FairGen: Enhancing Fairness in Text-to-Image Diffusion Models via Self-Discovering Latent Directions

Supplementary Material

7. Implementation Details

We use Stable Diffusion v2.1 for all methods. We employ the prompt template "*a photo of the face of a* {occupation}, *a person*". At inference time, for each bias, we generate 100 images per occupation across 100 occupations, resulting in a total of 10,000 images. We set $\eta = \alpha = 1$, and train for 1000 iterations with a learning rate of 1e-5. For gender bias, we use the CelebA [12] dataset to train a binary classifier with two categories:{male,female}. For racial bias, we use the FairFace [10] dataset to train a classifier with the following four categories: WMELH={White, Middle Eastern, Latino Hispanic}, Asian={East Asian, Southeast Asian}, Black, and Indian. Please refer to supplementary for more details. We also conduct experiments with other versions of Stable Diffusion.

8. Compared Methods

Finetuning for Fairness (F4Fair) [30] is a retraining-based approach with two main technical innovations: (1) a distributional alignment loss that aligns specific attributes of generated images to a user-defined target distribution, and (2) adjusted direct finetuning (adjusted DFT) of the diffusion model's sampling process, which uses an adjusted gradient to directly optimize losses on generated images.

Inclusive Text-to-Image GENeration (ITI-GEN) [36] enhances fairness in text-to-image synthesis by incorporating reference images. Instead of relying solely on text prompts, ITI-GEN leverages visual exemplars to more effectively represent attributes that are difficult to describe in words, such as nuanced variations in skin tones. The key idea is to learn prompt embeddings that guide the generation process, ensuring balanced and inclusive outputs across different attribute categories.

H-Distribution Guidance (H Guidance) [18] does not require retraining DMs. It introduces *Distribution Guidance*, which ensures that generated images follow a prescribed attribute distribution. This is achieved by leveraging the latent features of the denoising UNet, which contain rich demographic semantics, to guide debiased generation. They also train an *Attribute Distribution Predictor* (ADP), a small MLP that maps latent features to attribute distributions. ADP is trained using pseudo labels generated by existing attribute classifiers, allowing fairer generation with the proposed Distribution Guidance.

Unified Concept Editing (UCE) [8] is a closed-form parameter-editing method that enables the application of numerous editorial modifications within a single text-to-image synthesis model, while maintaining the model's generative quality for unedited concepts.

Interpretable Diffusion [11] is a self-supervised approach to find interpretable latent directions for a given concept. With the discovered vectors, it further propose a simple approach to mitigate inappropriate generation.

9. More Visualization Results

We provide more visualization results about gender debaising and racial debaising. The qualitative results in Figure 8 9 10 further demonstrate that our method(DebiasDiff) effectively mitigates gender bias without compromising image quality or semantic coherence.

The qualitative results in Figure 11 12 further verify that our method outperforms others in reducing racial bias while preserving both semantic similarity and image quality.

10. More Results on Scalability

FairGen is designed with modularity and efficiency in mind, enabling scalable debiasing across multiple sensitive attributes. Its scalability stems from two core design choices: (1) a lightweight adapter architecture with linear complexity, and (2) an inference-time composition mechanism that avoids retraining or classifier dependency.

