Identifying and Addressing Bias in AI Face Generators

Neil Nie
Department of Computer Science
Stanford University
neilnie@stanford.edu

Hannah Norman
Department of Computer Science
Stanford University
hnorman@stanford.edu

Abstract

We present a systematic benchmark of demographic bias in three widely-used face generators—StyleGAN2-ADA, Stable Diffusion v1.5, and FaceDiffusion—using a balanced benchmark derived from FairFace v1.3. Our evaluation spans all 6 (*race*, *gender*) combinations and quantifies performance disparities using Fréchet Inception Distance (FID), LPIPS perceptual similarity, and FaceNet-based identity alignment. We find substantial quality gaps, with dark-skinned and Southeast Asian faces consistently under-served across all models. To mitigate these disparities, we introduce a lightweight rejection sampling scheme that filters poor or mismatched outputs from Stable Diffusion v1.5 based on CLIP similarity and demographic classification. This simple post-processing step improves fidelity and diversity for marginalized subgroups, cutting FID by up to 33 points without any model retraining. Our benchmark, code, and findings offer a reproducible foundation for future fairness-driven interventions in generative face modeling.

1 Introduction

Breakthroughs in generative adversarial networks (GANs) such as *StyleGAN2* and advancements in diffusion models (e.g. *Stable Diffusion* and *FaceDiffusion*) have enabled easy-to-use and photo-realistic face synthesis across many applications and domains.

Today's image filters can "de-age" actors during post-production, game engines can generate character models at scale, and virtual try-on mirrors can let online shoppers preview cosmetics in real time—all by sampling these models. However, the core training corpora for generative models remain heavily imbalanced (FFHQ $\approx 83\%$ light-skinned; CelebA $\approx 75\%$ male). Consequently, the same pipeline that produces flawless portraits for well-represented groups often *blurs details, shifts skin tone, or distorts facial geometry* for darker-skinned, elderly, and/or female subjects. Once shipped inside commercial creative tool chains, these artifacts propagate stereotypes at Internet scale: for example, social-media filters that consistently lighten skin or avatar generators that masculinize ambiguous features.

Our project seeks to answer two practical questions about bias in such face-generation models:

- **Measurement.** *How wide is the performance gap* in image fidelity and identity preservation across sub-groups defined by gender, skin tone, and age, and how does this gap vary across leading generative model architectures?
- **Mitigation.** *Which principled, easy-to-implement solution*—requiring minimal additional compute and no full retraining—can most effectively narrow this gap while preserving inference speed and overall visual quality?

Rather than assume access to new training data or large-scale compute for full model retraining, we focus on practical interventions using existing checkpoints. In this work, we benchmark three

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

pretrained face generators on a balanced FairFace-derived testbed and introduce a lightweight rejection sampling filter for Stable Diffusion v1.5 that improves output quality for under-represented groups without modifying model weights. This approach offers a low-cost, deployable path to fairer face generation—one that product teams can adopt before releasing models into high-impact applications like photo filters, game avatars, or marketing tools.

2 Related Work

Recent advances in generative modeling have produced face synthesis tools with remarkable realism, but increasing scrutiny has revealed persistent biases along demographic lines. These disparities often arise from imbalances in the training corpora. For example, the FFHQ dataset used to train StyleGANs contains around 80% light-skinned faces, while CelebA is similarly skewed by gender [Karras et al. [2019], Karkkainen and Joo [2021]]. To address this, Karkkainen and Joo [2021] introduced the FairFace dataset to provide balanced coverage across race, gender, and age groups, enabling a more rigorous foundation for fairness auditing in face-related tasks. FairFace has since become a key benchmark for probing generative models, especially in evaluating representational harms across sensitive attributes.

Generative adversarial networks (GANs) like StyleGAN and its successors—first developed by Karras et al. [2019, 2020a,b]—have become foundational in high-quality face generation. These models introduced innovations such as style-based latent controls and adaptive data augmentation, which enabled stable training even with limited data. More recently, diffusion-based approaches such as Latent Diffusion [Rombach et al. [2022]] have surpassed GANs in sample diversity and photorealism by modeling the data generation process as a denoising trajectory through a latent space. Despite architectural differences, both model families are prone to reproducing the demographic skews present in their training data, motivating fairness evaluations that span architectures. Yet few studies systematically benchmark GANs and diffusion models using unified metrics and balanced demographic scaffolds.

To quantify fidelity and performance gaps, a range of metrics have been developed. The Fréchet Inception Distance (FID) [Heusel et al. [2017]] remains a dominant standard, capturing the distance between real and generated distributions in an Inception-v3 feature space. However, FID conflates image diversity and visual quality. To decouple these dimensions, Sajjadi et al. [2018] proposed a Precision–Recall framework that separately evaluates coverage and realism. Furthermore, LPIPS [Zhang et al. [2018]] assesses perceptual similarity by comparing deep visual features, while FaceNet embeddings [Schroff et al. [2015]] provide a proxy for identity preservation, supporting fairness audits that move beyond pixel-level similarity to more human-relevant attributes.

