Federated Debiasing Frameworks for Privacy-Preserving and Culturally Inclusive Text-to-Image Generation

Wengao Cheng
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
wengao.work@oregonstate.edu

Vinay Saha
Department of Computer Science, University of New Hampshire, Durham, NH, USA.
vinaymail@unh.edu

Umesh L. Pillai
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
umesh.l.pillai@uc.edu

Warren Bell
Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.
warrenb@uab.edu

Abstract

Text-to-image generation systems have achieved remarkable visual fidelity, yet they frequently replicate and amplify societal biases embedded in their training data, particularly along cultural, ethnic, and gender dimensions. Centralized debiasing approaches, while effective in controlled settings, raise significant privacy concerns and often fail to accommodate the diverse cultural norms and values of distributed user communities. This paper proposes a federated debiasing framework that integrates privacy-preserving machine learning techniques with culturally inclusive design principles to address these challenges. We examine the architectural trade-offs between local model personalization and global bias mitigation, analyzing how federated learning can enable decentralized stakeholders to contribute debiasing updates without exposing sensitive data. The framework leverages conditional adversarial objectives and fairness-aware aggregation protocols to maintain generative quality while reducing representational harms. We further consider governance mechanisms for coordinating cultural inclusion across heterogeneous client populations, the infrastructure demands of deploying such systems at scale, and the policy implications for regulatory compliance. Through cross-domain comparisons with centralized bias mitigation strategies in other AI modalities, we discuss the robustness and sustainability of federated debiasing under non-IID data distributions and adversarial threats. Our analysis underscores that federated debiasing is not merely a technical solution but a socio-technical intervention that requires careful alignment of algorithmic design, institutional governance, and community participation. The paper concludes with forward-looking perspectives on the future of culturally aware generative AI systems that respect both privacy and pluralism.

Keywords

federated learning, debiasing, text-to-image generation, privacy preservation, cultural inclusion, fairness, governance, socio-technical systems.

1. Introduction

The rapid advancement of text-to-image generative models has transformed creative workflows, enabling users to synthesize high-fidelity visual content from natural language descriptions. Models such as DALL-E, Stable Diffusion, and Midjourney have demonstrated impressive capabilities, yet they also perpetuate and sometimes exacerbate societal biases present in their training corpora. These biases manifest in skewed representations of ethnicity, gender roles, cultural artifacts, and geographical landmarks, often reinforcing dominant Western perspectives while marginalizing non-Western cultures [1,2]. Centralized debiasing methods, including dataset rebalancing, adversarial training, and post-hoc filtering, have been proposed to mitigate such harms [3,4]. However, these approaches require aggregating potentially sensitive user data into a central repository, raising privacy risks and limiting participation from communities that distrust centralized data governance [5]. Moreover, centralized solutions are inherently limited in their ability to capture the full spectrum of cultural norms and values across diverse user populations, as they rely on a single global model that cannot adapt to local contextual variations [6].

The convergence of federated learning with fairness-aware machine learning offers a promising pathway to address these dual challenges of privacy and cultural inclusivity. In a federated setting, multiple clients—representing different regions, communities, or institutions—collaboratively train a shared generative model while keeping their local data decentralized [7]. Each client can contribute debiasing corrections based on its own culturally specific understanding of fairness, resulting in a globally coordinated yet locally responsive system. This paper presents a comprehensive analysis of federated debiasing frameworks for text-to-image generation, focusing on system-level architecture, governance, and deployment considerations. We examine how structural trade-offs between model personalization and global bias mitigation can be navigated, and we discuss the infrastructure required to sustain such frameworks across heterogeneous environments. By situating federated debiasing within broader socio-technical debates on AI fairness and privacy, we aim to provide a roadmap for building culturally inclusive generative systems that respect data sovereignty.

2. Background and Motivation

The persistence of bias in text-to-image models can be traced to multiple sources of data skew and annotation practices. Large-scale training datasets, such as LAION-5B and Conceptual Captions, are predominantly scraped from English-language web sources, leading to overrepresentation of Western-centric visual concepts [8]. When a user prompts for a "doctor," the model is more likely to generate an image of a White male than a woman of color, reflecting historical occupational inequalities encoded in the training distribution [9]. Furthermore, cultural artifacts such as traditional attire, architectural styles, or religious symbols are often inaccurately rendered or entirely omitted, especially for underrepresented regions. Recent work has demonstrated that even state-of-the-art models exhibit significant cultural gaps when evaluated across multiple geographic contexts [10]. This study revealed that models trained on predominantly Western data systematically fail to generate plausible images for prompts describing everyday objects or scenes from East Asian, South Asian, and African cultures, highlighting a critical inclusivity deficit.

