Interpretable Latent Space Analysis of Cultural Symbol Representation in Generative Foundation Models

Ruben J. Barnett
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
rubenbarnett910@ku.edu

Ganghai Yan
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
yan958@binghamton.edu

Jeremy Bdams
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
hellojeremy@unr.edu

Abstract

The rapid deployment of generative foundation models in applications such as text-to-image synthesis has raised critical questions about how these systems represent cultural symbols. Latent spaces, which serve as the internal representation manifolds of such models, encode a vast array of conceptual structures, but the interpretability of these representations remains limited. This paper presents a systematic analysis of cultural symbol representation within latent spaces of generative foundation models, focusing on the structural trade-offs between model scalability, interpretability, and cultural fidelity. We argue that current model architectures and training paradigms often produce asymmetrical representations that favor dominant cultural contexts while marginalizing less frequent or historically underrepresented symbols. Through a cross-domain examination of interpretability techniques, infrastructure constraints, and governance frameworks, the paper highlights the need for integrated approaches that combine mechanistic interpretability, socio-technical auditing, and policy design. The discussion extends to deployment sustainability, fairness metrics, and the ethical implications of latent space opacity. By situating cultural symbol representation as a system-level challenge, this study contributes to the broader discourse on accountable AI and offers a roadmap for future research that bridges computer science, cultural studies, and public policy. The analysis underscores that interpretable latent space analysis is not merely a technical problem but a prerequisite for equitable and trustworthy generative systems.

Keywords

latent space interpretability, cultural symbols, generative foundation models, fairness, governance, socio-technical systems, model auditing.

1. Introduction

Generative foundation models, particularly those designed for text-to-image generation, have become ubiquitous in creative, educational, and commercial contexts. These models, built

upon large-scale transformer or diffusion architectures, learn representations of visual and textual concepts through vast training corpora [1,2]. The internal latent space—a high-dimensional manifold where concepts are embedded—is the principal substrate through which these models reason and produce outputs. Understanding how cultural symbols, such as national emblems, religious iconography, traditional clothing, and ritual artifacts, are encoded within these spaces is essential for ensuring that generative systems do not perpetuate historical biases or erase cultural diversity. Yet, the sheer complexity and opacity of modern latent spaces pose significant challenges to interpretability.

The problem of cultural symbol representation is inherently multidisciplinary. It involves technical questions about how latent dimensions correspond to semantic features, as well as socio-political concerns about whose culture is seen and whose is systematically omitted. Early work on model interpretability focused on local explanations or feature visualization [3,4], but these methods rarely addressed the cultural dimension. More recently, researchers have begun to document representational gaps in text-to-image models, revealing that certain cultural contexts are systematically under-represented or distorted [5,14]. These findings indicate that the latent spaces of foundation models are not neutral; they reflect the biases and blind spots of the data on which they were trained.

This paper adopts a system-level perspective to examine the structural trade-offs inherent in making latent spaces interpretable with respect to cultural symbols. We argue that interpretability is not a monolithic goal but involves competing objectives: expressiveness, scalability, and cultural fidelity. We explore how architectural choices, training data infrastructures, and deployment strategies shape these trade-offs. Furthermore, we consider governance and policy implications, proposing that latent space auditing should become a standard component of model governance. The subsequent sections develop these arguments by first establishing theoretical foundations, then analyzing cultural representation asymmetries, followed by architectural and infrastructural considerations, and finally governance and fairness challenges.

2. Theoretical Foundations of Latent Space Interpretability

Latent spaces in generative models can be understood as compressed, abstract representations that capture statistical regularities in training data. In a typical text-to-image diffusion model, the latent space is accessed through an encoder that maps text prompts to a learned distribution, from which image generation proceeds via iterative denoising [1]. Interpretability of such spaces relies on methods that map latent dimensions to human-interpretable concepts. Common techniques include probing classifiers, activation maximization, and concept activation vectors [3,6].

Probing classifiers train simple models to predict a concept label from a subset of latent dimensions, thereby revealing whether that concept is linearly separable in the latent space. While effective for well-defined categories, probes struggle with culturally nuanced concepts that are polysemous or context-dependent. For example, a symbol such as a yin-yang icon may carry different meanings in Taoist philosophy versus contemporary commercial logos. A linear probe may conflate these variations if the training data do not separate them. Another widely used method, concept activation vectors, defines directions in latent space that correspond to a particular concept by contrasting examples with and without that concept [6]. Again, the quality of the resulting vector depends heavily on the contrastive dataset, which is often derived from a globalized, English-centric perspective.

These interpretability approaches reveal a fundamental trade-off: the more expressive and high-dimensional the latent space, the harder it becomes to isolate cultural symbols in a stable, generalizable manner. As models scale in parameter count and training data diversity, the latent manifold becomes increasingly non-linear and entangled [7]. This entanglement means that a single symbolic concept may be distributed across many overlapping features, complicating any attempt at faithful interpretation. Conversely, attempts to enforce disentanglement—for instance, through variational autoencoders with regularization terms—often reduce the generative fidelity of the model [8]. Thus, cultural symbol interpretability is caught between the desire for highly expressive models and the need for reliable, steerable representations.

