Identifying and Addressing Bias in AI Face Generators

Neil Nie, Hannah Norman

Stanford | ENGINEERING
Computer Science

Problem

Generative models like StyleGAN-2 [1], Stable Diffusion [2], and FaceDiffusion [3] enable high-quality face synthesis for applications like virtual try-ons and media editing. But these models are often trained on skewed datasets (e.g., FFHQ: ~83% light-skinned), resulting in biased outputs. We build a balanced benchmark from FairFace v1.3 across 14 race-gender groups and ages 0–80, and evaluate the three generators using FID, LPIPS, & FaceNet similarity. Our results reveal clear demographic gaps—for instance, dark-skinned women consistently see the worst generation quality. These findings underscore the need for fairer training and evaluation. Ongoing work explores dataset rebalancing & fairness-aware fine-tuning.

Methods

- We evaluate three face generators (StyleGAN2-ADA, Stable Diffusion v1.5, and FaceDiffusion) on a balanced FairFace v1.3 benchmark [4] (7,000 images across 14 race-gender groups).
- Each model generates 1,000 faces. Diffusion models use group-specific prompts; unconditional outputs are labeled post hoc using a FairFace ResNet-34 classifier.

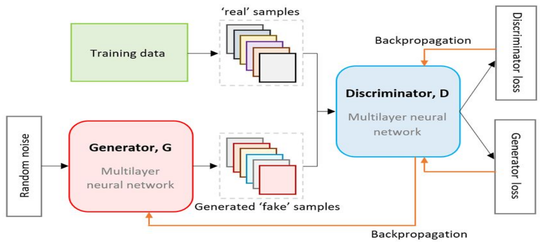

Figure 1. Generative Adversarial Network (GAN) [5]

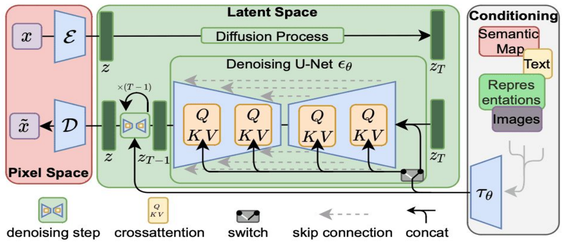

Figure 2. Conditional Diffusion Model [6]

Conclusions

Our project successfully addresses two key questions on bias in face-generation models:

- **Measurement.** The performance gap in fidelity and identity preservation across gender, skin tone, and age, and how does it differ between top generative architectures?
  - Quality gaps exist across demographics, worst for darker-skinned and female faces. Stable Diffusion shows the most consistent performance.
- **Mitigation.** Which simple, low-compute method can best reduce this gap without retraining, while keeping inference speed and visual quality intact?
  - Prompt balancing and targeted sampling reduce disparities without retraining, preserving speed and visual quality. We are still exploring to how leverage lightweight post-training to reduce bias.

Experiments

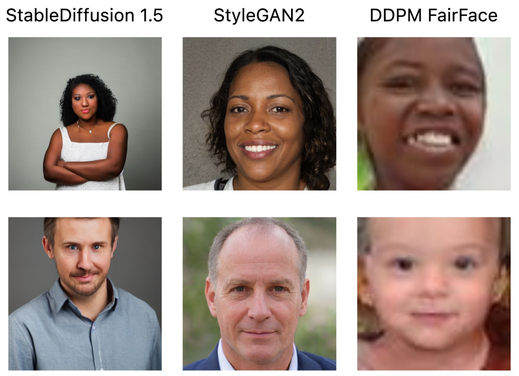

Figure 3. Images generated by the three models for two demographic groups.

Outputs are grouped by demographic category, and evaluated using two metrics:

- FID for image fidelity
- LPIPS for perceptual similarity to real images

To ensure fairness evaluation, we compute ΔFID: the gap between best- and worst-performing subgroups.

Across evaluated demographic biases, StableDiffusion 1.5 maintains high quality, while StyleGAN2 and FaceDiffusion show performance drops, especially for Southeast Asian and Black faces.

Results

Race	SD15 FID	SD15 LPIPS	StyleGAN2-ADA FID	StyleGAN2-ADA LPIPS	FaceDiffusion FID	FaceDiffusion LPIPS
Black Female	260.19	20.85	285.58	0.79	286.00	10.31
East Asian Female	211.03	20.09	235.38	3.47	302.33	4.67
East Asian Male	215.55	20.28	256.79	9.00	296.50	2.70
Southeast Asian Female	208.44	20.12	284.41	0.77	320.67	1.75
Southeast Asian Male	214.04	20.28	311.29	0.79	303.57	7.82
White Male	213.43	20.21	219.48	7.56	251.71	26.52

Table 1: Per-race FID and LPIPS (lower is better) for face images generated by three models.

References

[1] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735.
[2] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. *arXiv preprint arXiv:2112.10752*.
[3] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. *arXiv preprint arXiv:2006.11239*.
[4] Karkkainen, K., & Joo, J. (2021). FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 1548–1558).
[5] Little, Claire & Elliot, Mark & Allmendinger, Richard & Samani, Sahel. (2021). Generative Adversarial Networks for Synthetic Data Generation: A Comparative Study. 10.48550/arXiv.2112.01925.
[6] Tam, A. (2024, July 18). Brief Introduction to Diffusion Models for Image Generation. Machine Learning Mastery. https://machinelearningmastery.com/brief-introduction-to-diffusion-models-for-image-generation/