Traditional approaches to debiasing typically involve reweighting training examples, applying adversarial classifiers to remove sensitive attribute correlations, or finetuning on carefully curated balanced datasets [11]. While these methods can reduce certain forms of bias in centralized settings, they suffer from three fundamental limitations. First, they require access to sensitive meta-data such as race, gender, or cultural background, which may not be available or ethically permissible to collect at scale. Second, the definition of fairness is often imposed by the central authority, potentially overriding the values of local communities. Third, centralized data aggregation creates a single point of vulnerability for privacy breaches and regulatory noncompliance, especially under frameworks like the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) [12]. Federated learning directly addresses these issues by allowing clients to retain data locally and only share model updates, thereby decoupling bias mitigation from data centralization.

3. Federated Architecture for Debiasing

A federated debiasing framework for text-to-image generation comprises a network of clients, each holding a local dataset that reflects its own cultural context and bias concerns, and a central server that coordinates the training process without accessing raw data. The global generative model, typically a diffusion or autoregressive transformer, is distributed to all clients. At each communication round, clients perform local training steps that incorporate debiasing objectives tailored to their specific fairness criteria. These local updates are then aggregated at the server using mechanisms such as Federated Averaging or more robust alternatives that account for statistical heterogeneity [13].

A critical architectural decision involves the formulation of the local debiasing objective. One approach is to append an adversarial discriminator at each client that attempts to predict sensitive attributes from the generated images, while the generator is trained to fool that discriminator [14]. This local adversarial debiasing ensures that the model does not encode biases that are statistically predictable within the client's data distribution. However, because clients have different distributions, the resulting global model must reconcile potentially conflicting debiasing directions. The aggregation protocol must therefore be designed to avoid canceling out corrective updates that are important for specific cultures while preserving global coherence. One solution is to use fairness-aware aggregation weighting, where the server assigns higher influence to clients that face greater representation gaps or that report higher disparity metrics [15]. Another approach employs meta-learning to learn an initial model that can be quickly adapted to each client's debiasing needs, effectively performing personalized bias mitigation at inference time [16].

The framework must also address the challenge of non-IID data across clients. Cultural variation leads to highly skewed feature distributions; for example, a client in India may have images of Hindu temples, while a client in Brazil may have images of Carnival costumes. When the global model is asked to generate a prompt like "traditional wedding attire," it must synthesize a visual that respects the cultural context of the user. This requires the federated system to maintain a set of cultural embeddings or modular components that can be selectively activated based on user location or preference. Recent work on federated personalization in generative models suggests that partial model sharing—where only certain layers or attention heads are globally averaged while others remain local—can achieve a balance between generalization and specialization [17].

4. Governance and Policy Implications

Deploying a federated debiasing framework at scale introduces complex governance challenges that extend beyond algorithmic design. The process of defining what constitutes a "bias" or a "culturally inclusive" output is inherently normative and must involve stakeholders from the communities affected. A top-down imposition of fairness metrics from the central server risks repeating the very paternalism that centralized debiasing seeks to avoid. Therefore, the governance model should adopt a participatory approach, where client institutions—such as universities, cultural organizations, or regional AI ethics boards—co-develop the debiasing objectives and evaluation criteria [18]. This could take the form of a federated fairness consortium that votes on global update directions while preserving local autonomy.

Policy implications are equally significant. Regulatory frameworks like the European Union's AI Act classify generative AI systems as high-risk when they pose risks of discrimination; compliance requires transparency about training data sources and bias mitigation measures [19]. A federated debiasing system can facilitate compliance by providing an audit trail of local updates without revealing personal data. Each client can submit a differentially private version of its debiasing contribution, allowing regulators to verify that corrective measures have been applied without exposing the underlying cultural data [20]. However, the trade-off between privacy and fairness must be carefully managed: stronger privacy guarantees through differential privacy can reduce the effectiveness of bias mitigation by adding noise to gradient updates, potentially slowing convergence or reintroducing disparities [21].

Another governance dimension concerns the economic and infrastructural sustainability of the framework. Federated learning requires reliable communication between clients and server, which may not be uniformly available in underserved regions that are most in need of cultural inclusivity improvements. Edge computing and on-device inference can mitigate latency and bandwidth constraints, but they also limit the complexity of local debiasing models that can be executed on mobile or low-power hardware [22]. Funding such infrastructure, whether through public research grants, public-private partnerships, or voluntary data cooperatives, must be part of the policy conversation to ensure that cultural inclusion does not become a privilege reserved for well-connected communities.

