# JUST AS FAIR: BALANCING DIFFUSION TRAJECTORY FOR FAIR TEXT-TO-IMAGE GENERATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Diffusion models exhibit persistent demographic biases, systematically favoring some attributes over others in text-to-image generation. While prior work largely address bias in prompts or representations, we ask a different question: does the denoising process itself explore demographic attributes unevenly? We find that it does, biased generation is closely associated with an imbalance in the exploration of denoising trajectories, which reflects statistical biases in the training data and persists throughout sampling. This motivates a *trajectory-level* view of fairness, where fair generation is treated as balanced exploration of demographic attribute modes, rather than solely as a property of the final outputs. Based on this view, we develop a trajectory-level debiasing framework for fair text-to-image generation. We first introduce *trajectory bias*, together with an observation mechanism based on Tweedie's formula, which estimates the semantic outcome of intermediate noisy latents and makes attribute preferences measurable throughout denoising. Building on this signal, we dynamically regulate the denoising trajectories with a balanced allocation budget, steering the sampling process toward fair attribute coverage. A key advantage of our approach is its *architecture-agnostic* nature, which intervenes directly in sampling dynamics and overcomes a core bottleneck of existing debiasing methods. Across four representative diffusion architectures, including U-Net, DiT, and rectified flow models, our method consistently improves demographic balance while preserving generation quality. Our code is available at `https://anonymous.4open.science/r/Just_As_Fair/`.

## 1 INTRODUCTION

Diffusion models (DMs) generate images by progressively refining noise through a sequence of intermediate latent states (Saharia et al., 2022). These denoising trajectories explore the targeted semantics, typically under a textual guidance, and transform initial noise into coherent and detailed visual content. However, when generating images involving sensitive social attributes, these trajectories are far from being fairly explored, resulting in systematically *biased generation* towards certain demographic modes (e.g., generating CEOs exclusively as male figures). As diffusion-generated imagery becomes increasingly ubiquitous in modern media (GoVisualize, 2026; PicHeadline, 2026), such unchecked biases may reinforce harmful stereotypes and pose a serious ethical challenge.

A natural source of such imbalance is the statistical bias of the training data. Existing studies have shown that DMs tend to reproduce, and sometimes amplify, the demographic biases present in their training corpora (Roos et al., 2026; Seshadri et al., 2024). When a target concept $Y$ (e.g., CEO) is spuriously associated with a demographic attribute $B$ (e.g., male) in the training data, this association can be encoded into the learned denoising dynamics and persist throughout the sampling process. This suggests that the denoising trajectories themselves may carry an inherent imbalance that shapes unfair generation. In this work, we dive into these imbalanced denoising dynamics and revisit the biased generation problem. We point out that demographic bias is not only manifested at the final output, but can already emerge along the denoising trajectories, where different attribute modes are explored with unequal frequency.

Yet, existing debiasing methods rarely address the intrinsic dynamics of diffusion process. One line of work modifies the conditioning signals by debiasing text embeddings of sensitive concepts



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(Fu et al., 2026; Li et al., 2025), but leaves the underlying denoising process unchanged and may compromise prompt alignment. Another intervenes inside the denoiser via steering vectors (Jiang et al., 2025; Shi et al., 2025) or fine-tuning (Qi et al., 2025; Shen et al., 2024), but often relies on architecture-specific heuristics (e.g., semantically meaningful hidden spaces in U-Nets) and incurs substantial training or interpretation overhead. This gap calls for a shift in perspective, from modifying the input or the model, to directly regulating diffusion trajectories where imbalance emerges.

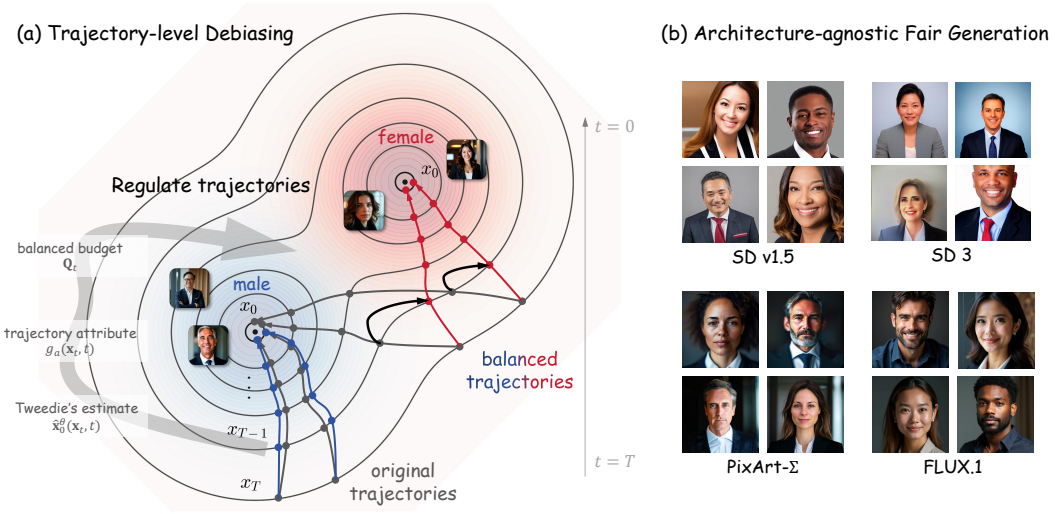



Figure 1: (a) Illustration of our trajectory-level debiasing framework. (b) Qualitative results across four mainstream DMs with distinct architectures.

**Contribution.** In this work, we propose a *trajectory-level* debiasing framework for fair diffusion generation. We formulate fairness as the balanced exploration of diffusion trajectories, shifting the focus from biased final outputs to the underlying denoising process. To make trajectory bias observable and measurable, we introduce an observation mechanism based on Tweedie's formula (Efron, 2011; Luo, 2022), which maps intermediate denoising latents to their estimated semantic outcomes and quantifies demographic preferences throughout generation. This perspective reveals substantial imbalances in trajectory exploration for real-world sensitive attributes (Section 3.1). Moreover, the observed trajectory bias is consistent with statistical biases of the LAION training data (Schuhmann et al., 2022), suggesting that these biases are inherited and amplified the learning of generative dynamics (Mandal et al., 2024; Seshadri et al., 2024).

Building on this observation, we develop an architecture-agnostic debiasing framework that directly intervenes in the denoising process. Our method is guided by a balanced allocation budget, which dynamically estimates and regulates denoising trajectories on the fly, encouraging them to fairly cover diverse attributes (Figure 1 (a)). The framework enables a seamless integration across different diffusion architectures, which addresses a major bottleneck in debiasing frontier model architectures (Shokrolahi et al.) (Figure 1 (b)). Intervening at key denoising steps substantially reduces the trajectory imbalance, while preserving generation quality on general, non-sensitive prompts.

To validate its effectiveness and broad applicability, we conduct comprehensive experiments across four representative DMs, spanning U-Net (Ronneberger et al., 2015), Diffusion Transformer (DiT) (Peebles & Xie, 2023), and Rectified Flow Transformer (Esser et al., 2024) architectures. Across different attribute groups, our method consistently achieves effective debiasing, even outperforming existing model-specific or training-based approaches. As a case study, we further apply our framework to fair medical image generation, taking dermatological image synthesis with diverse skin colors as an example (Appendix F). This study illustrates the potential of trajectory-level debiasing in downstream applications beyond general-purpose image generation.

**Related work.** Strategies for debiasing DMs have received growing attention in recent years. Some works address biased semantics in textual conditions through prompt tuning (Kim et al., 2023), mapping networks (Li et al., 2025), and projections (Fu et al., 2026), aiming to disentangle attribute-related components from text embeddings. However, these methods may only partially resolve

entangled biases, while potentially compromising prompt alignment or generation diversity. Another line of work performs debiasing at the model level by modifying model parameters or internal representations. Some approaches fine-tune DMs using augmented datasets (Shen et al., 2024) or reweighting strategies (Kim et al., 2024), but require substantial training and data curation costs. Others intervene at inference time by steering hidden representations (Jiang et al., 2025) or attention patterns (Park et al., 2025b), but often rely on model-specific semantic spaces. More broadly, most existing methods are restricted to a single DM architecture, limiting their generalizability across different architectures, while imprecise interventions may further compromise semantic alignment.

Another research area related to ours is generating minority samples for rare concepts (Park et al., 2025a). These methods aim to improve generation in low-density regions, either by retraining diffusion models with augmented data (Shao et al., 2024) or by designing tailored samplers for inference (Um et al., 2024). Although such approaches may also promote the generation of bias-conflicting samples, their primary goal is to address semantic rarity rather than demographic imbalance, with a greater emphasis on improving quality of underrepresented samples than ensuring balanced demographic distributions.

