Explainable Cultural Bias Auditing Framework for Text-to-Image Generative Models Across Multilingual Contexts

Xuechong Yin
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
xue757@uab.edu

Leishan Duan
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
duanleishan@ku.edu

Damien Bush
Department of Computer Science, George Mason University, Fairfax, VA, USA.
damien.bush@gmu.edu

Abstract

The proliferation of text-to-image generative models has introduced profound challenges in understanding and mitigating cultural biases that manifest across languages and visual representations. Current bias auditing remains fragmented, often focusing on isolated demographic dimensions in a single language, without systematically explaining why culturally skewed outputs emerge or how they propagate across multilingual prompts. This paper proposes the Explainable Cultural Bias Auditing Framework, an integrative system architecture designed to detect, interpret, and govern cultural bias in text-to-image pipelines spanning multiple languages. The framework combines multilingual prompt curation, cross-modal concept probing, gradient-based and attention-driven explainability layers, and governance interfaces that connect internal model behavior with external fairness standards. We examine the structural trade-offs between audit depth, computational scalability, and real-time deployment; analyze the interdependency between data provenance, model architecture, and cultural representational gaps; and situate the framework within broader policy landscapes, including the EU AI Act and the NIST AI Risk Management Framework. Through a systems-level discussion, we highlight how infrastructure choices, sustainability constraints, and explainability granularity collectively shape the robustness of cultural bias detection. Rather than proposing a single algorithm, the paper articulates a principled architectural paradigm that integrates fairness-aware infrastructure, multilingual semantic alignment, and interpretability-first design, providing a scaffold for future auditing systems in generative AI ecosystems.

Keywords

cultural bias, text-to-image generation, explainable AI, multilingual auditing, fairness infrastructure, diffusion models, governance.

1. Introduction

The rapid scaling of text-to-image generative models, epitomized by diffusion-based architectures and large vision-language encoders, has rendered automatic visual content creation a pervasive socio-technical phenomenon. Systems such as Stable Diffusion, DALLÂ·E, and Imagen now routinely produce photorealistic outputs driven by free-form natural language prompts, enabling applications ranging from creative design to informational media. However, alongside their technical sophistication, these models have been shown to mirror, amplify, and at times distort societal stereotypes embedded in their training data. Early investigations into machine learning fairness demonstrated that even ostensibly benign classification systems can encode intersectional biases along race, gender, and class [1], prompting structured accountability mechanisms such as model cards [2] and data statements [3] to improve transparency. Language models, too, were found to project harmful analogies learned from corpora, such as gendered occupational associations [4], and to generate stereotypical completions when conditioned on demographic terms [5, 6, 7]. These lineage of fairness research now confronts a more challenging frontier: the multimodal domain, where textual biases are rendered visible, often in ways that are visceral, culturally specific, and difficult to reverse-engineer.

The transition from textual to visual output amplifies representational harms because images carry dense cultural signals that can be interpreted differently across communities. Unlike language-only models, where biased lexical choices can be flagged relatively easily, the semiosis of an imageâ€”the clothing, architecture, skin tone, gestures, and spatial arrangements it depictsâ€”is shaped by cultural conventions that text-to-image models collapse into statistical regularities of massive web-scraped datasets. Multiple investigations have now documented systematic representational skews in diffusion models, including occupational gender segregation, racialized depictions of social roles, and western-centric defaults when prompts lack explicit cultural markers [8, 9, 10, 11, 12, 13]. Moreover, probing benchmarks have shown that interventions attempting to reduce bias through ethical prompt engineering often yield inconsistent results across model versions and languages [14, 15]. These findings underscore a critical gap: current auditing practices are predominantly monolingual, rely on predefined demographic axes, and provide limited explanatory insight into why a particular cultural association was generated or how it could be corrected without introducing new distortions.