A growing body of work has applied these metrics to audit demographic bias in generative models. These studies consistently find that darker-skinned, older, and female faces are more likely to suffer from visual degradation, identity inconsistency, or stylistic instability. Teo et al. [2023] argued that subgroup performance gaps can be distorted by biased attribute classifiers and proposed CLEAM (Classifier Error-Aware Measurement) to correct for this issue. Their work underscores the importance of auditing not only the generators but also the tools used to evaluate them. However, despite increasingly refined diagnostics, most fairness audits stop short of offering mitigation techniques.

Early mitigation efforts often relied on retraining models with fairness-aware objectives. Fairness GAN Sattigeri et al. [2018] and FairGAN Xu et al. [2018], for example, use adversarial losses to promote demographic parity but require full model access and high compute. More recent approaches work with frozen models—for instance, FairGen Tan et al. [2020] edits latent codes, and Tibet Chinchure et al. [2024] rewrites prompts to mitigate bias—but typically target conditional generators or assume controllable latent spaces. We extend this line of work with a simple post hoc rejection sampling method for Stable Diffusion v1.5, filtering outputs by CLIP similarity and race classification. Unlike retraining-based approaches, our method requires no model updates and reduces demographic bias without compromising diversity or realism.

Beyond academic benchmarking, the ethical and commercial stakes of fair face generation are growing. Generative models are increasingly embedded in consumer tools, from filters and avatars to shopping assistants and biometric systems. When these systems reproduce or amplify demographic bias, the consequences scale quickly; lightening skin, masculinizing features, or erasing age cues not only alienate users but reinforce harmful stereotypes. As generative models spread across creative

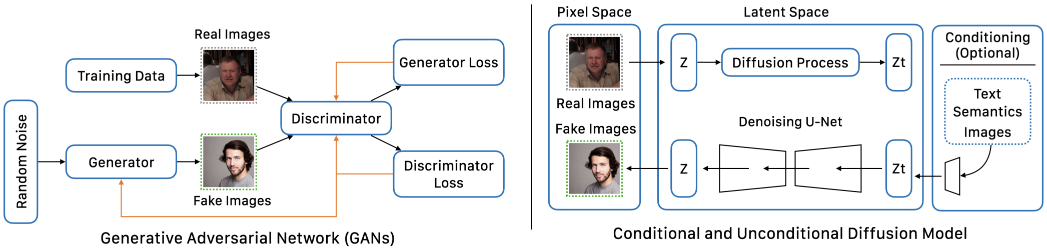

Figure 1: Overview of the generative architectures benchmarked in our study. **Left:** Generative Adversarial Networks (GANs) synthesize images by learning to fool a discriminator trained to distinguish real from fake samples. **Right:** Diffusion models generate images by iteratively denoising latent variables sampled from a Gaussian prior. Conditioning (e.g., text prompts) can guide generation, as in Stable Diffusion, or be omitted for unconditional models like FaceDiffusion.

industries and advertising, deployable fairness interventions—like those we propose—are essential. Our work contributes to the broader conversation on algorithmic accountability, emphasizing that technical excellence must go hand-in-hand with representational equity.

3 Methods

3.1 Overview

Benchmarking Our goal for this project is to quantify how three widely-used face generators—StyleGAN2-ADA, Stable Diffusion v1.5, and FaceDiffusion—vary in quality and demographic balance. The architecture of the models is shown in Fig 1. The evaluation pipeline is as follows. First, we draw *real* reference images from a balanced slice of the FairFace v1.3 dataset. Next, each generator produces an equal-sized set of *synthetic* faces under identical sampling budgets. For unconditional models, a lightweight FairFace 7-race classifier assigns race labels so that outputs can be sorted into folders that mirror the ground-truth taxonomy. Finally, image-quality metrics (FID and LPIPS) are computed independently for every race folder, and a single fairness indicator ΔFID—the difference between the worst- and best-performing sub-groups—summarizes the largest observed gap. All experiments run on a single NVIDIA GPU. The full codebase and pre-trained checkpoints are publicly released to ensure reproducibility.

Bias Mitigation by Rejection Sampling Some failure modes, such as low-quality or duplicated faces, often appear disproportionately in under-represented racial groups, exaggerating measured bias. Rather than retrain a large model, we apply a lightweight *rejection-sampling* pass to Stable Diffusion v1.5. A sample is retained only if (i) a FairFace classifier confirms the target (*race*, *gender*), (ii) its CLIP similarity to the subgroup prompt exceeds 0.27. We keep at most $M = 300$ images per bucket, trimming only ~10% of outputs yet markedly shrinking the ensuing ΔFID gap.