**Challenges.** Our work adopts a novel trajectory-oriented perspective for debiasing. We attribute the limited exploration of this direction to two fundamental challenges. The first lies in measuring bias from noisy intermediate latents. Unlike model-level debiasing methods that characterize bias from fully generated images, trajectory-level regulation must infer the attributes of the eventual generation from the current noisy latent alone, rendering existing bias evaluation metrics difficult to apply. The second challenge concerns regulating the intricate step-by-step denoising process. It is well recognized that directly manipulating the sampling trajectory can degrade image quality or induce off-manifold deviations (Cho et al., 2026; Chung et al., 2025). Determining when and how strongly to intervene while preserving generation quality remains non-trivial challenges.

## 2 BACKGROUND

Before presenting our approach, we briefly review the key concepts of diffusion models, and their conditional generation toward target semantics. We then introduce Tweedie's denoising formula, which provides the basis for observing bias at the trajectory level.

**Diffusion models.** DMs are a class of generative models that synthesize images by progressively denoising latent variables along a learned diffusion trajectory (Ho et al., 2020; Song et al., 2021). For a given $D$-dimensional image $\mathbf{x}_0$, the forward diffusion process stepwisely corrupts the image with Gaussian noise, gradually transforming $\mathbf{x}_0$ into a Gaussian prior:

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t}, \mathbf{x}_{t-1}, (1-\alpha_t)\mathbf{I}), \tag{1}$$

where $\alpha_t \in (0,1)$ is a predefined noise-schedule parameter. The reverse sampling process, in contrast, starts from a Guassian sample and traverses the diffusion trajectory in the opposite direction. To approximate this reverse transition, a denoiser network $\boldsymbol{\varepsilon}_\theta$ is trained to predict the injected noise $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{I})$ injected during the forward process:

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}\big(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)\big), \tag{2}$$

The denoiser is trained with the standard noise-prediction objective, which learns the reverse diffusion dynamics that connect the Gaussian prior and the data distribution:

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{x}_0, \boldsymbol{\varepsilon}, t}\left[\left\|\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}_\theta(\mathbf{x}_t, t)\right\|^2\right], \tag{3}$$

During inference, sampling begins by drawing $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$ and iteratively applying the reverse transitions $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$, producing a latent trajectory $\{\mathbf{x}_T, \mathbf{x}_{T-1}, \ldots, \mathbf{x}_0\}$ that terminates with a generated sample $\mathbf{x}_0$. We revisit the biased generation problem via the lens of these latent trajectories, which directly shape the distribution of generated samples.

**Classifier-free guidance (CFG).** DMs also support conditional generation of target semantics given a text prompt $c$. Let $\boldsymbol{\varepsilon}_\theta(\mathbf{x}_t, t, \varnothing)$ and $\boldsymbol{\varepsilon}_\theta(\mathbf{x}_t, t, c)$ denote the unconditional and conditional noise predictions, respectively. Classifier-free guidance (Ho & Salimans, 2022) defines the guided denoising prediction as:

$$\tilde{\boldsymbol{\varepsilon}}_\theta(\mathbf{x}_t, t, c) \;:=\; \boldsymbol{\varepsilon}_\theta(\mathbf{x}_t, t, \varnothing) + \gamma\big(\boldsymbol{\varepsilon}_\theta(\mathbf{x}_t, t, c) - \boldsymbol{\varepsilon}_\theta(\mathbf{x}_t, t, \varnothing)\big), \tag{4}$$

where $\gamma > 0$ denotes the guidance scale. While this mechanism steers diffusion trajectories toward regions consistent with the conditioning signal, it may also amplify the demographic biases in final samples (Kim et al., 2025; 2026).

**Tweedie's formula for denoising.** Regulating the latent trajectory for fair generation requires estimating the final attributes of $\mathbf{x}_0$ from noisy intermediate latents $\mathbf{x}_t$, which is barely covered by existing bias-related metrics. We bridge this gap by leveraging Tweedie's formula (Efron, 2011), which provides the Bayesian optimal denoising estimate (i.e., the posterior mean) through the score function. Under the Gaussian corruption process of DMs, Tweedie's formula gives a closed-form estimate of the clean sample $\mathbf{x}_0$ from its noisy observation $\mathbf{x}_t$:

$$\mathbb{E}[\mathbf{x}_0 \mid \mathbf{x}_t] = \frac{\mathbf{x}_t + (1-\alpha_t)\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)}{\sqrt{\alpha_t}}, \tag{5}$$

which depends on the score of the noisy marginal distribution $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t)$. In practice, this score can be approximated from the denoiser's prediction as $\nabla \mathbf{x}_t \log p(\mathbf{x}_t) \approx -\boldsymbol{\varepsilon}_\theta(\mathbf{x}_t, t)/\sqrt{1-\alpha_t}$, which yields a clean-sample proxy that can be computed at any intermediate denoising step (see Appendix A.1 for the derivation):

$$\hat{\mathbf{x}}_0^\theta(\mathbf{x}_t, t) \;:=\; \frac{\mathbf{x}_t - \sqrt{1-\alpha_t}\,\boldsymbol{\varepsilon}_\theta(\mathbf{x}_t, t)}{\sqrt{\alpha_t}}. \tag{6}$$

This approximation enables the probing of the final images without running a full sampling process. Due to its forecasting capability, we leverage these intermediate estimates to investigate trajectory-level bias and reveal severe imbalances in path exploration.

## 3 METHOD

In this section, we present our debiasing framework that quantifies and regulates the trajectory imbalance in exploring data manifolds. We first develop an observation mechanism to estimate the final attribute along current denoising trajectory (Section 3.1). Our proposed metric, *trajectory bias*, captures the imbalance in trajectory exploration across parallel generations. We show that this imbalance is already pronounced at early denoising timesteps, causing minority data modes to be under-explored. Motivated by this observation, we propose an inference-time regulating strategy that directly optimizes intermediate latents to steer parallel trajectories toward a balanced exploration budget (Section 3.2).

### 3.1 THE TRAJECTORY-LEVEL BIAS

Tweedie's denoised estimator (Eq. 6) provides an estimate of the corresponding clean sample from an intermediate denoising state. Although imperfect, this estimate is directly compatible with existing attribute classifiers, allowing us to probe the attribute evidence implied by the current denoising trajectory. Given a set of sensitive attribute classes $\mathcal{A} = \{a_1, a_2, \ldots, a_C\}$, we quantify the evidence for each attribute class at every intermediate denoising state.

**Definition 3.1** (Trajectory Attribute Score). *For an intermediate latent $\mathbf{x}_t \in \mathbb{R}^{C_z \times H_z \times W_z}$ at timestep $t$ and an attribute class $a$, we define the Trajectory Attribute Score at timestep $t$ as*

$$g_a(\mathbf{x}_t, t) := \left[ f\left(\varphi\left(\hat{\mathbf{x}}_0^\theta(\mathbf{x}_t, t)\right)\right)\right]_a \in \mathbb{R}. \tag{7}$$

*where $\varphi : \mathbb{R}^{C_z \times H_z \times W_z} \to \mathbb{R}^{3\times H\times W}$ is the fixed decoder of DM, and $f : \mathbb{R}^{3\times H\times W} \to \mathbb{R}^K$ is a pre-trained $K$-class attribute classifier (e.g., $K = 2$ for gender classification). The score $g_a(\mathbf{x}_t, t)$ corresponds to the logit of attribute class $a$.*

This score turns each intermediate latent into an observable semantic signal, allowing us to track not only which attributes a trajectory explores, but also when they emerge. Therefore, we examine this signal across four representative DMs with diverse architectures and 100 stereotyped occupations from WinoBias (Zhao et al., 2018). [1]

Figure 3 visualizes Tweedie denoising estimates across timesteps. Remarkably, the estimated samples begin to exhibit demographic semantics (e.g., gender, race, and age) at early denoising steps,

[1] We use the same model and prompt configurations as in our main experiments (Section 4.1).

and rapidly develop toward the key semantic structure of the final samples, becoming clearly identifiable as early as timestep 5. Taking gender as an example, Figure 4 further shows that trajectory attribute scores remain highly accurate and stable even at these early timesteps. These results reveal that intermediate denoising latents already carry reliable demographic semantics, allowing its evolution throughout the generation trajectory to be directly observed and quantitatively measured.

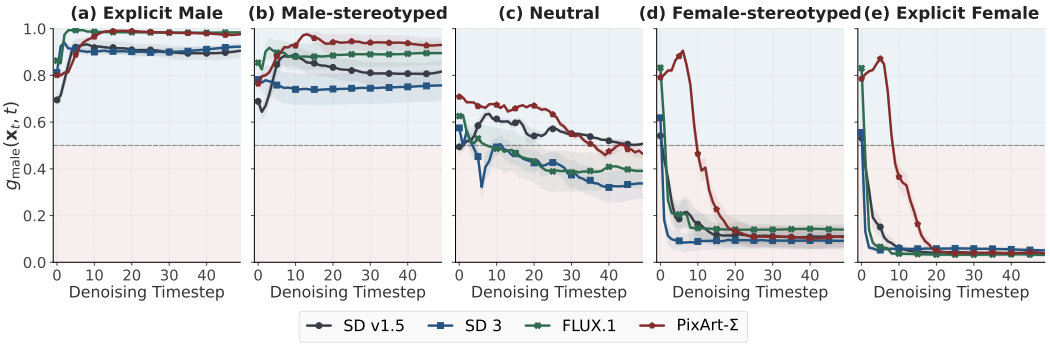



Figure 2: Mean trajectory attribute scores under five prompt conditions.