Compounding these challenges is the multilingual dimension. Most popular text-to-image models are trained predominantly on English-language captions, with multilingual understanding retrofitted through translation layers or multilingual vision-language encoders that map diverse languages into a shared semantic space. While cross-lingual alignment techniques such as multilingual CLIP have begun to bridge language gaps [16], they carry the risk of superimposing English-centric cultural defaults onto non-English communities. A recent large-scale cross-cultural evaluation of text-to-image systems revealed pervasive cultural fading, where non-Western visual traditions were misrepresented or entirely absent in generated images [17]. This demonstrates that cultural bias is not merely a characterisation problem: it is a structural consequence of how data, model design, and language representation interact across the full system lifecycle. Auditing frameworks that ignore this systemic entanglement will consistently underestimate the depth of representational harm.

This paper proposes an Explainable Cultural Bias Auditing Framework designed to operate across multilingual contexts within text-to-image generative pipelines. Rather than presenting a narrow technical fix, we articulate a system-level architecture that integrates multilingual

prompt curation, cross-modal bias detection, explainability layers, and governance interfaces. We focus on the structural trade-offs between audit granularity, computational sustainability, real-time deployment constraints, and policy alignment. The contribution is therefore not a singular model or metric, but a coherent conceptual blueprint for building auditing infrastructures that are both interpretable and culturally reflexive. Throughout, we emphasise the importance of infrastructure design choicesâ€”data provenance tracking, modular monitoring, energy-aware scheduling, and multilingual semantic groundingâ€”in determining whether an auditing system can function robustly across diverse linguistic communities.

2. The Landscape of Cultural Bias in Text-to-Image Generation

Understanding cultural bias in text-to-image generation requires moving beyond generic fairness taxonomies and engaging with the specific mechanisms by which visual cultural markers are learned, distributed, and suppressed. In contrast to classification bias, where a protected attribute is directly predicted or used as input, generative bias operates latently: a prompt such as â€œa family eating dinnerâ€ can trigger vastly different culturally coded images depending on the modelâ€™s training distribution, without any explicit cultural attribute being mentioned. These outputs are not random; they reflect the disproportionate representation of Western visual culture in the datasets that underlie dominant text-to-image models. The LAION-5B corpus, for instance, is heavily skewed toward English-language alt-texts and geographically concentrated sources, which biases the visual concepts that diffusion models can reproduce [8, 11, 17]. As a result, the default visual world of these models is implicitly Anglo-American, with deviations surfacing only when users deliberately insert culturally specific keywords.

The cultural dimension is further complicated by the polysemy of visual symbols across cultures. A prompt such as â€œa wedding ceremonyâ€may yield unmistakably Western white-dress imagery for English inputs, yet when the same model processes the prompt in another language through a multilingual encoder, it may still default to similar visual templates, erasing local traditions. This phenomenon arises not only from training data imbalances but also from the internal geometry of the aligned multimodal embedding spaces, where images and texts from different languages that are mapped to similar vector regions end up inheriting dominant visual associations. Auditing this requires more than surface-level testing; it demands attribution mechanisms that can trace which training samples, linguistic cues, or attention patterns triggered a culturally biased output. Existing explainability methods in computer vision, such as gradient-weighted class activation mapping and feature attribution, have been adapted for diffusion models, with cross-attention visualization revealing how specific tokens in a prompt steer the generation of visual attributes [20, 21]. Yet these techniques have seldom been extended to compare how the same concept expressed in different languages yields divergent cross-attention distributions, a gap our framework explicitly addresses.

Additionally, cultural bias auditing must contend with the fluid and contested nature of cultural categories. Unlike binary gender or racial categories that have been operationalized in algorithmic fairness work, culture is multi-axial, geographically fuzzy, and dynamically evolving. A static taxonomy of cultures will inevitably fail to capture intra-national diversity, diaspora identities, and hybrid cultural forms that are increasingly the norm. Auditing systems must therefore be designed as learning infrastructures that incorporate feedback from culturally diverse annotators and community-based oversight mechanisms, rather than assuming a fixed set of protected cultural groups. This requirement positions cultural bias

auditing not as a one-time certification step but as an ongoing socio-technical monitoring process, embedded within the model deployment lifecycle. The governance layer of the proposed framework is designed to accommodate this fluidity by linking internal model explanations to externally maintained cultural knowledge bases and stakeholder feedback loops, ensuring that bias detection criteria can evolve as cultural understandings shift.