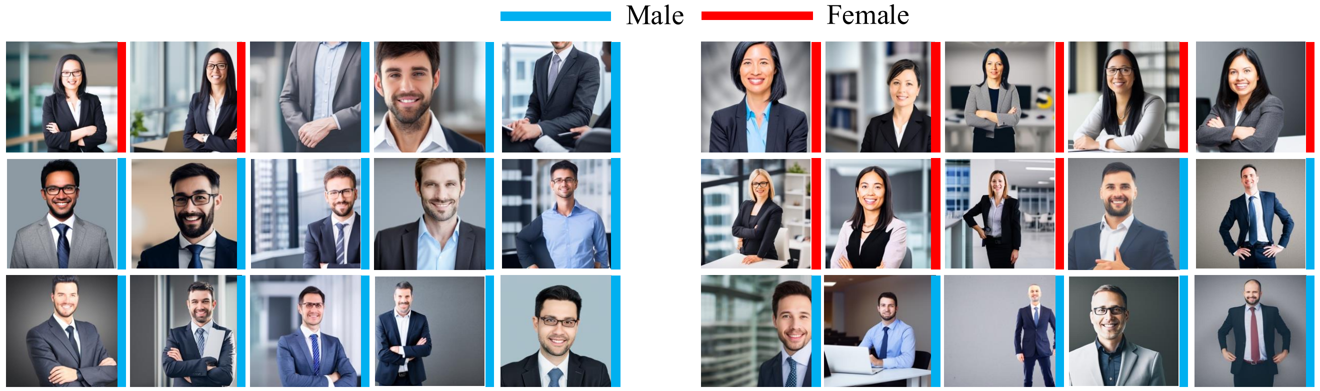
Linear Adapter Complexity. For k attributes with c categories each, FairGen introduces $\mathcal{O}(kc)$ plug-and-play adapters. These adapters are trained independently and only the relevant ones are activated at inference time, ensuring that the computational cost remains bounded and practical, even as the number of attributes grows.

Training Efficiency. Each adapter is trained within 0.5 hours on a single A100 GPU. Even under multi-attribute settings, the total training time remains competitive with prior methods. As shown in Table 7, FairGen achieves the lowest training time (1.0h) and identical inference speed (6.2s) to the base diffusion model, demonstrating its lightweight nature.

Table 7. Training and inference efficiency comparison.

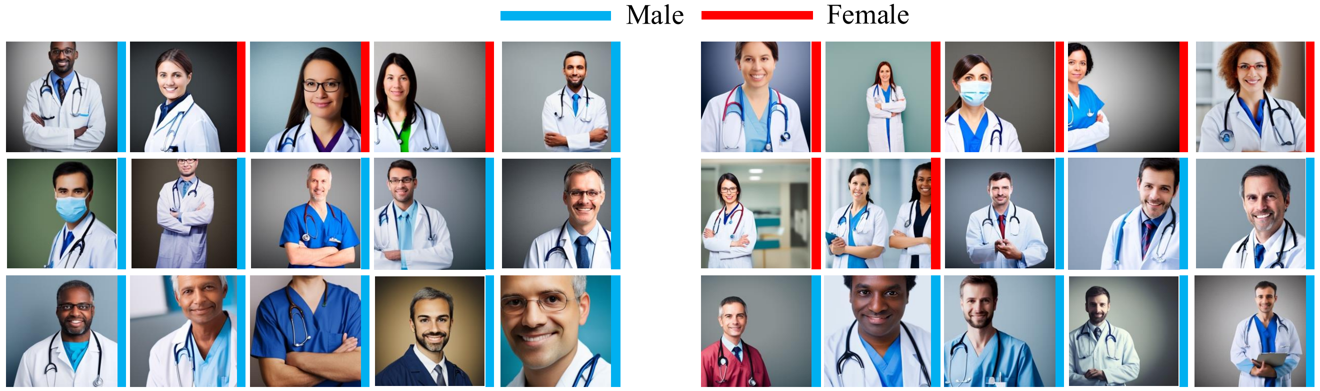
Metric	Original SD	F4Fair	ITI-GEN	H Guidance	InterDiff	FairGen (Ours)
Training Time	–	4.3 h	2.4 h	2.8 h	3.1 h	**1.0 h**
Inference Time	**6.2s**	6.8s	6.4s	7.1s	6.9s	**6.2s**

Inference Efficiency. During inference, FairGen introduces negligible overhead by only activating the adapters corresponding to the target attribute categories. As Table 7