With the observation mechanism, we next examine the mean trajectory attribute scores under five prompt conditions[2]. As shown in Figure 2, generating gender-stereotyped occupations (Figure 2(b), (d)) exhibit a pronounced trajectory imbalance, with the mean attribute scores quickly shift toward the stereotyped gender in the early denoising steps and remain strongly skewed thereafter. Notably, these trajectories closely resemble those induced by explicit gender instructions (Figure 2(a), (e)), suggesting that the gender-skewed generation is already encoded in the denoising dynamics. In other words, even without an explicit gender instruction, the sampling process already develops a comparable gender preference. In contrast, under the neutral condition (Figure 2(c)), the mean attribute scores remain near the balanced region without consistently favoring either gender. These observations reveal that demographic bias can be traced through the denoising trajectory itself, motivating our trajectory-level characterization of bias, defined below.

**Definition 3.2** (Trajectory Bias)**.** *Given $N$ parallel denoising trajectories $\{\mathbf{x}_t^i\}_{i=1}^N$ at timestep $t$ and $C$ sensitive attribute classes $\mathcal{A} = \{a_1, a_2, \ldots, a_C\}$, we define the mean trajectory attribute score for attribute $a_j$ as*

$$\bar{g}_{a_j}(t) = \frac{1}{N} \sum_{i=1}^{N} g_{a_j}(\mathbf{x}_t^{(i)}, t). \tag{8}$$

*The trajectories are considered **balanced** in attribute exploration when the mean scores remain sufficiently close across all sensitive attributes, i.e.,*

$$\left| \bar{g}_{a_j}(t) - \frac{1}{C} \sum_{k=1}^{C} \bar{g}_{a_k}(t) \right| \leq \epsilon, \qquad j = 1, \ldots, C, \tag{9}$$

*where $\epsilon > 0$ is a predefined tolerance. Accordingly, we define **trajectory bias** at timestep $t$ as the deviation of the mean trajectory attribute scores from their cross-attribute average:*

$$\mathcal{B}(t) = \left[ \frac{1}{C} \sum_{j=1}^{C} \left( \bar{g}_{a_j}(t) - \frac{1}{C} \sum_{k=1}^{C} \bar{g}_{a_k}(t) \right)^2 \right]^{1/2}. \tag{10}$$

*A larger $\mathcal{B}(t)$ indicates a stronger imbalance in attribute exploration, while $\mathcal{B}(t) = 0$ corresponds to balanced exploration across all sensitive attributes.*

As illustrated in Figure 5, we track the evolution of gender trajectory bias across DMs. It is observed that trajectory bias is established early in the denoising process and persists throughout the trajectory,

[2]Figure 2 (a,e) use explicit gender prompts, *A photo of the face of a male/female* {*occupation*}, (b,d) use gender-stereotyped occupation prompts, *A photo of the face of a/an* {*occupation*}, and (c) uses neutral prompts without human figures, e.g., *A photo of a beautiful landscape.*

with only its speed and magnitude varying across architectures. Moreover, the observed trajectory bias during inference closely aligns with the statistical bias of the corresponding training data, as discussed in Appendix B.3 and Figures 7-8. It reveals that biased generation originates from skewed denoising trajectories learned during training, which fundamentally constrains the sampling process to imbalanced exploration. Building on this trajectory-based insight, we introduce an inference-time debiasing framework that regulates trajectory explorations, which substantially reduces trajectory bias throughout the denoising process (as illustrated in Figure 6).

### 3.2 Balancing Diffusion Trajectories

Here raises a natural question: given the severe imbalance among denoising trajectories, how can we regularize their dynamics to ensure fair coverage across different attributes? To address this question, we introduce a balanced exploration budget that assigns an equal share to each attribute. We then enforce this budget during the denoising process through a balance-aware objective, which adaptively optimizes the intermediate latents at key denoising steps to encourage balanced trajectory exploration. Below, we describe how to construct the exploration budget and apply this trajectory-level regularization on the fly.

Consider a set of attribute classes $\mathcal{A} = \{a_1, \dots, a_C\}$ and a batch of $N$ parallel denoising trajectories. We regard these trajectories as a finite exploration budget and allocate an equal share, $N/C$, to each attribute. At denoising timestep $t$, let $\mathbf{X}_t = \{\mathbf{x}_t^i\}_{i=1}^N$ denote the intermediate latents. Following the observation mechanism introduced in Section 3.1, we first obtain the attribute score of each trajectory as:

$$p_{i,c}^t = g_{a_c}(\mathbf{x}_t^i, t), \qquad i = 1, \dots, N, \quad c = 1, \dots, C, \tag{11}$$

where $p_{i,c}^t$ denotes the probability that the $i$-th trajectory explores attribute $a_c$. Collectively, these scores form the matrix $\mathbf{P}_t \in [0,1]^{N \times C}$.

To ensure the fair exploration, a straightforward approach would be to directly enforce the batch-averaged attribute distribution to match a uniform distribution. However, this aggregate constraint does not specify how the exploration budget should be distributed among individual trajectories. We therefore formulate budget allocation as a balanced assignment problem. Specifically, we compute a soft assignment matrix $\mathbf{Q}_t \in \mathbb{R}_+^{N \times C}$ by solving

$$\begin{aligned} \mathbf{Q}_t = \underset{\mathbf{Q} \geq 0}{\arg\max} \quad & \sum_{i=1}^{N} \sum_{c=1}^{C} Q_{i,c} \log p_{i,c}^t + \tau \mathcal{H}(\mathbf{Q}) \\ \text{s.t.} \quad & \mathbf{Q}\mathbf{1}_C = \mathbf{1}_N, \mathbf{Q}^\top \mathbf{1}_N = \frac{N}{C} \mathbf{1}_C, \end{aligned} \tag{12}$$

where $\mathcal{H}(\mathbf{Q}) = -\sum_{i,c} Q_{i,c} \log Q_{i,c}$ is the entropy regularizer and $\tau > 0$ is the assignment temperature. The row constraint of $\mathbf{Q}_t$ assigns one unit of exploration budget to each trajectory, while the column constraint ensures that each attribute receives exactly $N/C$ units. Meanwhile, the objective favors assignments consistent with the current semantic tendency of each trajectory. Thus, the resulting allocation achieves batch-level balance without prescribing a fixed attribute label for each initial noise sample. In implementation, we efficiently solve for $\mathbf{Q}_t$ through iterative Sinkhorn algorithm (Sinkhorn & Knopp, 1967), which is designed to iteratively rescale a nonnegative matrix to match prescribed row and column marginals.

Based on this balanced allocation, we define the loss function of balanced exploration as:

$$\mathcal{L}_{\mathrm{BE}}(\mathbf{X}_t) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} \mathrm{sg}\big(Q_{i,c}^t\big) \log p_{i,c}^t, \tag{13}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient operation. This objective encourages each trajectory to follow the attribute assigned by the balanced allocation. As the assignments enforce an equal budget across attributes, minimizing $\mathcal{L}_{\mathrm{BE}}$ promotes balanced exploration across the batch, preventing trajectories from collapsing toward the same dominant attribute mode.