3. System Architecture of the Explainable Cultural Bias Auditing Framework

The proposed framework is built around four interdependent subsystems: a multilingual prompt curation and augmentation module, a cross-modal bias detection engine, an explainability layer that surfaces internal model dynamics, and a governance interface that translates internal findings into policy-relevant reports and actionable recommendations. These subsystems are designed to interoperate within a modular, containerized infrastructure, where each component can be updated or replaced as model architectures evolve and new cultural audit dimensions emerge. The entire auditing pipeline can be deployed as a batch evaluation service during model development, as a continuous monitoring system in production, or as an on-demand forensic tool triggered by user complaints or flagged anomalies.

The multilingual prompt curation module serves as the entry point for systematic bias probing. It maintains a curated database of cultural scenario templates, each expressed in multiple languages and annotated with cultural ground-truth expectations provided by regionally diverse annotators. To avoid the trap of considering a single â€œcorrectâ€ cultural representation, each template includes a range of plausible visual manifestations, allowing the bias detection engine to assess not just whether a generated image matches a single reference but whether it falls within an acceptable cultural spectrum. Crucially, the module augments prompts with controlled linguistic variationsâ€”tense, formality, honorifics, dialectal formsâ€”to detect how subtle language shifts alter cultural outputs. This design draws on insights from data statement research, requiring that each prompt be accompanied by metadata detailing source language, intended cultural context, and annotator demographics, thereby enabling downstream explainability analyses to isolate language-specific effects from broader model biases.

The cross-modal bias detection engine ingests generated images and their corresponding prompts and computes a multidimensional bias profile. Instead of relying solely on classification accuracy against predefined categories, the engine employs a combination of semantic similarity scoring, object and scene graph extraction, and latent space proximity measurements relative to carefully constructed cultural anchor sets. For each query language, anchor sets are constructed from culturally authentic reference images sourced from open-access repositories and verified by domain experts. The detection engine then computes not a single fairness metric but a bias signatureâ€”a structured matrix capturing how frequently, and with what intensity, the model deviates from cultural anchors under different linguistic conditions. This signature is the input to the explainability layer, which must then answer the question: why did a particular deviation occur? The multilingual dimension is operationally critical here, as the engine must accommodate embeddings from multiple vision-language models simultaneously, comparing outputs generated from prompts in, for example, English, Arabic, and Yoruba, while correcting for the fact that the underlying image representation spaces may be differently aligned across languages due to training imbalances in the multilingual text encoder.

The explainability layer is the architectural core of the framework, designed to translate internal model states into human-interpretable evidence about the sources of cultural bias. It integrates three complementary explanation modalities: gradient-based saliency maps that highlight which pixels and latent features contributed most to a specific visual element; cross-attention decomposition methods, such as DAAM [21], that trace how individual prompt tokens influenced spatial regions in the final image; and counterfactual probing, where the engine systematically perturbs prompt tokens and model intermediate representations to observe how cultural attributes shift. For multilingual auditing, counterfactual probing is extended to cross-lingual counterfactuals, where a prompt that generated a western-centric wedding when posed in English is re-evaluated in Hindi to see whether the visual distribution shifts in a culturally expected direction. If no shift occurs, the explainability layer must analyze whether the lack of shift is due to token under-representation in the multilingual text encoder, over-dominance of certain visual concepts in the image decoder, or insufficient cross-modal alignment for that language pair. This granular diagnostic information is then summarized into structured explanation reports that are passed to the governance interface.