Moreover, latent spaces are not static; they evolve during fine-tuning, reinforcement learning from human feedback, and continual learning. Each adaptation process may alter the geometry of cultural symbol representations, potentially erasing or reinforcing existing biases. A crucial gap exists in the literature regarding longitudinal studies of latent space drift as models are deployed and updated. Understanding these dynamics is essential for building governance mechanisms that monitor cultural representation over time.

3. Cultural Symbol Representation in Foundation Models: Structural Asymmetries and Representational Gaps

Empirical studies have increasingly documented the systematic under-representation of non-Western cultural symbols in generative foundation models [5,9,14]. For instance, when prompted to generate images of traditional weddings, models often default to Western bridal attire, while South Asian, East Asian, or African ceremonies produce visibly less detailed or stereotyped outputs. Similarly, religious symbols such as the Christian cross are rendered with high fidelity, whereas symbols associated with Indigenous spiritual traditions may be absent or inaccurately depicted. These asymmetries are not merely aesthetic; they have real-world consequences for cultural identity, representation, and the perpetuation of colonial legacies in digital spaces.

The root cause of these representational gaps lies in the training data infrastructure. Foundation models are typically trained on web-scale datasets that are dominated by English-language content and images from North America and Europe [10]. Cultural symbols from these regions appear in millions of examples, ensuring that the latent space learns robust, multi-modal representations. In contrast, symbols from smaller or historically marginalized cultures appear infrequently and often in low-quality or stereotyped contexts. The result is a latent space that is highly discriminative toward frequent symbols and poorly generalizable for rare ones. This is an instance of the long-tail problem, but with a socio-cultural dimension: the long tail consists of entire cultural traditions.

Recent work has attempted to quantify this gap using culturally diverse prompts and human evaluation [14]. The findings indicate that even when models can generate recognizable depictions of diverse cultural symbols, the generated images often lack contextual accuracy—for example, mixing elements from different traditions or anachronistically combining symbols. Such errors reflect the latent space's inability to capture the co-occurrence patterns that define a specific cultural context. This is a structural limitation: the latent space encodes statistical correlations, but cultural symbols often gain meaning through specific narrative, ritual, or historical frameworks that are not reducible to co-occurrence statistics alone.

Addressing representational gaps involves more than simply adding more training data. The latent space must be restructured to enable culturally aware separability. One approach is to use targeted fine-tuning with curated, culturally specific datasets [11]. Another is to employ contrastive learning objectives that explicitly penalize representations that collapse distinct cultural symbols into the same region. However, these interventions introduce new trade-offs. Fine-tuning on a narrow cultural subset can degrade performance on the original distribution, and contrastive objectives may require expensive human annotation to define cultural categories. Moreover, because culture is dynamic and contested, any fixed definition of a cultural symbol risks essentializing it. Balancing the need for precise representation with the fluidity of cultural meaning remains an unsolved challenge.

4. Architectural and Infrastructural Considerations for Culturally Aware Latent Spaces

The architecture of a generative foundation model profoundly influences its capacity to represent cultural symbols. Transformer-based models, which rely on self-attention mechanisms, can learn long-range dependencies between tokens, but they treat input text as a sequence of subwords, which may not align with culturally salient units. For example, a term like “qipao” may be tokenized into multiple subword units, potentially breaking the semantic coherence of the cultural concept [2]. Diffusion models, on the other hand, generate images by iteratively denoising a latent representation, and the latent space is typically learned via a variational autoencoder. The compression ratio of the autoencoder determines how much fine-grained semantic information is preserved; higher compression rates may discard culturally specific textures or patterns.

Infrastructural choices also play a critical role. Training data curation pipelines often employ automated filtering and deduplication, which can inadvertently remove culturally unique images that do not fit Western aesthetic norms. For instance, images of non-Western architecture might be flagged as low quality if they contain irregular lighting or non-standard compositions. The hardware infrastructure—namely the availability of high-performance computing clusters—determines the scale at which models are trained, and this scale often correlates with a homogenization of training data due to the ease of sourcing large English-centric datasets [10]. Consequently, the infrastructure itself reinforces a cycle where culturally diverse data remain scarce and expensive to collect and process.

Deployment strategies further shape cultural representation. Many foundation models are released as APIs with fine-tuning capabilities, but the fine-tuning process is typically user-driven and unmonitored. This creates a risk of adversarial or inadvertently harmful cultural modifications to the latent space. For example, a benign fine-tuning dataset intended to add knowledge of Diwali could, if poorly curated, embed stereotypes of the festival. Moreover, the computational cost of fine-tuning for every cultural context is unsustainable, particularly for resource-constrained communities. Therefore, architectural innovations are needed that enable modular, plug-in representations of cultural knowledge without full model retraining.

One promising direction is the use of adapter layers or hypernetworks that inject culturally specific information into the latent space while keeping the base model frozen [12]. This reduces the computational burden and allows multiple cultural adaptations to coexist. However, adapters may interfere with each other if their latent perturbations overlap, leading to unpredictable interactions. Another infrastructural solution is the development of culturally annotated benchmarks and evaluation suites that are integrated into the model training pipeline as regular checkpoints. These benchmarks would allow continuous monitoring of cultural representation quality, providing early warnings of drift or bias. Such an

infrastructure requires sustained investment from both academic institutions and industry stakeholders.