We enforce the balanced exploration objective by directly updating the intermediate latents only at selected denoising timesteps. Let $\mathcal{T}_{\mathrm{int}}$ denote the set of intervention timesteps. For each $t \in \mathcal{T}_{\mathrm{int}}$, let

$$\mathbf{d}_t^i = \nabla_{\mathbf{x}_t^i} \mathcal{L}_{\mathrm{BE}}(\mathbf{X}_t), \qquad t \in \mathcal{T}_{\mathrm{int}}. \tag{14}$$

denote the gradient for the $i$-th trajectory. To make the intervention less sensitive to the scale of gradients at different timesteps, we normalize the gradients with their mean batch-wise norm:

$$\bar{r}_t = \frac{1}{N} \sum_{i=1}^{N} \left\| \mathbf{d}_t^i \right\|_2 . \tag{15}$$

The intermediate latent is then regularized as

$$\widetilde{\mathbf{x}}_t^i = \mathbf{x}_t^i - \lambda \frac{\mathbf{d}_t^i}{\bar{r}_t + \epsilon}, \tag{16}$$

where $\lambda$ controls the strength of balance-aware guidance and $\epsilon$ is a small constant for numerical stability. After the update, we recompute the denoising prediction from $\widetilde{\mathbf{x}}_t^i$ and apply the original sampler transition:

$$\mathbf{x}_{t-1}^i = \mathcal{S}_t \left( \widetilde{\mathbf{x}}_t^i, \, \varepsilon_\theta(\widetilde{\mathbf{x}}_t^i, t, \mathbf{y}) \right), \tag{17}$$

where $\varepsilon_\theta$ is the pretrained denoiser conditioned on prompt $\mathbf{y}$ and $\mathcal{S}_t$ denotes the original sampling update. This allows each subsequent transition to directly proceed from the fairness-regularized latent state. Repeating this procedure aross the intervention timesteps progressively directs the denoising trajectories toward balanced attribute exploration, without modifying the model parameters or sampling formulation. We establish a theoretical proposition for our trajectory-level regularization, proving that it reduces trajectory bias $\mathcal{B}(t)$ to achieve fairer generation:

**Proposition 1** (Balanced Exploration Controls Trajectory Bias)**.** *Let $\mathbf{p}_i^t$ denote the attribute probability vector of trajectory $i$, and let $\mathbf{Q}_t$ be the balanced assignment obtained from Eq. 12. Since $\mathbf{Q}_t^\top \mathbf{1}_N = \frac{N}{C} \mathbf{1}_C$, its average target distribution is uniform. Then the trajectory bias satisfies*

$$\mathcal{B}(t) \leq \sqrt{\frac{2}{C} \left( \mathcal{L}_{\mathrm{BE}}(\mathbf{X}_t) - \frac{1}{N} \sum_{i=1}^{N} H(\mathbf{q}_i^t) \right)}, \tag{18}$$

*where $\mathbf{q}_i^t$ is the $i$-th row of $\mathbf{Q}_t$. Thus, the balanced exploration loss provides a upper bound on trajectory-level imbalance, where minimizing $\mathcal{L}_{\mathrm{BE}}$ drives the aggregate attribute distribution toward the uniform exploration budget. See Appendix A.2 for the proof.*

# 4 Experiments

## 4.1 Setup

In our experiments, we focus on evaluating fairness with respect to gender and race. Following prior work, we evaluate on 100 stereotyped occupations from the extended Winobias dataset (Zhao et al., 2018), using the prompt *a photo of a face of a/an* {*occupation*}. We consider binary gender, and categorize race into four groups: WMELH (White, Middle Eastern, and Latino Hispanic), Asian (East Asian and Southeast Asian), Black, and Indian. All experiments are repeated three times with different random seeds for robustness. Further implementation details are in Appendix C.1.

**Models**. To evaluate the model-agnostic nature of our debiasing framework, we consider four text-to-image DMs with different denoising architectures: SD v1.5 (U-Net) (Rombach et al., 2022), SD 3 (MMDiT) (Esser et al., 2024), PixArt-$\Sigma$ (DiT) (Chen et al., 2024), and FLUX.1 (Rectified Flow Transformer) (Labs et al., 2025). Architectures backgrounds are provided in Appendix C.2.

**Baselines.** We consider representative *model-agnostic* debiasing baselines that focus on modifying text conditions. These include **Controlled Prompt** that uses fairness-aware instructions (e.g., *Ensure diversity across gender and ethnicity.*) (Raza et al., 2025), **FairImagen** that achieves state-of-the-art performance in debiasing text embeddings (Fu et al., 2026), and **Weak Guidance** that injects guidance toward minority attributes into the text embeddings (Kim et al., 2025). To further compare against *model-specific* debiasing methods for SD v1.5, we include representative approaches from three categories: **FairGen** (Jiang et al., 2025) and **DiffLens** (Shi et al., 2025) with steering vectors, **EFA** (Park et al., 2025b) and **Balancing Act** (Parihar et al., 2024) with representation manipulation, and **BiasNeuron** (Qi et al., 2025) with fine-tuning. For all baselines, we follow the configurations provided in their official repositories or research papers.

Table 1: Results of fairness, accuracy, and quality on four DMs. Best results are marked in bold.

| Method | Gender | | | | Race | | | |
|---|---|---|---|---|---|---|---|---|
| | FD ↓ | Fairness ↑ | CLIP-T ↑ | MUSIQ ↑ | FD ↓ | Fairness ↑ | CLIP-T ↑ | MUSIQ ↑ |
| **SD v1.5** | | | | | | | | |
| Base | 0.386±0.036 | 0.352±0.024 | 0.722±0.000 | 0.726±0.001 | 0.364±0.001 | 0.392±0.043 | 0.722±0.000 | 0.726±0.001 |
| Controlled | 0.225±0.007 | 0.602±0.219 | 0.655±0.008 | **0.739**±0.009 | 0.344±0.001 | 0.420±0.055 | 0.655±0.008 | **0.739**±0.009 |
| Weak Guid. | 0.218±0.005 | 0.545±0.145 | 0.619±0.002 | 0.656±0.016 | 0.291±0.005 | 0.579±0.144 | 0.635±0.001 | 0.645±0.016 |
| Fairimagen | 0.194±0.008 | 0.608±0.212 | 0.627±0.000 | 0.590±0.001 | **0.249**±0.005 | 0.484±0.123 | 0.622±0.004 | 0.531±0.010 |
| *Ours* | **0.174**±0.008 | **0.738**±0.286 | **0.698**±0.002 | 0.705±0.001 | 0.253±0.001 | **0.627**±0.182 | **0.701**±0.007 | 0.699±0.003 |
| **SD 3** | | | | | | | | |
| Base | 0.543±0.022 | 0.175±0.180 | 0.720±0.000 | 0.733±0.011 | 0.478±0.011 | 0.258±0.110 | 0.720±0.000 | 0.733±0.011 |
| Controlled | 0.453±0.072 | 0.315±0.074 | 0.708±0.006 | 0.722±0.014 | 0.404±0.011 | 0.363±0.050 | 0.708±0.006 | 0.722±0.014 |
| Weak Guid. | 0.500±0.067 | 0.261±0.114 | **0.710**±0.004 | **0.740**±0.002 | 0.414±0.015 | 0.303±0.044 | 0.707±0.000 | **0.762**±0.002 |
| Fairimagen | **0.293**±0.010 | 0.503±0.112 | 0.657±0.030 | 0.650±0.004 | 0.365±0.003 | 0.384±0.007 | 0.699±0.002 | 0.654±0.003 |
| *Ours* | 0.301±0.008 | **0.572**±0.135 | **0.710**±0.000 | 0.713±0.002 | **0.278**±0.009 | **0.559**±0.169 | **0.736**±0.024 | 0.727±0.009 |
| **PixArt-Σ** | | | | | | | | |
| Base | 0.536±0.005 | 0.151±0.186 | 0.733±0.001 | 0.723±0.003 | 0.417±0.011 | 0.273±0.079 | 0.733±0.001 | 0.723±0.003 |
| Controlled | 0.422±0.002 | 0.286±0.061 | **0.714**±0.004 | **0.713**±0.005 | 0.414±0.008 | 0.351±0.024 | **0.714**±0.002 | 0.709±0.005 |
| Weak Guid. | 0.313±0.008 | 0.425±0.035 | 0.645±0.003 | 0.374±0.006 | 0.375±0.045 | 0.325±0.035 | 0.627±0.070 | 0.533±0.122 |
| Fairimagen | 0.259±0.007 | 0.413±0.135 | 0.660±0.020 | 0.655±0.020 | 0.302±0.015 | 0.382±0.064 | 0.700±0.002 | 0.650±0.003 |
| *Ours* | **0.216**±0.007 | **0.554**±0.182 | 0.677±0.021 | 0.663±0.020 | **0.237**±0.021 | **0.462**±0.085 | 0.722±0.002 | **0.710**±0.002 |
| **FLUX.1** | | | | | | | | |
| Base | 0.527±0.013 | 0.222±0.148 | 0.703±0.001 | 0.715±0.005 | 0.327±0.007 | 0.244±0.136 | 0.703±0.001 | 0.715±0.005 |
| Controlled | 0.424±0.006 | 0.344±0.053 | 0.696±0.001 | **0.734**±0.001 | 0.395±0.002 | 0.400±0.014 | 0.696±0.000 | 0.710±0.002 |
| Weak Guid. | 0.553±0.003 | 0.172±0.186 | **0.704**±0.000 | 0.713±0.005 | 0.550±0.003 | 0.186±0.174 | **0.705**±0.000 | 0.713±0.005 |
| Fairimagen | 0.275±0.020 | 0.504±0.120 | 0.685±0.003 | 0.705±0.005 | 0.237±0.030 | 0.259±0.091 | 0.690±0.000 | 0.710±0.001 |
| *Ours* | **0.267**±0.022 | **0.563**±0.160 | 0.683±0.003 | 0.707±0.003 | **0.272**±0.063 | **0.496**±0.123 | 0.694±0.000 | **0.714**±0.001 |

**Metrics.** We evaluate the fairness, accuracy, and image quality of debiased generation. For *fairness*, we adopt two metrics, **Fairness Discrepancy (FD)** (Parihar et al., 2024) and **Fairness Score** (Fu et al., 2026), based on a pretrained facial attribute classifier from Deepface (Taigman et al., 2014). FD measures the Euclidean distance between the generated attribute distribution and a uniform distribution, while Fairness Score measures the balance across attribute groups. For *accuracy*, we measure prompt alignment using **CLIP-T**, which quantifies the semantic similarity between the generated image and the text prompt. For *image quality*, we use **MUSIQ** (Ke et al., 2021) to assess the perceptual quality of the generated images. More details about metrics are in Appendix C.3.