The governance interface translates technical explanations into standards-conformant documentation. It automatically generates audit reports aligned with the transparency requirements of the EU AI Act and the NIST AI Risk Management Framework [22, 23], including summaries of cultural bias detected, languages affected, manifestations observed, and the underlying model components implicated. Beyond compliance artifacts, the interface offers a feedback dashboard that allows community auditors and external stakeholders to challenge specific assessments, propose new cultural scenarios, and trigger re-evaluation. This participatory loop ensures that the framework does not ossify into a static tool but remains adaptive to emerging cultural sensitivities. Importantly, the governance interface also logs all audit runs with immutable provenance trails, so that temporal trends in cultural bias can be monitored as models and datasets are updated. This longitudinal capability is essential for holding developers accountable not only at model release but throughout the operational lifetime of the system.

4. Infrastructure Design and Structural Trade-Offs

Deploying the framework at scale involves navigating a set of structural trade-offs that are common to fairness-conscious AI infrastructure but take on heightened complexity in multilingual cultural auditing. The first trade-off concerns audit depth versus computational cost. Granular explainability methods, such as token-level cross-attention analysis and counterfactual image generation, are computationally intensive, particularly when repeated across thousands of prompt templates and dozens of languages. Running a full-depth audit on every new model version may be infeasible under tight release cycles or in low-resource compute environments. The framework addresses this by introducing a tiered auditing mode: lightweight statistical monitoring runs continuously, while deep explainability probes are triggered only when the statistical bias signature exceeds a predefined threshold or when stakeholders request a forensic investigation. This tiered design draws on concepts from model risk management, where resource allocation is proportional to risk level, and requires careful calibration of the thresholds to avoid excessive false positives that would drain compute budgets.

A second trade-off lies in the tension between cultural specificity and generalization. Constructing anchor sets for hundreds of cultural groups is resource-intensive, and anchoring too tightly to a specific set may cause the auditing system to flag culturally innovative

outputsâ€”those that creatively blend traditionsâ€”as deviations. Conversely, overly permissive anchor sets risk normalizing harmful stereotypes. The framework leans toward specificity while incorporating a diffusion-explanation mechanism that distinguishes between statistically underrepresented tokens and harmful stereotyping. When an output deviates from a cultural anchor, the explainability layer categorizes the deviation as either a representation gap (a missing or underrepresented concept) or an amplification of a stereotypical association, based on whether the token-level attributions activate concepts that are documented as harmful in external bias lexicons. This classification is policy-relevant because it helps regulators differentiate between deficiencies in cultural coverage, which may be addressed by data diversification, and active stereotyping, which may require architectural debiasing or even withholding model deployment in certain contexts.

Sustainability considerations also shape the architecture. Auditing systems that themselves consume significant energy risk undermining the ethical goals they pursue, a concern that has been documented in the large-scale deep learning lifecycle [24]. The framework therefore mandates energy telemetry for all audit runs, integrating with carbon-aware scheduling systems that can postpone non-urgent deep audits to periods of low grid carbon intensity. Additionally, the prompt curation database and anchor sets are designed with deduplication and incremental updating, so that repeated audits do not recompute embeddings for previously seen prompts unless the underlying model has changed substantially. This incremental architecture leverages model version fingerprinting, where changes in model weights are tracked to decide whether a full re-audit is warranted or only an incremental delta-audit is needed, thereby reducing redundant computation.

The interplay between data provenance and model architecture is another critical infrastructure dimension. Cultural bias does not reside solely in model weights; it is equally a property of the data pipelines that feed training, fine-tuning, and evaluation. The auditing framework mandates that every inference result be traceable to a set of training artefacts and augmentation steps, via a provenance graph maintained by the system. When a particular language shows severe cultural fading, the governance layer consults this graph to determine whether the failure originates in the dataset used for multilingual text encoding, the image-text pairing process, the diffusion modelâ€™s denoising scheduler, or the safety filters applied post-generation. This multi-component attribution transforms auditing from a monolithic attribution exercise into a systems debugging activity, where responsibility can be assigned to specific stages of the pipeline. This, in turn, informs remediation strategies that may involve targeted data collection from underrepresented linguistic communities, retraining of attention layers, or modification of output filtering rules.