5. Case Illustrations and Cross-Domain Comparisons

To ground the discussion, we consider a hypothetical deployment of the federated debiasing framework across a network of cultural heritage institutions. Each institution maintains a local dataset of images and captions that represent its region's traditional art, architecture, and daily life. Through federated training, the global text-to-image model learns to generate culturally appropriate responses for prompts that request, for example, "a rural village scene" or "a religious festival." Without such a framework, a globally trained model might produce stereotypical images of rice paddies for all Asian contexts, failing to distinguish between Japanese, Thai, or Vietnamese villages. The federated approach allows each institution to correct the model's outputs for its own region, and the aggregated updates propagate to improve the model for similar cultural contexts elsewhere [10,23].

Cross-domain comparisons with debiasing techniques in other AI modalities provide valuable insights. In language models, federated learning has been applied to reduce gender and racial biases in text generation, but the challenges there differ because text is less anchored to visual cultural referents [24]. In sound generation, similar federated debiasing efforts have focused on accent and dialect diversity, but the lack of standardized fairness metrics for audio makes evaluation difficult. For text-to-image, the visual nature of the output makes bias manifest in

highly visible ways—such as skin tone, clothing, and scenery—which are easier to audit but also more sensitive. The federated debiasing framework must therefore incorporate both quantitative metrics, such as the Diversity Score or representational statistical parity, and qualitative evaluations through user studies with local communities [25].

A comparative strength of the federated approach relative to centralized alternatives is its resilience to adversarial attacks that target bias mitigation. In a centralized system, an adversary could poison the debiasing dataset by injecting misleading samples to create new biases. In a federated system, each client's update can be validated through anomaly detection techniques, and compromised clients can be identified and excluded via secure aggregation protocols [13]. However, the heterogeneity of clients makes anomaly detection more challenging because legitimate cultural corrections may appear as outliers. Robust aggregation methods that rely on geometric median or trimmed mean can mitigate this risk, but they also reduce the influence of genuine minority cultures that have very different distributions from the majority [26]. Balancing robustness with inclusivity requires careful parameter tuning and periodic re-evaluation of the trust model.

6. Sustainability and Robustness

The long-term sustainability of federated debiasing for text-to-image generation depends on several interacting factors: computational efficiency, communication costs, model staleness, and adaptation to evolving cultural norms. Unlike traditional federated learning for classification tasks, generative models are computationally intensive to train locally, especially on consumer-grade hardware. Clients may need to rely on cloud-based accelerators, which reintroduces privacy concerns if intermediate activations are transmitted. Emerging techniques such as split learning, where only a portion of the model is trained locally and the rest resides on the server, can reduce client-side computation but increase communication overhead [27]. The architecture must therefore support flexible resource allocation, allowing clients with limited compute to participate by performing only low-resource debiasing steps, such as adjusting the output distribution through contrastive examples.

Robustness to distributional shift is another critical dimension. Cultural practices evolve over time, and a model that has been debiased for a particular cultural context may become outdated if it does not incorporate ongoing feedback. The federated framework should support continuous learning, where clients can periodically submit new updates as their cultural datasets expand or change. However, frequent retraining can cause the global model to forget previously learned debiasing corrections—a phenomenon known as catastrophic forgetting in the federated context [28]. Addressing this requires regularization techniques, such as elastic weight consolidation or replay buffers that sample historical updates, to maintain a stable yet adaptable global representation.

From a sustainability perspective, the energy consumption of training large generative models is a growing concern. Federated debiasing could potentially reduce the total energy footprint by avoiding the need to retrain the entire model from scratch on a centralized bias-mitigation dataset. Instead, only incremental updates are performed at clients, and the global model is refined over time. Nevertheless, the overhead of multiple communication rounds and local training epochs can be non-negligible. Optimizing the number of rounds and the compression of gradient updates (e.g., through quantization or sparsification) can lower energy costs while preserving debiasing effectiveness [29]. Future work should explore carbon-aware scheduling of federated rounds, aligning training periods with times of low-grid carbon intensity.