3.2 Data

We evaluate against the **FairFace v1.3** dataset, which provides annotations for seven race categories, two genders, and coarse age brackets.[1] To eliminate sampling bias, a deterministic script selects exactly 500 images for every (*race*, *gender*) combination, yielding 6 balanced groups and a total of 7 000 photographs. Each image is center-cropped and resized to 512×512 px, and its metadata are stored in `metadata.csv` to facilitate downstream analysis.

3.3 Baseline Generative Models

Our study examines three publicly available face generators in their original form. The three models represent three important approaches in image generation: generative adversarial networks

[1] `https://github.com/joojs/fairface`

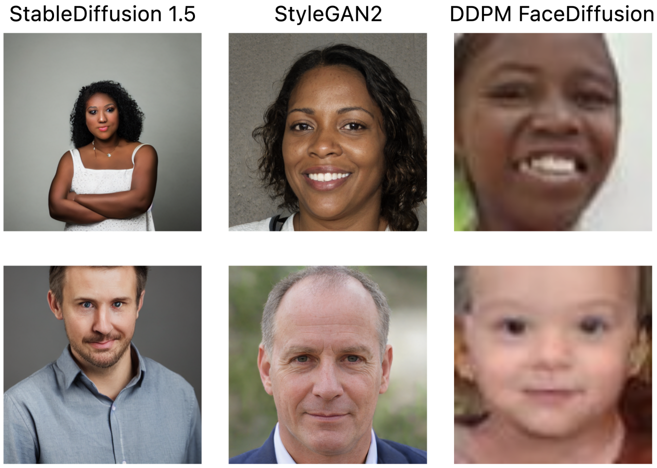

Figure 2: **Face Generation** images generated by the three models for two demographic groups.

(**StyleGAN2-ADA**); domain specific diffusion models (**FaceDiffusion**); and large, general diffusion models (**Stable Diffusion v1.5**). **StyleGAN2-ADA** is evaluated using the official `ffhq.pkl` checkpoint, trained on FFHQ with adaptive discriminator augmentation. We evaluate two diffusion-based models. We employ **Stable Diffusion v1.5** with the safety checker disabled to avoid content filtering that could mask demographic artifacts. We also use the unconditional DDPM-based **FaceDiffusion** model for face generation.

All three checkpoints remain strictly *frozen*; we neither fine-tune nor inject supplemental data at this stage. By holding the weights fixed, we can probe and quantify the latent demographic biases that each model inherits from its original training corpus. This setup isolates architectural and training-data effects, allowing us to attribute disparities to model internals rather than tuning artifacts. The insights gleaned here will guide a subsequent mitigation phase—beyond the scope of the present section—in which we fine-tune the generators with bias-aware objectives and balanced data to reduce the disparities identified in our baseline evaluation.

3.4 Generation Procedure

We generate 1000 images with each model. This sample size balances computational feasibility with statistical coverage across all 6 demographic subgroups. For Stable Diffusion v1.5, we manually craft a single prompt for each FairFace race×gender bucket—for example, "*studio portrait photo of a Black woman, 85 mm DSLR*"—to guide the generation process. To ensure reproducibility across runs, we fix a global random seed, which synchronizes latent vectors or noise inputs across different sampling iterations. For unconditional generators like StyleGAN2-ADA and FaceDiffusion, we apply off-line classification using a FairFace-trained ResNet-34 model (top-1 accuracy $\approx 90\%$), assigning each output to the appropriate race category based on predicted labels.

3.5 Implementation Details

All experiments are conducted in Python 3.10 using PyTorch 2.3 and CUDA 12.1. We run generation and evaluation pipelines on a single GPU with at least 16 GB of VRAM; results reported here were tested on an NVIDIA A100-40GB.

4 Experiments

4.1 Evaluation

We evaluate generated faces along two axes: image fidelity and identity preservation. **Image fidelity** is quantified using the Fréchet Inception Distance (FID), computed between each race-specific folder of real images and its synthetic counterpart. We use `torch-fidelity` with batched feature extraction

(32 images per GPU pass) to ensure efficient and consistent measurement. **Identity preservation** is assessed using the LPIPS metric based on AlexNet features. We pair the first N real images in each demographic group with the first N generated samples and compute the average perceptual distance. Together, these metrics reveal both coarse-grained and fine-grained differences in model performance across demographic groups.