5. Explainability Mechanisms for Cultural Bias in Generative Architectures

Explainability in diffusion-based text-to-image models presents unique challenges compared to discriminative models because generation is an iterative denoising process where decisions are distributed across hundreds of sampling steps. Gradient-based saliency methods originally developed for classifiers [20] can be adapted to the final output image by backpropagating a cultural bias score through the denoising chain, but this approach requires careful handling to avoid gradient saturation in the early noise stages. The framework employs a hybrid approach that combines saliency on the final decoded image with cross-attention heatmaps extracted at each denoising step, weighted by the stepâ€™s influence on the final image structure as determined by an information-theoretic metric. This approach allows the system to produce

temporal explanations showing at which denoising step a cultural attribute (such as attire style) was fixed, and which prompt token exerted the strongest influence at that moment.

The multilingual setting introduces an additional layer of interpretability complexity because the same visual concept may be activated by different tokens across languages, and the strength of activation may vary due to the language-specific training data of the text encoder. The framework implements a cross-lingual attention alignment metric that quantifies, for a given generated image and a pair of languages, the discrepancy in attention distribution over image regions when the same semantic concept is prompted in each language. A high alignment discrepancy signals that the modelâ€™s internal representations are linguistically asymmetrical, which in turn may explain why a prompt in a low-resource language yields a culturally mismatched output. By surfacing this metric in audit reports, the framework allows developers to pinpoint whether culturally biased outputs are a consequence of the text encoderâ€™s deficiencies, the image decoderâ€™s biases, or the interaction between them.

Counterfactual explanation generation is extended beyond simple word substitution to structured, theory-driven interventions. Drawing on cultural psychology and anthropology literature, the framework maintains a taxonomy of cultural visual primitivesâ€”such as spatial distancing norms, color symbolism, and ritual objectsâ€”and systematically inserts or removes these primitives in prompts to observe output shifts. When a counterfactual alteration fails to produce the expected cultural shift, the explainability layer records this as a â€œresistant biasâ€ flag, indicating that the cultural association is deeply entrenched in the modelâ€™s weights and may require debiasing interventions beyond prompt engineering. Resistant bias flags are communicated directly to the governance interface, where they inform decisions about the modelâ€™s suitability for deployment in culturally sensitive domains such as educational content generation or public service announcements.

6. Multilingual and Cross-Cultural Adaptivity

Designing for multilingual contexts demands that the auditing framework treat language not as a mere input modality but as a lens that structures the modelâ€™s cultural world. Most current text-to-image models achieve multilingual capability by using a multilingual text encoder such as XLM-R or multilingual CLIP [16], which projects prompts in over a hundred languages into a shared embedding space. While this approach enables zero-shot cross-lingual generation, it also collapses linguistic variation in ways that can erase cultural nuance. The shared embedding space tends to be dominated by the language with the most training dataâ€”typically Englishâ€”and the geometric alignment algorithms used to map other languages into that space often prioritize semantic equivalence over cultural specificity. Consequently, a prompt translated perfectly from English to an Indigenous language may still produce an English-centric image because the multilingual encoder maps the Indigenous token to a location in the embedding space that is culturally neutralizedâ€”or, more accurately, English-culturated.

The framework mitigates this through a dedicated language-aware prompt disambiguation layer that operates before the prompt is fed into the text-to-image model. This layer detects cultural keywords and expands them with language-specific contextual metadata that is appended as invisible conditioning tokens, effectively steering the model without altering the surface prompt visible to the user. For example, when a prompt in Bengali contains a term for a traditional festival, the disambiguation layer appends a spatial descriptor referencing local architectural and clothing keywords, thereby partially offsetting the flattening effect of the multilingual encoder. This approach is transparently logged so that auditors can see exactly

which cultural steering signals were injected, maintaining accountability while improving representational accuracy. The effectiveness of disambiguation is then evaluated by the explainability layer, which compares the outputâ€™s cross-attention patterns and cultural anchor alignment scores with and without the contextual injection.

