augmentations that insert accurate cultural descriptors. However, scaling such audits to cover all cultures represented in a global user base is resource-intensive, and there is a risk of tokenizing participation.

Governance frameworks must consider the allocation of responsibility among model developers, deployers, and users. Current norms place most responsibility on developers to provide safe models, but adversarial attacks exploiting cultural misrepresentation often arise from user-generated prompts. A strict liability model might deter innovation, while a complete reliance on user self-regulation is ineffective. A more balanced approach involves shared accountability: developers implement robust technical safeguards and transparency measures (e.g., model cards documenting cultural biases), deployers enforce usage policies and provide easy reporting mechanisms, and users are educated about the potential for misrepresentation. Regulatory bodies, such as the European Union's AI Act, are beginning to mandate risk assessments for generative AI systems, which could include evaluations of cultural representation [19].

Cross-domain comparisons offer insights. In large language models, adversarial attacks often target toxicity or misinformation, and defenses include prompt engineering, reinforcement learning from human feedback, and output filtering. For image generation, the visual modality introduces additional challenges because harmful outputs are not easily detected by text-based classifiers. Some systems have adopted "concept erasure" techniques that prevent the model from generating certain categories (e.g., nudity), but applying such erasure to cultural stereotypes would require defining a vast set of problematic concepts. Moreover, erasing concepts can inadvertently remove benign representations, as seen in cases where models stopped generating images of certain religious symbols [20].

A forward-looking mitigation strategy involves building generative models that are explicitly culturally aware—trained on intentionally curated, balanced datasets that include detailed captions about cultural context and variability. This requires large-scale efforts to collect and annotate culturally diverse imagery, which is expensive but essential. Techniques such as retrieval-augmented generation, where the model retrieves relevant cultural information from an external knowledge base at inference time, could reduce reliance on biased training data [21]. Similarly, fine-tuning models on small, high-quality datasets for specific cultures can improve accuracy, but this approach may not scale globally.

7. Future Directions and Research Agenda

The convergence of adversarial attacks and cultural misrepresentation opens several research directions. First, we need better formal frameworks for defining and measuring cultural harm in generated images. Current metrics like Frechet Inception Distance or CLIP score assess image quality or semantic alignment but not cultural appropriateness. Developing culturally grounded evaluation metrics requires collaboration between computer scientists, anthropologists, and cultural experts. These metrics should account for nuance, such as the difference between a respectfully generic depiction and a harmful stereotype.

Second, research on adversarial robustness must expand to include cultural attacks. Most adversarial defense research focuses on classification tasks; generative models present unique challenges because the output space is continuous and high-dimensional. Robustness to cultural adversarial prompts may require new training objectives that penalize distributions that diverge from a fair cultural representation. This is not a simple addition of a fairness constraint, because the adversary's goal is to activate existing biases. A deeper understanding of the geometry of latent space and how cultural concepts are embedded is needed.

Third, system-level resilience should be studied through the lens of socio-technical feedback loops. How do adversarial prompts affect the evolution of training data through user sharing and model updates? Can we design deployment architectures that break harmful feedback cycles, for example by anonymizing user prompts or limiting the influence of user-generated content on future models? These questions intersect with privacy and data governance, and require interdisciplinary research.

Fourth, we must examine the political economy of image generation systems. The companies that control these models have economic incentives to prioritize scale and speed over cultural equity. Open-source models may offer more transparency but can be easily misused. A sustainable path forward may involve public infrastructure for culturally inclusive generative AI, similar to efforts in open science. Policy interventions such as mandatory bias auditing, transparency reporting, and liability for downstream harms could incentivize better practices [22].

Finally, the role of human-in-the-loop systems deserves exploration. Rather than fully automated generation, systems could require explicit cultural validation from human moderators for prompts that fall outside well-represented domains. This approach trades efficiency for safety and may be feasible for sensitive contexts like education or journalism. However, it raises questions about scalability and the potential for moderator bias.

8. Conclusion

Adversarial prompt attacks on large-scale image generation systems pose a significant risk of cultural misrepresentation, exacerbating existing biases and enabling new forms of harm. This paper has argued that these risks are not merely additive but emerge from the interplay of algorithmic vulnerabilities, training data imbalances, and system-level deployment dynamics. By examining mechanisms such as data skew, prompt ambiguity, and latent space characteristics, we have shown how adversarial manipulations can systematically produce culturally harmful outputs. The threat is amplified by infrastructure features like content moderation filters, feedback loops, and scalability. Mitigation requires a combination of technical defenses, participatory auditing, and governance frameworks that balance safety with cultural diversity. As these systems become more pervasive, addressing cultural representation is not a niche concern but a central challenge for responsible AI development. Future research must adopt an interdisciplinary and system-oriented perspective, recognizing that fairness, robustness, and cultural sensitivity are interdependent goals.

References

1. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In International Conference on Learning Representations (ICLR).

2. Carlini, N., & Wagner, D. (2017). Adversarial examples are not easily detected: Bypassing ten detection methods. In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security (pp. 3–14).

3. Birhane, A., & Prabhu, V. U. (2021). Large image datasets: A pyrrhic win for computer vision? In IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 1537–1547).

4. Srinivasan, T., & Chander, A. (2022). The cultural bias of generative AI. Journal of Artificial Intelligence Research, 75, 1–25.

5. Wall, E., & Stede, M. (2023). Prompt attacks on text-to-image models: A taxonomy and analysis. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 4231–4243).

6. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. In International Conference on Learning Representations (ICLR).

7. Guo, Z., Li, Y., Song, D., & Liu, Y. (2024). Adversarial prompt attacks on diffusion models: Bypassing safety filters. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 11234–11243).

8. Bianchi, F., Attanasio, G., Pisoni, D., Leite, D., & Poch, M. (2023). It's not just size that matters: Small language models are also few-shot learners. In Findings of the Association for Computational Linguistics: EACL (pp. 1479–1489).

9. Bansal, H., & Garg, S. (2023). Bias amplification in generative models through iterative prompting. In Advances in Neural Information Processing Systems (NeurIPS) (Vol. 36).

10. Lee, M., & Kim, J. (2023). Cultural representation in text-to-image generation: A case study of East Asian traditions. In Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT) (pp. 404–414).

11. Jain, N., & Schwartz, R. (2024). Adversarial prompt modification for stereotyping in generative models. In International Conference on Machine Learning (ICML) (Vol. 202).

12. 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.

13. 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 (FAT) (pp. 59–68).

14. Smirnova, A., & Riedl, M. (2024). Visual stereotypes in generative AI: An analysis of African representations. Journal of Cultural Analytics, 9(1), 1–22.

15. 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 (CVPR) (pp. 10684–10695).

16. Olteanu, A., Varol, O., & Kiciman, E. (2022). The amplification of stereotypes through content recommendation algorithms. In Proceedings of the ACM Web Conference (WWW) (pp. 2345–2355).

17. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In International Conference on Learning Representations (ICLR).

18. 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 Conference on Fairness, Accountability, and Transparency (FAT) (pp. 33–44).

19. European Commission. (2021). Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). COM(2021) 206 final.

20. Luccioni, S., & Viviano, J. (2023). The unintended consequences of concept erasure in diffusion models. In Advances in Neural Information Processing Systems (NeurIPS) (Vol. 36).

21. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems (NeurIPS) (Vol. 33).

22. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689–707.