4.2 Results

	SD15		**StyleGAN2-ADA**		**FaceDiffusion**		**SD15-Balanced**	
Race	**FID**	**LPIPS**	**FID**	**LPIPS**	**FID**	**LPIPS**	**FID**	**LPIPS**
Black Female	260.19	20.85	285.58	0.79	286.00	10.31	257.63	8.07
East Asian Female	211.03	20.09	235.38	3.47	302.33	4.67	179.30	57.97
East Asian Male	215.55	20.28	256.79	9.00	296.50	2.70	181.58	25.49
Southeast Asian Female	208.44	20.12	284.41	0.77	320.67	1.75	174.96	26.52
Southeast Asian Male	226.04	20.28	311.29	0.79	303.57	7.82	182.70	47.90
White Male	213.43	20.21	219.48	7.56	251.71	26.52	213.43	34.46

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

Table 1 reports per-race Fréchet Inception Distance (FID) and LPIPS scores for the three baseline generators - StyleGAN2-ADA, Stable Diffusion v1.5, and FaceDiffusion - alongside our rejection-sampled *SD15-Balanced* variant. Figure 2 complements these metrics with representative images from two demographic groups. Together, the table and figure provide the quantitative and qualitative foundation for the bias analysis that follows.

4.3 Analysis

Demographic Disparities Persist. Across all three baseline generators, darker-skinned and Southeast-Asian faces receive the poorest scores. StyleGAN-2-ADA and FaceDiffusion record the worst FID values—exceeding 300 for *Southeast-Asian female/male*—while delivering the flattest LPIPS curves, indicating simultaneously low quality and low diversity. Even Black-female faces, already disadvantaged in SD15 (260.2 FID), deteriorate further under StyleGAN2-ADA (285.6) and FaceDiffusion (286.0).

Larger Diffusion Models Help but Do Not Solve Bias. Stable Diffusion v1.5, trained on a far broader image corpus, cuts average FID by roughly 20–30 points relative to the GAN and domain-specific diffusion baseline. Nevertheless, its best–worst gap remains sizeable ($\Delta FID \approx 52$), showing that scale alone cannot guarantee parity across sensitive attributes.

Lightweight Rejection Sampling Narrows the Gap. Applying the rejection sampling filter to SD15 (*SD15-Balanced*) reduces FID for five of six demographics, with the largest absolute drops for the East Asian (-32 points) and Southeast Asian (-33 points) groups. LPIPS increases for these same groups, signaling richer intraclass variety rather than mode collapse. The worst-case FID (Black female) also edges down, evidencing that a simple post hoc screen can meaningfully improve both fidelity and diversity without expensive retraining. Residual gaps motivate the more principled fine-tuning strategies we outline for future work.

5 Conclusion

Using our new benchmark, we showed that three popular face generators inherit stark demographic disparities: StyleGAN2-ADA and FaceDiffusion produce the worst fidelity and diversity for darker-skinned and Southeast-Asian faces, while the larger Stable Diffusion v1.5 narrows—but cannot close—the gap (worst–best $\Delta FID \approx 52$). A post-hoc rejection-sampling filter on SD15 improves both FID and LPIPS for five of six groups and lowers the worst-case score, demonstrating that bias can be mitigated without retraining. Our open pipeline, metrics, and findings provide a quantitative baseline and a practical first step toward bias-aware finetuning and loss-driven interventions that aim for true demographic parity in face generation.

Future extensions may include adaptive prompt engineering, classifier-free guidance tuning, or latent rejection during sampling. We also encourage the community to build on our benchmark by incorporating intersectional and region-specific subgroups underrepresented in current datasets. Ultimately, we hope this work serves as both a diagnostic tool and a call to action for more inclusive generative modeling practices.

References

Aditya Chinchure, Pushkar Shukla, Gaurav Bhatt, Kiri Salij, Kartik Hosanagar, Leonid Sigal, and Matthew Turk. Tibet: Identifying and evaluating biases in text-to-image generative models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pages 6626–6637, 2017.

Kimmo Karkkainen and Jungseock Joo. 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 (WACV)*, pages 1548–1558, 2021.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2019.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *arXiv preprint arXiv:2006.06676*, 2020a.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8107–8116, 2020b.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.

Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pages 5234–5243, 2018.

Prasanna Sattigeri, Samuel C. Hoffman, Vijil Chenthamarakshan, and Kush R. Varshney. Fairness gan. *arXiv preprint arXiv:1805.09910*, 2018.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Shuhan Tan, Yujun Shen, and Bolei Zhou. Improving the fairness of deep generative models without retraining. *arXiv preprint arXiv:2012.04842*, 2020.

Christopher T. H. Teo, Milad Abdollahzadeh, and NgaiMan Cheung. On measuring fairness in generative models. *CoRR*, abs/2310.19297, 2023.

Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Fairgan: Fairness-aware generative adversarial networks. In *IEEE International Conference on Big Data*, pages 570–575, 2018. doi: 10.1109/BigData.2018.8622525.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. doi: 10.1109/CVPR.2018.00068.