### 4.2 Results

**Results of Fair Generation.** Table 1 reports the fairness, accuracy, and generation quality of our debiasing framework, compared with model-agnostic baselines. Across both gender and race debiasing tasks, our framework achieves stronger fairness performance, with consistent improvements in both FD and fairness scores. These results suggest that regulating the denoising trajectory provides a more direct way to promote fair generation. Qualitative results are shown in Figure 9 and Figure 10 for gender and race debiasing, respectively. We also observe several limitations of existing baselines. For example, Controlled Prompt may generate multiple figures within a single image to satisfy its diversity instructions, leading to lower CLIP-T scores. Methods that modify text embeddings, such as Weak Guidance and FairImgen, also show noticeable degradation in image quality and prompt alignment, potentially due to perturbations to the semantics encoded by the original prompt.

We also evaluate the generalization of our framework to the intersectional regulation of gender $\times$ racial bias (Appendix E.1), where demographic diversity across multiple attribute groups is considered (Shukla et al., 2025). Moreover, our trajectory-level debiasing preserves generation quality on general, non-sensitive prompts (Appendix E.2), supporting its practicality for broad applications.

Beyond stereotyped occupations, we further evaluate our framework on fair medical image generation, using dermatology image synthesis with different skin colors as a case study (Appendix F). The

Table 2: Comparison with model-specific baselines on SD v1.5.

| | Gender | | | | Race | | | |
|---|---|---|---|---|---|---|---|---|
| **Method** | **FD ↓** | **Fairness ↑** | **CLIP-T ↑** | **MUSIQ ↑** | **FD ↓** | **Fairness ↑** | **CLIP-T ↑** | **MUSIQ ↑** |
| Base | $0.386_{\pm 0.036}$ | $0.352_{\pm 0.024}$ | $0.722_{\pm 0.000}$ | $0.726_{\pm 0.001}$ | $0.364_{\pm 0.001}$ | $0.392_{\pm 0.043}$ | $0.722_{\pm 0.000}$ | $0.726_{\pm 0.001}$ |
| Balanc.Act | $0.190_{\pm 0.072}$ | $0.633_{\pm 0.257}$ | $0.664_{\pm 0.002}$ | $0.732_{\pm 0.005}$ | $0.269_{\pm 0.022}$ | $\mathbf{0.673}_{\pm 0.200}$ | $0.664_{\pm 0.012}$ | $\mathbf{0.733}_{\pm 0.005}$ |
| FairGen | $0.340_{\pm 0.066}$ | $0.450_{\pm 0.037}$ | $0.693_{\pm 0.001}$ | $\mathbf{0.748}_{\pm 0.005}$ | $0.307_{\pm 0.026}$ | $0.645_{\pm 0.250}$ | $0.690_{\pm 0.001}$ | $0.731_{\pm 0.003}$ |
| DiffLens | $0.335_{\pm 0.003}$ | $0.410_{\pm 0.039}$ | $\mathbf{0.718}_{\pm 0.002}$ | $0.720_{\pm 0.002}$ | $0.358_{\pm 0.000}$ | $0.410_{\pm 0.018}$ | $0.719_{\pm 0.002}$ | $0.720_{\pm 0.002}$ |
| EFA | $0.358_{\pm 0.000}$ | $0.410_{\pm 0.010}$ | $0.716_{\pm 0.000}$ | $0.708_{\pm 0.003}$ | $0.332_{\pm 0.001}$ | $0.449_{\pm 0.054}$ | $\mathbf{0.724}_{\pm 0.000}$ | $0.710_{\pm 0.001}$ |
| BiasNeuron | $0.373_{\pm 0.039}$ | $0.416_{\pm 0.014}$ | $0.680_{\pm 0.008}$ | $0.667_{\pm 0.002}$ | $0.345_{\pm 0.002}$ | $0.408_{\pm 0.031}$ | $0.643_{\pm 0.042}$ | $0.677_{\pm 0.008}$ |
| *Ours* | $\mathbf{0.174}_{\pm 0.008}$ | $\mathbf{0.738}_{\pm 0.286}$ | $\mathbf{0.718}_{\pm 0.002}$ | $0.705_{\pm 0.001}$ | $\mathbf{0.253}_{\pm 0.001}$ | $0.627_{\pm 0.182}$ | $0.701_{\pm 0.007}$ | $0.729_{\pm 0.003}$ |

results show that our trajectory-level regulation also transfers to this specialized setting, leading to more diverse image generation to augment critical downstream datasets. This case study highlights the potential of fairness-aware generation as a data-centric tool, where diverse generation results are essential for reliable model development.

**Comparison with model-specific baselines.** We compare our debiasing framework with model-specific baselines for SD v1.5, where debiasing is extensively studied due to the semantic bottleneck layer in its U-Net architecture. Table 2 reports the comparison results, with our method substantially improves debiasing effectiveness for both gender and race attributes. In contrast, existing model-specific methods face several limitations in fair generation. Methods that rely on self-discovered steering vectors may fail to fully capture the semantics of the target attribute, while the best-performing baseline, Balancing Act, requires substantial additional training at the cost of prompt alignment. Unlike these approaches that focus on final images, our method directly regulates the denoising trajectory, enabling more effective debiasing without these model-specific limitations.

### 4.3 Analysis

**Ablation Studies.** To examine the robustness of our trajectory-level debiasing framework, we conduct ablation studies on intervention steps, prompt variants, and pretrained attribute classifiers in Appendix E.4. The debiasing effectiveness remains stable across prompt formats, while different classifiers consistently yield effective trajectory bias assessment. Across different intervention timesteps, regulating the denoising process consistently improves fairness, with interventions spanning key stages (e.g., $0.10$-$0.80$) achieving the strongest performance. This highlights the importance of these stages for semantic development and trajectory-level bias.

**Cost Analysis.** To assess the practicality of our debiasing framework, we analyze generation time and GPU memory usage in Appendix E.3. Our method is memory-efficient, requiring only $1.5\times$ the memory of the corresponding base models. While the debiased generation incurs additional sampling time, the overhead primarily comes from estimating trajectory bias to enable dynamic regulation during denoising. This design trades additional computation for direct, adaptive control of trajectory-level bias. We leave further acceleration to future work, including multi-GPU execution, early stopping, and lightweight bias estimators.

## 5 Conclusion and Limitations

This work revisits biased generation in diffusion models from a trajectory-level perspective, framing fairness as balanced trajectory exploration rather than solely an output property. We develop a Tweedie's-formula-based observation mechanism to reveal trajectory imbalance across denoising steps, and propose an architecture-agnostic framework to regulate it through balanced allocation and latent-state optimization during inference. Our method improves fairness while preserving prompt semantics and image quality, overcoming the model-specific limitations of existing baselines. A limitation of our framework is the additional inference cost associated with dynamically measuring and regulating trajectories. We leave the exploration of acceleration strategies as an interesting direction for further improving the practicality of trajectory-level debiasing.

## AI USE STATEMENT

In this work, we used generative AI tools to improve the readability of the manuscript and to assist with literature search and identification of relevant papers. We did not use generative AI tools for developing the methodology, designing experiments, implementing methods, interpreting results, or any other research tasks requiring disclosure under the ICLR AI policy. All AI-assisted content was reviewed by the authors, who take full responsibility for the final content of this work.

## ETHICS STATEMENT

Our work studies fair text-to-image generation with diffusion models. To mitigate potential misuse and the risk of reinforcing harmful stereotypes, we do not publicly release the generated images. The dermatological case study investigates fairness in synthetic medical image generation. The generated images are intended solely for research purposes and are not used as clinical evidence or as substitutes for medically validated diagnostic procedures. All datasets and pretrained models are used in accordance with their applicable licenses and usage terms.

## REPRODUCIBILITY STATEMENT

To facilitate the reproducibility of our experiments, we provide detailed implementation settings in Section C.1. For all referenced baselines, we use their official implementations and default parameters. For Weak Guidance (Kim et al., 2025), we reproduce the method based on the implementation details provided in the original paper, as no code is publicly available. Our code is available at the anonymous repository: `https://anonymous.4open.science/r/Just_As_Fair/`.

## REFERENCES

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-$\alpha$: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *International conference on learning representations*, volume 2024, pp. 57611–57640, 2024.