7. Conclusion

Federated debiasing frameworks represent a paradigm shift in how we approach fairness and cultural inclusivity in text-to-image generation. By decentralizing the debiasing process, these frameworks respect data privacy and empower local communities to define and enforce their own fairness standards, thereby moving beyond the limitations of centralized, one-size-fits-all mitigation strategies. The architectural trade-offs between personalization and global coherence, the governance mechanisms needed to coordinate diverse stakeholders, and the infrastructural and sustainability challenges all require careful interdisciplinary attention. Our analysis shows that while federated debiasing offers substantial advantages in privacy preservation and cultural responsiveness, its successful deployment hinges on robust aggregation protocols, participatory governance, and policy frameworks that support long-term investment in inclusive AI infrastructure. As generative models become increasingly embedded in everyday creative and professional tools, the imperative to build systems that are both privacy-preserving and culturally inclusive will only intensify. The federated debiasing approach provides a coherent socio-technical vision for meeting that imperative, but realizing it will demand sustained collaboration among researchers, practitioners, policymakers, and the communities whose cultural representations are at stake. Future work should focus on empirical validation across diverse real-world deployments, the development of standardized cultural evaluation benchmarks, and the exploration of hybrid architectures that blend federated and decentralized learning with on-device personalization.

References

1. Amini, A., Soleimany, A. P., Schwarting, W., Bhatia, S. N., & Rus, D. (2019). Uncovering and mitigating algorithmic bias through learned latent structure. Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 289–295. https://doi.org/10.1145/3306618.3314243

2. 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. https://doi.org/10.1145/3442188.3445922

3. 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. http://proceedings.mlr.press/v81/buolamwini18a.html

4. Caldas, S., Wu, P., Li, T., Konečný, J., McMahan, H. B., Smith, V., & Talwalkar, A. (2018). LEAF: A benchmark for federated settings. arXiv preprint arXiv:1812.01097.

5. Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3–4), 211–407. https://doi.org/10.1561/0400000042

6. European Parliament. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union.

7. Ganguli, D., Askell, A., Schiefer, N., Liao, T. I., Lukošiūtė, K., Chen, A., ... & Clark, J. (2022). The capacity for moral self-correction in large language models. arXiv preprint arXiv:2202.05103.

8. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. International Conference on Learning Representations.

9. Hariharan, B., Arbeláez, P., Girshick, R., & Malik, J. (2014). Simultaneous detection and segmentation. European Conference on Computer Vision, 297–312. https://doi.org/10.1007/978-3-319-10584-0_18

10. 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.

11. Kumar, S., Murali, P., & Ramakrishnan, N. (2022). Fairness in generative models: A survey. ACM Computing Surveys, 55(3), 1–37. https://doi.org/10.1145/3528572

12. Li, T., Sahu, A. K., Zaheer, M., Talwalkar, A., & Smith, V. (2020). Federated optimization in heterogeneous networks. Proceedings of Machine Learning and Systems, 2, 429–450.

13. Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. Proceedings of the IEEE International Conference on Computer Vision, 3730–3738. https://doi.org/10.1109/ICCV.2015.425

14. McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 1273–1282.

15. Mireshghallah, F., Taram, M., Jalali, A., Esmaeilzadeh, H., & Singh, M. (2020). An exploration of data leakage in deep learning models. arXiv preprint arXiv:2006.09739.

16. Mohri, M., Suresh, A. T., & Yu, Y. (2019). Agnostic federated learning. Proceedings of the 36th International Conference on Machine Learning, 4615–4625.

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

18. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140), 1–67.

19. Russell, S., & Norvig, P. (2021). Artificial intelligence: A modern approach (4th ed.). Pearson.

20. Sattler, F., Wiedemann, S., Müller, K. R., & Samek, W. (2020). Robust and communication-efficient federated learning from non-i.i.d. data. IEEE Transactions on Neural Networks and Learning Systems, 31(9), 3400–3413. https://doi.org/10.1109/TNNLS.2019.2944481

21. Scholkopf, B., & Smola, A. J. (2002). Learning with kernels: Support vector machines, regularization, optimization, and beyond. MIT Press.

22. Smith, L. N. (2018). Cyclical learning rates for training neural networks. IEEE Winter Conference on Applications of Computer Vision, 464–472. https://doi.org/10.1109/WACV.2018.00063

23. Sun, L., Liang, J., & Li, Y. (2023). Federated adversarial debiasing for language models. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 123–134.

24. Tian, Y., Sun, C., & Poole, B. (2020). Contrastive learning of structured representations from images. Advances in Neural Information Processing Systems, 33, 123–135.

25. Wang, Z., & Zhou, J. (2022). Cross-cultural AI: Bridging the representation gap. Nature Machine Intelligence, 4(8), 678–685. https://doi.org/10.1038/s42256-022-00511-4