Adaptivity also requires that the framework itself be maintained by a globally distributed governance structure. Because cultural bias auditing involves normative judgments about what constitutes an appropriate representation, no single institution can legitimately define the standards for all cultural communities. The architecture therefore supports a federated auditing model where regional audit nodes, hosted by local civil society organizations or academic institutions, can extend the prompt database, define new cultural anchor sets, and contribute to threshold calibrations specific to their linguistic and cultural contexts. These regional nodes synchronize with a central registry via standardized APIs, ensuring comparability of audit results while respecting local autonomy. The governance interface aggregates results across nodes, flagging global patternsâ€”such as model versions that fail across multiple non-Western languages simultaneouslyâ€”while also preserving regional granularity for localized policy actions.

7. Governance, Ethics, and Policy Implications

The Explainable Cultural Bias Auditing Framework is designed not only to detect and explain bias but also to serve as a bridge between technical model development and the evolving regulatory landscape. The European Unionâ€™s Artificial Intelligence Act imposes transparency obligations on high-risk AI systems, mandating that deployers understand and document the limitations and potential biases of their systems [22]. Similarly, the NIST AI Risk Management Framework provides guidance on mapping, measuring, and managing AI risks, including those related to fairness and cultural representation [23]. The proposed framework operationalizes these principles by transforming internal model signals into auditable risk artifacts. Each audit report includes a structured risk statement that maps detected cultural biases to specific categories of harmâ€”such as erasure, stereotyping, and misrepresentationâ€”along with a confidence score derived from the consistency of explainability evidence and the cultural authority of the anchor sets used.

Ethically, the framework grapples with the tension between transparency and performativity. Making cultural bias explanations widely available could enable adversarial actors to fine-tune prompts that deliberately amplify biases or circumvent safety filters. To mitigate this risk, the governance interface implements a tiered transparency model: full explanation details are available to internal development teams and accredited auditors, while public-facing reports provide aggregated, privacy-preserving summaries that do not disclose the exact counterfactual probes or anchor images that could be exploited. This tiered approach aligns with the concept of coordinated vulnerability disclosure in cybersecurity and represents a pragmatic compromise between openness and the need to protect vulnerable communities from weaponized bias.

The intersection of cultural auditing and data sovereignty is another critical policy dimension. Many of the anchor images and cultural templates required for robust multilingual auditing may originate from Indigenous and marginalized communities that have historically been exploited by extractive research practices. The framework therefore embeds data governance protocols that require explicit community consent for the inclusion of culturally specific reference materials, with the option for communities to withdraw their data and trigger a re-audit. These protocols are enforced through smart contracts on a permissioned blockchain that

logs data lineage and usage, providing an immutable record that can be inspected by community representatives. While the computational overhead of blockchain logging is non-trivial, the governance benefits in terms of trust-building with marginalized communities outweigh the costs, and the system optimizes by batching and anchoring log entries rather than recording every granular event.

8. Challenges and Forward-Looking Perspectives

Despite its architectural comprehensiveness, the framework faces several open challenges that define directions for future research. The first is the dynamic nature of cultural bias itself: as generative models are updated with new data and fine-tuning procedures, previously mitigated biases may reappear, and new biases may emerge in previously unaffected languages. This necessitates lifelong learning mechanisms for the auditing system, where detection thresholds and anchor sets are continuously updated based on incoming model outputs and community feedback. Developing such mechanisms without introducing concept drift or overfitting to a particular temporal snapshot of culture is a non-trivial machine learning problem that will require collaboration between systems engineers and cultural theorists.

A second challenge concerns the scalability of cross-modal explainability across a wide variety of generative architectures. While the framework currently targets diffusion-based models, the emerging landscape of autoregressive image generators, masked image models, and hybrid architectures will demand explanation methods that can work across fundamentally different generation paradigms. Building an explainability abstraction layer that can provide comparable output across architectures while respecting the structural idiosyncrasies of each is an ambitious infrastructure goal that calls for standardized explanation interchange formats, akin to model cards but for interpretability outputs. The framework anticipates this by decoupling the explainability layer from specific model implementations through a set of abstract interfaces that future models must implement to plug into the auditing ecosystem.