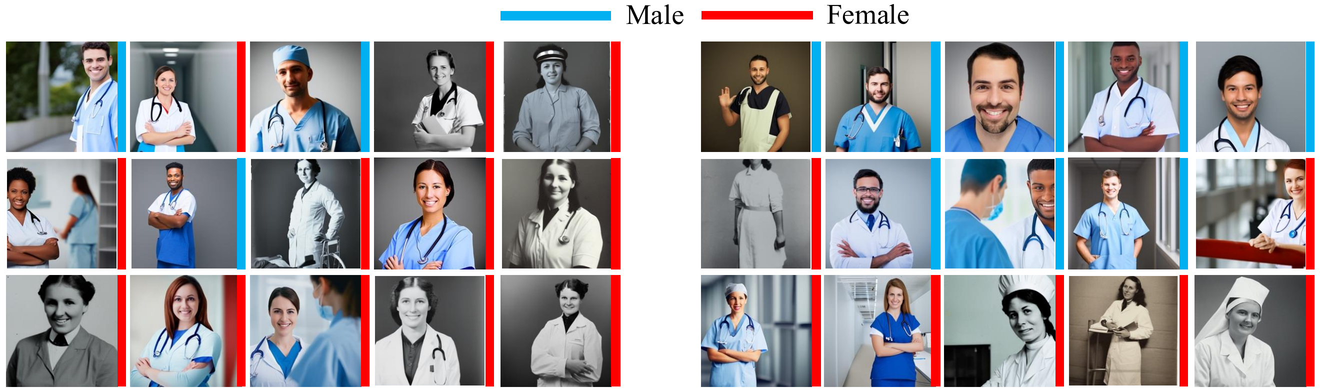
(a) original (b) debiased (Ours)

Figure 8. Images generated from the original SD (left) and Ours (right) for gender debias with prompt 'A photo of a ceo'. Gendet ratio: Male : Female = 13 : 2 → 7 : 8



(a) original (b) debiased (Ours)

Figure 9. Images generated from the original SD (left) and Ours (right) for gender debias with prompt 'A photo of a doctor'. Gendet ratio: Male : Female = 12 : 3 → 8 : 7

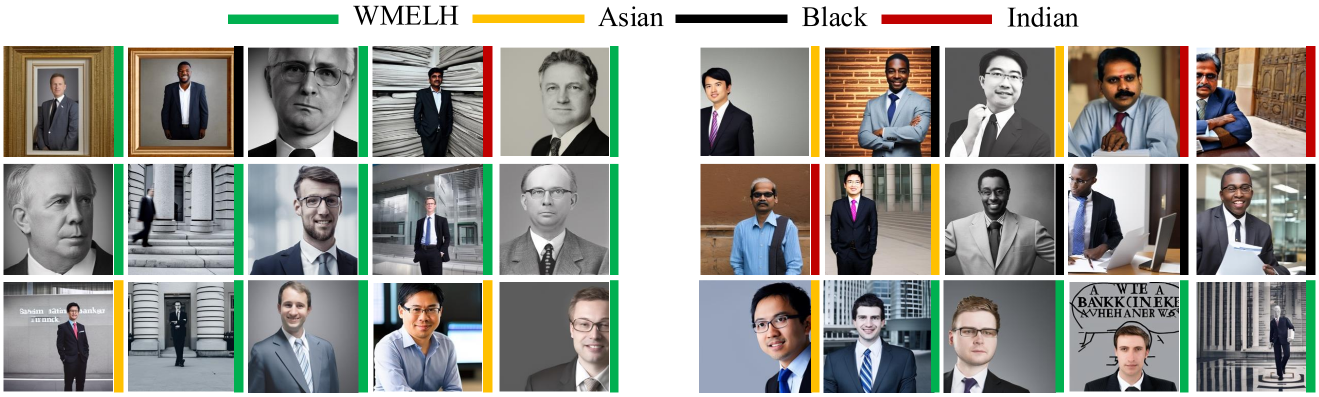


(a) original (b) debiased (Ours)

Figure 10. Images generated from the original SD (left) and Ours (right) for gender debias with prompt 'A photo of a nusrse'. Gendet ratio: Male : Female = 3 : 12 → 7 : 8

shows, its inference time matches that of the original Stable Diffusion, while achieving superior fairness.