Hyunmin Cho, Donghoon Ahn, Susung Hong, Jee Eun Kim, Seungryong Kim, and Kyong Hwan Jin. TAG: Tangential amplifying guidance for hallucination-resistant sampling. In *Forty-third International Conference on Machine Learning*, 2026. URL `https://openreview.net/forum?id=1XJqPy2LhA`.

Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. Cfg++: Manifold-constrained classifier free guidance for diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 30824–30850, 2025.

Roxana Daneshjou, Kailas Vodrahalli, Roberto A Novoa, Melissa Jenkins, Weixin Liang, Veronica Rotemberg, Justin Ko, Susan M Swetter, Elizabeth E Bailey, Olivier Gevaert, et al. Disparities in dermatology ai performance on a diverse, curated clinical image set. *Science advances*, 8(31): eabq6147, 2022.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Felix Friedrich, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Patrick Schramowski, Sasha Luccioni, and Kristian Kersting. Fair diffusion: Instructing text-to-image generation models on fairness. *arXiv preprint arXiv:2302.10893*, 2023.

Zihao Fu, Ryan Brown, Shun Shao, Kai Rawal, Eoin Delaney, and Chris Russell. Fairimagen: Post-processing for bias mitigation in text-to-image models. *Advances in Neural Information Processing Systems*, 38:29488–29512, 2026.

Ada Görgün, Fawaz Sammani, Nikos Deligiannis, Bernt Schiele, and Jonas Fischer. Temporal concept dynamics in diffusion models via prompt-conditioned interventions. In *International Conference on Learning Representations*, volume 2026, pp. 66640–66669, 2026.

GoVisualize. AI-Native Design Studio for Marketers & Creatives, 2026. URL `https://govisualize.ai`.

Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1820–1828. IEEE, 2021.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Yilei Jiang, Wei-Hong Li, Yiyuan Zhang, Minghong Cai, and Xiangyu Yue. Fairgen: Enhancing fairness in text-to-image diffusion models via self-discovering latent directions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18411–18420, 2025.

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*, pp. 1548–1558, 2021.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5128–5137. IEEE, 2021.

Eunji Kim, Siwon Kim, Chaehun Shin, and Sungroh Yoon. De-stereotyping text-to-image models through prompt tuning. In *ICML 2023 Workshop on Deployable Generative AI*, Honolulu, Hawaii, USA, 2023.

Eunji Kim, Siwon Kim, Minjun Park, Rahim Entezari, and Sungroh Yoon. Rethinking training for de-biasing text-to-image generation: Unlocking the potential of stable diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13361–13370, 2025.

Myeongsoo Kim, Eunji Kim, Minwoo Chae, and Sangwoo Mo. Stay fair! ensuring group fairness in diffusion models across guidance scales. *arXiv preprint arXiv:2605.28036*, 2026.

Yeongmin Kim, Byeonghu Na, Minsang Park, JoonHo Jang, Dongjun Kim, Wanmo Kang, and Il-Chul Moon. Training unbiased diffusion models from biased dataset. In *The Twelfth International Conference on Learning Representations*, 2024.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Jia Li, Lijie Hu, Jingfeng Zhang, Tianhang Zheng, Hua Zhang, and Di Wang. Fair text-to-image diffusion via fair mapping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 26256–26264, 2025.

Calvin Luo. Understanding diffusion models: A unified perspective. *arXiv preprint arXiv:2208.11970*, 2022.

Zhiyuan Ma, Yuzhu Zhang, Guoli Jia, Liangliang Zhao, Yichao Ma, Mingjie Ma, Gaofeng Liu, Kaiyan Zhang, Ning Ding, Jianjun Li, et al. Efficient diffusion models: A comprehensive survey from principles to practices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(9):7506–7525, 2025.

Abhishek Mandal, Susan Leavy, and Suzanne Little. Generated bias: Auditing internal bias dynamics of text-to-image generative models. In *European Conference on Computer Vision*, pp. 96–111. Springer, 2024.

Nusrat Munia and Abdullah-Al-Zubaer Imran. Dermdiff: generative diffusion model for mitigating racial biases in dermatology diagnosis. *arXiv preprint arXiv:2503.17536*, 2025.

Rishubh Parihar, Abhijnya Bhat, Abhipsa Basu, Saswat Mallick, Jogendra Nath Kundu, and R Venkatesh Babu. Balancing act: Distribution-guided debiasing in diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6668–6678, 2024.

Dongmin Park, Sebin Kim, Taehong Moon, Minkyu Kim, Kangwook Lee, and Jaewoong Cho. Rare-to-frequent: Unlocking compositional generation power of diffusion models on rare concepts with llm guidance. In *International conference on learning representations*, volume 2025, pp. 14650–14676, 2025a.

Jeonghoon Park, Juyoung Lee, Chaeyeon Chung, Jaeseong Lee, Jaegul Choo, and Jindong Gu. Fair generation without unfair distortions: Debiasing text-to-image generation with entanglement-free attention. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17567–17576. IEEE, 2025b.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

PicHeadline. PicHeadline: AI-Powered Visuals from Just a Headline, 2026. URL `https://www.picheadline.com`.

Fan Qi, Zhan Wang, Changsheng Xu, and Huaiwen Zhang. Fine-tuning bias neurons for fair text-to-image generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 9657–9666, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Shaina Raza, Maximus Powers, Partha Pratim Saha, Mahveen Raza, and Rizwan Qureshi. Prompting away stereotypes? evaluating bias in text-to-image models for occupations. In *Proceedings of the 1st Workshop on Generative and Protective AI for Content Creation (GenProCC)*, San Diego, CA, USA, December 2025. Co-located with the 39th Conference on Neural Information Processing Systems (NeurIPS 2025); arXiv:2509.00849.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Nathan Roos, Ekaterina Iakovleva, Ani Gjergji, Vito Paolo Pastore, and Enzo Tartaglione. How i met your bias: Investigating bias amplification in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5374–5383, 2026.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Preethi Seshadri, Sameer Singh, and Yanai Elazar. The bias amplification paradox in text-to-image generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6367–6384, 2024.

Jie Shao, Ke Zhu, Hanxiao Zhang, and Jianxin Wu. Diffult: Diffusion for long-tail recognition without external knowledge. *Advances in neural information processing systems*, 37:123007–123031, 2024.

Xudong Shen, Chao Du, Tianyu Pang, Min Lin, Yongkang Wong, and Mohan Kankanhalli. Fine-tuning text-to-image diffusion models for fairness. In *Proceedings of the 12th International Conference on Learning Representations*, 2024. Oral Presentation.

Yingdong Shi, Changming Li, Yifan Wang, Yongxiang Zhao, Anqi Pang, Sibei Yang, Jingyi Yu, and Kan Ren. Dissecting and mitigating diffusion bias via mechanistic interpretability. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8192–8202, 2025.

Sayedmoslem Shokrolahi, Jae-Mo Kang, and Il-Min Kim. Responsible text-to-image diffusion: Interpretable and linearly controllable semantics for fair and safe generation. In *Forty-third International Conference on Machine Learning*.

Pushkar Shukla, Aditya Chinchure, Emily Diana, Alexander Tolbert, Kartik Hosanagar, Vineeth N Balasubramanian, Leonid Sigal, and Matthew Turk. Mitigate one, skew another? tackling intersectional biases in text-to-image models. *arXiv preprint arXiv:2505.17280*, 2025.

Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=St1giarCHLP`.

Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1701–1708, 2014.

Soobin Um, Suhyeon Lee, and Jong Chul Ye. Don't play favorites: Minority guidance for diffusion models. In *International Conference on Learning Representations*, volume 2024, pp. 21632–21670, 2024.

Yuzhe Yang, Haoran Zhang, Judy W Gichoya, Dina Katabi, and Marzyeh Ghassemi. The limits of fair medical imaging ai in real-world generalization. *Nature medicine*, 30(10):2838–2848, 2024.

Ruichen Zhang, Yuguang Yao, Zhen Tan, Zhiming Li, Pan Wang, Huan Liu, Jingtong Hu, Sijia Liu, and Tianlong Chen. Fairskin: fair diffusion for skin disease image generation. *arXiv preprint arXiv:2410.22551*, 2024.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 15–20, 2018.

# A Proofs

## A.1 Derivation of Tweedie's Denoising Estimates (Eq. 6)

We derive the Tweedie's denoising estimator used in Eq. equation 6, starting from the standard form of Tweedie's formula, and showing it can be approximated by the denoiser output.