5. Governance, Fairness, and Policy Implications

The opacity of latent spaces poses a fundamental challenge to governance and accountability. Without interpretable access to how cultural symbols are represented, regulators, auditors, and affected communities cannot verify whether models comply with fairness standards or anti-discrimination laws. Existing AI governance frameworks, such as the EU AI Act, emphasize transparency and explainability, but they rarely specify how these principles apply to the latent space of a generative model [13]. Developing standards for latent space auditing is therefore an urgent policy need.

Fairness in cultural representation cannot be reduced to demographic parity of generated outputs because culture is not a binary attribute. Instead, fairness metrics must capture the fidelity and diversity of symbolic meanings. For example, a model that generates images of a Hindu deity equally often as a Christian saint may still be unfair if the deity is depicted inaccurately or without reverence. Thus, fairness assessments must involve both quantitative measures of representational balance and qualitative evaluations by cultural insiders. This requires participatory governance mechanisms that bring community representatives into model development and auditing processes [15].

Policy interventions should also address the infrastructural asymmetries that produce cultural gaps. Public funding for culturally diverse data collection, open-source curation platforms, and compute subsidies for underrepresented language communities could help level the playing field. At the same time, policy must guard against essentialism: mandating that models represent a fixed set of cultural symbols could freeze cultural evolution and exclude emerging traditions. Adaptive governance frameworks that allow for dynamic updating of fairness criteria as cultures evolve are needed [16].

International coordination is essential because cultural symbols cross borders. A latent space representation that is considered respectful in one context may be offensive in another. Lessons can be drawn from content moderation systems on social media, where context-dependent policies have been implemented through regional advisory boards and local knowledge bases [17]. A similar approach for generative models would involve establishing regional cultural advisory committees that oversee the fine-tuning and auditing of latent spaces for their jurisdiction. The technical challenge of deploying multiple culturally adapted models across different regions is significant, but it aligns with the broader trend toward regional AI regulation.

Finally, interpretability itself is a form of power. Who gets to interpret the latent space—engineers, policymakers, or affected communities—determines whose values are embedded in the model. Democratizing latent space analysis through accessible visualization tools and plain-language explanations is a prerequisite for meaningful public oversight. Research on model interpretability must therefore prioritize methods that are not only faithful to the model but also usable by non-technical stakeholders [18].

6. Conclusion

This paper has argued that interpretable latent space analysis of cultural symbol representation in generative foundation models is a critical socio-technical challenge that spans architecture, infrastructure, governance, and fairness. We have shown that current models exhibit structural

asymmetries that favor dominant cultural contexts, a problem rooted in training data imbalances and architectural entanglements. Interpretability methods, while valuable, are limited by the trade-off between expressiveness and cultural fidelity. Addressing these gaps requires holistic approaches that combine architectural innovations, such as modular adapters, with robust governance frameworks that include participatory auditing and dynamic fairness metrics.

Looking forward, research should focus on developing interpretability techniques that are culturally aware rather than universal, recognizing that the meaning of a symbol is always situated. Continual learning approaches that allow latent spaces to adapt to cultural change without catastrophic forgetting are needed. Interdisciplinary collaboration between computer scientists, anthropologists, ethicists, and policy experts is essential to ensure that generative models serve diverse cultural communities with dignity and accuracy. The latent space is not merely a technical artifact; it is a repository of cultural values, and its interpretability is a prerequisite for accountability in an increasingly AI-mediated world.

References

1. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33, 6840-6851.

2. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

3. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144.

4. 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). International Conference on Machine Learning, 2668-2677.

5. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610-623.

6. Ghorbani, A., Wexler, J., Zou, J., & Kim, B. (2019). Towards automatic concept-based explanations. Advances in Neural Information Processing Systems, 32.

7. Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8), 1798-1828.

8. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., ... & Lerchner, A. (2017). beta-VAE: Learning basic visual concepts with a constrained variational framework. International Conference on Learning Representations.

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

10. Dodge, J., Sap, M., Marasovic, A., Agnew, W., Ilharco, G., Groeneveld, D., ... & Choi, Y. (2021). Documenting large webtext corpora: A case study on the Colossal Clean Crawled Corpus. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 1286-1305.

11. Ilharco, G., Ribeiro, M. T., Wortsman, M., Gururangan, S., Schmidt, L., Hajishirzi, H., & Farhadi, A. (2022). Editing models with task arithmetic. arXiv preprint arXiv:2212.04089.

12. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., ... & Gelly, S. (2019). Parameter-efficient transfer learning for NLP. International Conference on Machine Learning, 2790-2799.

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

14. 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.

15. Sloane, M., Moss, E., Awomolo, O., & Forlano, L. (2020). Participation is not a design fix for machine learning. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 1-13.

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

17. Gillespie, T. (2018). Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press.

18. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

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

20. Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., & Denuyl, S. (2020). Social biases in NLP models as barriers for information access. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 163-173.