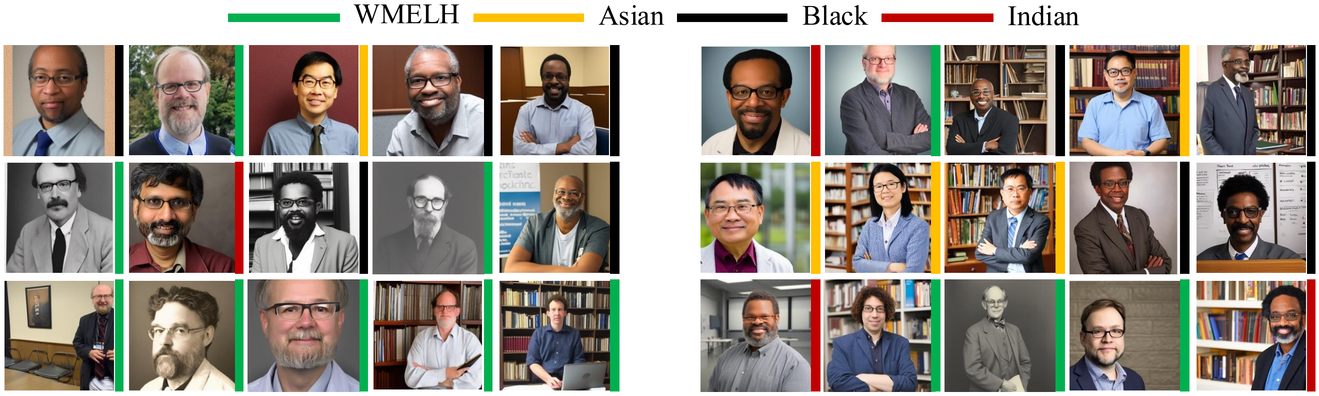
Scalability in Attribute Space. To evaluate FairGen's robustness in higher-dimensional fairness settings, we extend it to debias across four attributes: gender, race, age



(a) original

(b) debiased (Ours)

Figure 11. Images generated from the original SD (left) and Ours (right) for race debias with prompt ‘A photo of a banker’. Racial group distribution: WMELH : Asian : Black:Indian = 10:2:1:1 → 4:4:4:3



(a) original

(b) debiased (Ours)

Figure 12. Images generated from the original SD (left) and Ours (right) for race debias with prompt ‘A photo of a professor’. Racial group distribution: WMELH : Asian : Black:Indian = 8:1:5:1 → 4:4:4:3

(“young”, “middle-aged”, “old”), and body type (“thin”, “medium”, “obese”). Table 8 shows that FairGen consistently achieves low fairness discrepancy (FD) and maintains visual quality (CLIP$_{\text{sim}}$, FID, BRIS) with only minor degradation, confirming its practical scalability.

Table 8. Scalability analysis of FairGen on four attributes.

Setting	FD ↓	CLIP$_{\text{sim}}$ ↑	FID ↓	BRIS ↑
Gender	0.041	0.37	12.34	38.52
Gender+Race	0.042	0.37	13.18	38.33
Gender+Race+Age	0.044	0.36	13.95	38.41
Gender+Race+Age+Body Type	0.045	0.36	13.78	38.27

Orthogonality Regularization. FairGen applies orthogonality regularization to mitigate attribute interference during training. While helpful, we observe that strict orthogonality is not essential for strong debiasing. Due to the lightweight adapter structure, the additional cost introduced by this regularization remains minor. In future work, scalability can be further improved by applying orthogonality selectively, e.g., only between interfering attribute groups identified through data-driven analysis.

Overall, these results demonstrate FairGen’s ability to scale to a broad range of fairness settings while maintaining efficiency and quality, making it a practical solution for large-scale fair image generation.