For a Gaussian corruption process, Tweedie's formula (Efron, 2011) gives the posterior mean of the clean signal as

$$\mathbb{E}[\boldsymbol{\mu} \mid \mathbf{y}] = \mathbf{y} + \sigma^2 \nabla_{\mathbf{y}} \log p(\mathbf{y}), \tag{19}$$

where $\mathbf{y} = \boldsymbol{\mu} + \sigma\boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{I})$, and $p(\mathbf{y})$ is the marginal density of the noisy observation. In diffusion models, the forward process at timestep $t$ is

$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\sqrt{\alpha_t}\mathbf{x}_0,\, (1-\alpha_t)\mathbf{I})\,, \tag{20}$$

or equivalently,

$$\mathbf{x}_t = \sqrt{\alpha_t}\mathbf{x}_0 + \sqrt{1-\alpha_t}\boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{I}). \tag{21}$$

Applying Tweedie's formula with $\boldsymbol{\mu} = \sqrt{\alpha_t}\mathbf{x}_0$ and $\sigma^2 = 1 - \alpha_t$ gives

$$\mathbb{E}[\mathbf{x}_0 \mid \mathbf{x}_t] = \frac{\mathbf{x}_t + (1-\alpha_t)\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t)}{\sqrt{\alpha_t}}, \tag{22}$$

where $p_t(\mathbf{x}_t)$ denotes the marginal density of $\mathbf{x}_t$ at timestep $t$. Meanwhile, rearranging the forward process and taking the conditional expectation given $\mathbf{x}_t$ yields

$$\mathbb{E}[\sqrt{\alpha_t}\mathbf{x}_0 \mid \mathbf{x}_t] = \mathbf{x}_t - \sqrt{1-\alpha_t}\,\mathbb{E}[\boldsymbol{\varepsilon} \mid \mathbf{x}_t, t]. \tag{23}$$

Combining this relation with Eq. 22 gives the exact score-noise identity

$$\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t) = -\frac{\mathbb{E}[\boldsymbol{\varepsilon} \mid \mathbf{x}_t, t]}{\sqrt{1-\alpha_t}}. \tag{24}$$

Under the standard $\boldsymbol{\varepsilon}$-prediction objective of diffusion model training (Eq. 3):

$$\min_{\boldsymbol{\varepsilon}_\theta} \mathbb{E}\Big[\|\boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}_\theta(\mathbf{x}_t, t)\|_2^2\Big]\,, \tag{25}$$

the population-optimal predictor is the conditional mean, so that

$$\boldsymbol{\varepsilon}_\theta(\mathbf{x}_t, t) \approx \mathbb{E}[\boldsymbol{\varepsilon} \mid \mathbf{x}_t, t]. \tag{26}$$

Therefore,

$$\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t) \approx -\frac{\boldsymbol{\varepsilon}_\theta(\mathbf{x}_t, t)}{\sqrt{1-\alpha_t}}. \tag{27}$$

Substituting Eq. 27 into Eq. 22 yields

$$\hat{\mathbf{x}}_0^\theta(\mathbf{x}_t, t) := \frac{\mathbf{x}_t - \sqrt{1-\alpha_t}\,\boldsymbol{\varepsilon}_\theta(\mathbf{x}_t, t)}{\sqrt{\alpha_t}}. \tag{28}$$

which is the clean-sample proxy used in Eq. 6.

### A.2 Proof of Proposition 1

*Proof.* For simplicity, we omit the timestep superscript $t$ when it is clear from context. Let $\mathbf{p}_i$ denote the attribute probability vector of the $i$-th trajectory, and let $\mathbf{q}_i$ denote the $i$-th row of the balanced assignment matrix $\mathbf{Q}$. Define the batch-averaged attribute distributions as

$$\bar{\mathbf{p}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{p}_i, \qquad \bar{\mathbf{q}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{q}_i. \tag{29}$$

First, the column constraint of the balanced assignment in Eq. 12 gives

$$\mathbf{Q}^\top \mathbf{1}_N = \frac{N}{C}\mathbf{1}_C. \tag{30}$$

Therefore,

$$\bar{\mathbf{q}} = \frac{1}{N}\mathbf{Q}^\top \mathbf{1}_N = \frac{1}{C}\mathbf{1}_C = \mathbf{u}, \tag{31}$$

where $\mathbf{u}$ denotes the uniform distribution over the $C$ sensitive attributes. Next, for each trajectory, the KL divergence satisfies:

$$D_{\mathrm{KL}}(\mathbf{q}_i|\mathbf{p}_i) = \sum_{c=1}^{C} q_{i,c} \log \frac{q_{i,c}}{p_{i,c}} = \sum_{c=1}^{C} q_{i,c} \log q_{i,c} - \sum_{c=1}^{C} q_{i,c} \log p_{i,c}. \tag{32}$$

Using the entropy

$$H(\mathbf{q}_i) = -\sum_{c=1}^{C} q_{i,c} \log q_{i,c}, \tag{33}$$

we obtain

$$-\sum_{c=1}^{C} q_{i,c} \log p_{i,c} = D_{\mathrm{KL}}(\mathbf{q}_i|\mathbf{p}_i) + H(\mathbf{q}_i). \tag{34}$$

Averaging over all $N$ trajectories yields

$$\mathcal{L}_{\mathrm{BE}}(\mathbf{X}_t) = \frac{1}{N}\sum_{i=1}^{N} D_{\mathrm{KL}}(\mathbf{q}_i|\mathbf{p}_i) + \frac{1}{N}\sum_{i=1}^{N} H(\mathbf{q}_i), \tag{35}$$

and hence

$$\mathcal{L}_{\mathrm{BE}}(\mathbf{X}_t) - \frac{1}{N}\sum_{i=1}^{N} H(\mathbf{q}_i) = \frac{1}{N}\sum_{i=1}^{N} D_{\mathrm{KL}}(\mathbf{q}_i|\mathbf{p}_i). \tag{36}$$

We next bound the deviation of the aggregate attribute distribution from the uniform target. Since $\bar{\mathbf{q}} = \mathbf{u}$,

$$\bar{\mathbf{p}} - \mathbf{u} = \bar{\mathbf{p}} - \bar{\mathbf{q}} = \frac{1}{N}\sum_{i=1}^{N} (\mathbf{p}_i - \mathbf{q}_i)\,. \tag{37}$$

By the triangle inequality and the relation $|\mathbf{v}|_2 \le |\mathbf{v}|_1$,

$$|\bar{\mathbf{p}} - \mathbf{u}|_2 \le \frac{1}{N}\sum_{i=1}^{N} |\mathbf{p}_i - \mathbf{q}_i|_2 \;\le \frac{1}{N}\sum_{i=1}^{N} |\mathbf{p}_i - \mathbf{q}_i|_1 \tag{38}$$

By Pinsker's inequality,

$$|\mathbf{p}_i - \mathbf{q}_i|_1 \le \sqrt{2 D_{\mathrm{KL}}\left(\mathbf{q}_i|\mathbf{p}_i\right)}. \tag{39}$$

Substituting the above inequality into Eq. 38 and applying Jensen's inequality to the concave function $\sqrt{\cdot}$ gives

$$|\bar{\mathbf{p}} - \mathbf{u}|_2 \le \frac{1}{N}\sum_{i=1}^{N} \sqrt{2 D_{\mathrm{KL}}\left(\mathbf{q}_i|\mathbf{p}_i\right)} \;\le \sqrt{\frac{2}{N}\sum_{i=1}^{N} D_{\mathrm{KL}}\left(\mathbf{q}_i|\mathbf{p}_i\right)}. \tag{40}$$

By definition, the trajectory bias $\mathcal{B}(t)$ can be written as

$$\mathcal{B}(t) = \left[\frac{1}{C}\sum_{c=1}^{C}\left(\bar{p}_c - \frac{1}{C}\right)^2\right]^{1/2} = \frac{1}{\sqrt{C}}\,|\bar{\mathbf{p}} - \mathbf{u}|_2\,. \tag{41}$$

Combining this equality with Eq. 40 yields

$$\mathcal{B}(t) \le \sqrt{\frac{2}{C}\cdot\frac{1}{N}\sum_{i=1}^{N} D_{\mathrm{KL}}\left(\mathbf{q}_i|\mathbf{p}_i\right)} \;= \sqrt{\frac{2}{C}\left(\mathcal{L}_{\mathrm{BE}}(\mathbf{X}_t) - \frac{1}{N}\sum_{i=1}^{N} H(\mathbf{q}_i)\right)}, \tag{42}$$

where the second equality follows from Eq. 36. □

# B Analysis of Trajectory-level Bias

## B.1 Accuracy of Trajectory-level Observation

To visualize the Tweedie denoised estimates at intermediate latent states, Figure 3 presents their evolution across timesteps for two representative prompts. Even when $\mathbf{x}_t$ remains highly noisy at early denoising stages, the estimates already reveal key semantics of the final generation, including demographic attributes such as gender, race, and age. Moreover, these estimates rapidly converge to the final generated images, with the estimates at timestep 5 already closely resembling the final




Figure 3: Illustration of Tweedie's denoising estimates $\varphi(\hat{\mathbf{x}}_0^\theta(\mathbf{x}_t, t))$ across timesteps.

outputs. This observation suggests that trajectory bias can emerge early and persist throughout the denoising process.