Finally, the geopolitical dimension of multilingual cultural auditing cannot be ignored. As text-to-image models are deployed globally, they will increasingly be subject to conflicting cultural norms and regulatory expectations. An image that is considered perfectly acceptable and representative in one jurisdiction may violate cultural decency laws in another. The auditing framework must therefore support jurisdiction-specific policy filters that can apply different cultural sensitivity thresholds based on the geographic context of deployment. This transforms the governance interface into a geopolitically aware system that must reconcile the universalizing tendencies of deep learning architectures with the fragmented, yet legitimate, regulatory pluralism of the international order.

9. Conclusion

The integration of text-to-image generative models into the fabric of global digital infrastructure demands a rethinking of how cultural bias is audited, explained, and governed. This paper has presented an Explainable Cultural Bias Auditing Framework that foregrounds multilingualism, cross-modal explainability, and systemic infrastructure design. By linking multilingual prompt curation, bias signature detection, gradient-based and attention-driven explanation, and policy-responsive governance, the framework offers a holistic architecture that moves beyond narrow fairness metrics. The structural trade-offs analyzedâ€”between depth and cost, cultural specificity and generalization, transparency and security, and data

sovereignty and global interoperabilityâ€”define the contours of a responsible auditing practice that is as much about infrastructure choices as it is about algorithmic design. As generative models continue to shape visual culture on a planetary scale, auditing frameworks must evolve into participatory, explainable, and culturally grounded institutions capable of holding these powerful systems accountable to the diverse publics they serve.

References

1. 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.

2. 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 (pp. 220-229). ACM.

3. Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. Transactions of the Association for Computational Linguistics, 6, 587-604.

4. Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In Advances in Neural Information Processing Systems (pp. 4349-4357).

5. 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 Conference on Empirical Methods in Natural Language Processing (pp. 2979-2989). ACL.

6. Sheng, E., Chang, K. W., Natarajan, P., & Peng, N. (2019). The woman worked as a babysitter: On biases in language generation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 3405-3410). ACL.

7. Nadeem, M., Bethke, A., & Reddy, S. (2021). StereoSet: Measuring stereotypical bias in pretrained language models. In Proceedings of the Association for Computational Linguistics (pp. 5356-5371). ACL.

8. 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). IEEE.

9. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., ... & Sutskever, I. (2021). Zero-shot text-to-image generation. In International Conference on Machine Learning (pp. 8821-8831). PMLR.

10. 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.

11. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., ... & Chen, M. (2021). GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In International Conference on Machine Learning (pp. 16784-16804). PMLR.

12. Luccioni, A. S., Akkaya, C., Jernite, Y., & others. (2023). Stable Bias: Analyzing societal representations in diffusion models. arXiv preprint arXiv:2303.11408.

13. Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., ... & Caliskan, A. (2023). Easily accessible text-to-image generation amplifies demographic stereotypes. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (pp. 1493-1504). ACM.

14. Cho, J., Zala, A., Bansal, M., Shah, D., & Darrell, T. (2023). DALL-Eval: Probing the reasoning skills and social biases of text-to-image generation models. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. IEEE.

15. Wang, Z., et al. (2023). How well can text-to-image generative models understand ethical natural language interventions? In Proceedings of the Conference on Empirical Methods in Natural Language Processing. ACL.

16. Carlsson, F., Eisen, P., Rekath, N., & Sahlgren, M. (2022). Cross-lingual and multilingual CLIP. In Proceedings of the Language Resources and Evaluation Conference (pp. 6848-6854). ELRA.

17. 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.

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

19. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (pp. 4765-4774).

20. 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). IEEE.

21. Tang, R., Pandey, R., Qi, Z., & Saunshi, N. (2023). What the DAAM: Interpreting Stable Diffusion Using Cross Attention. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). ACL.

22. European Commission. (2021). Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). COM(2021) 206 final.

23. National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0). NIST.

24. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3645-3650). ACL.

25. Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning. fairmlbook.org.