We further evaluate the accuracy of gender prediction from trajectory attribute scores computed on Tweedie denoising estimates, presented in Figure 4. Across four DMs with diverse denoiser architectures, prediction accuracy rapidly exceeds 80% at early denoising stages and remains consistently high throughout generation. This result is consistent with the visualized Tweedie estimates, where gender-related semantics are already observable at early steps. These findings demonstrate that trajectory attribute scores provide a reliable signal for monitoring and regulating denoising trajectories dynamically.

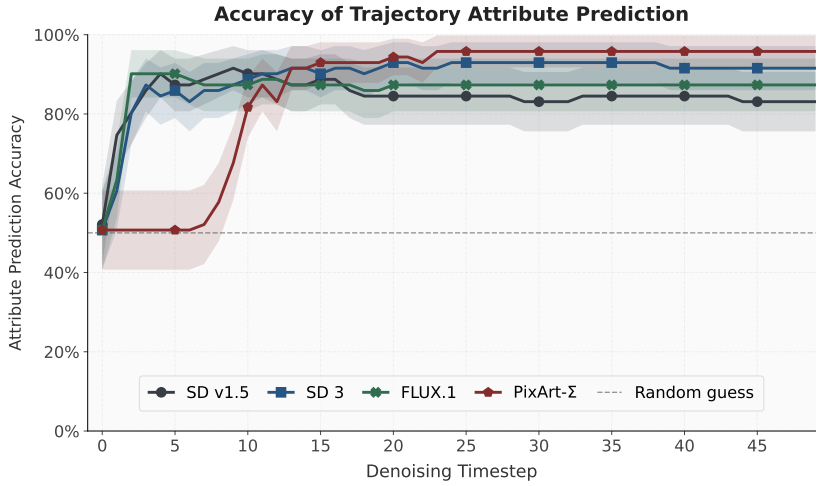



Figure 4: Accuracy of gender prediction based on trajectory attribute score across timesteps.

### B.2 Trajectory Bias

Based on the trajectory attribute scores, we further reveal the severe imbalance in trajectory exploration in Figure 5. The heatmap shows that trajectory bias emerges from the early denoising steps and becomes pronounced within the first 10 steps. Once established, the bias persists throughout the remaining denoising process, suggesting that early imbalance in trajectory exploration can carry through to the final generated samples. We notice that PixArt-$\Sigma$ develops trajectory bias later than the other three DMs, which may be related to its comparatively later emergence of semantic information (Görgün et al., 2026).

We further visualize the trajectory bias after debiasing with our trajectory-level framework in Figure 6. Although the imbalance emerges at the early denoising stages, our framework intervenes in these trajectories by enforcing a balanced exploration budget and optimizing the intermediate latents. As a result, the generated trajectories exhibit substantially more balanced attribute exploration, demon-

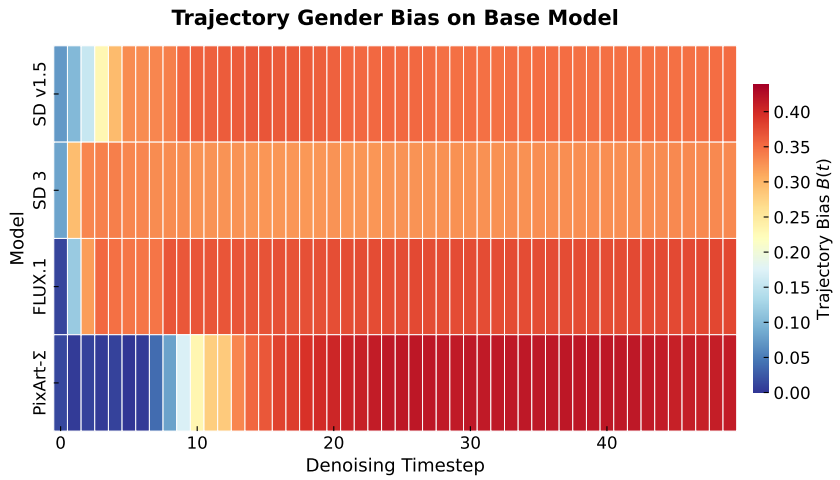



Figure 5: Gender trajectory bias $\mathcal{B}(t)$ across models and timesteps on base model.

strating the effectiveness of our trajectory-level regulation. These results also support the theoretical analysis that our balanced allocation strategy constrains the development of trajectory-level bias, as formalized in Proposition 1.

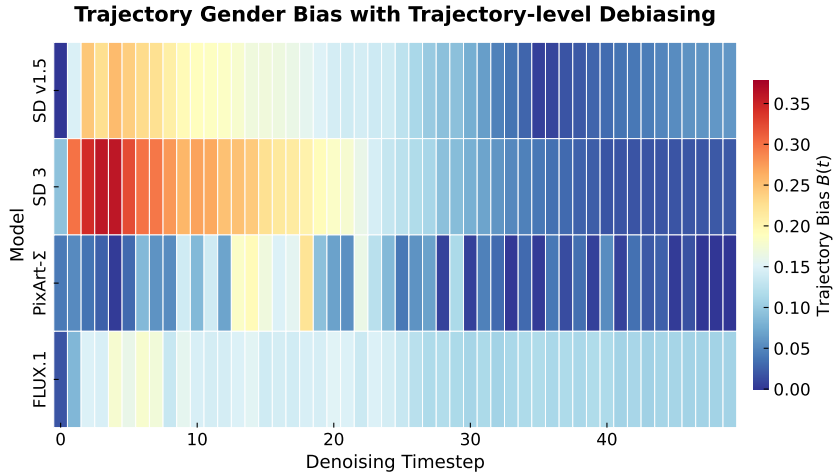



Figure 6: Gender trajectory bias $\mathcal{B}(t)$ across models and timesteps with our debiasing framework.

### B.3 Alignment with Training Corpus

To trace the origin of trajectory-level biases, we compare the trajectory attribute scores with the gender ratios of samples in the LAION dataset, using the gender statistics for the occupation subset reported in Friedrich et al. (2023) . We present the comparison results on SD v1.5 and SD 3 in Figure 7 and 8, respectively, as both of which use LAION as part of their training corpus.

The strong alignment suggests that trajectory-level bias is closely associated with gender imbalance in the training corpus and that intermediate trajectory observations align well with final outputs. We also observe several counter-stereotype cases in the cumulative bias gap. These cases mainly occur in generations with multiple figures, blurred or incomplete faces (e.g., masks or helmets), or back views, where gender-related visual cues are less explicit.

[2]LAION dataset is the training corpus of multiple Stable Diffusion models, and the subset contains over 1.8M images across 150 occupations Friedrich et al. (2023). Missing ratios are not included in the subset.

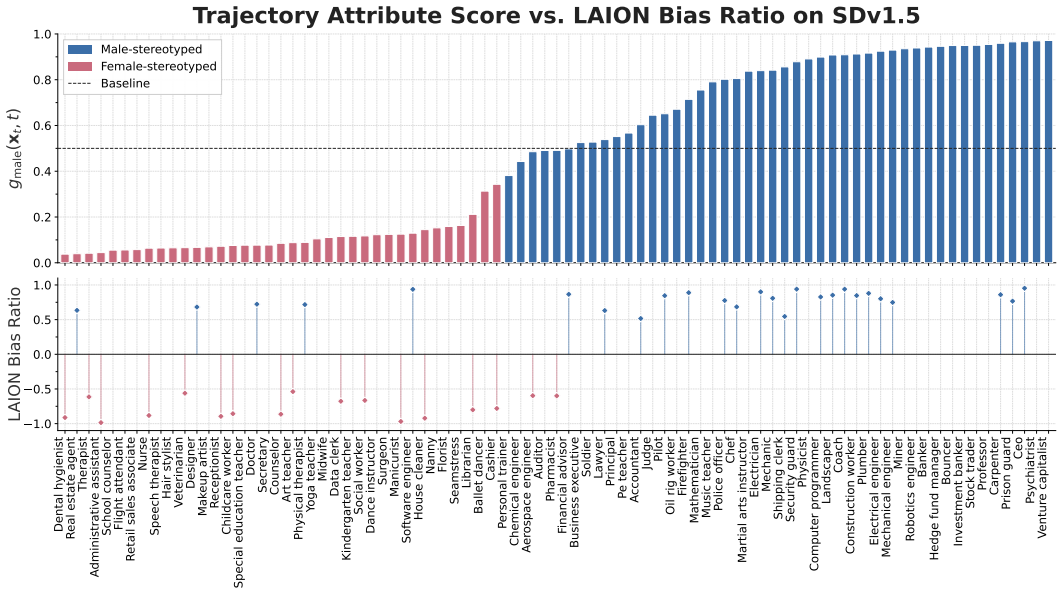



Figure 7: Comparison of trajectory attribute scores with the gender ratios of LAION samples on SD v1.5. Stereotype categories are defined based on the generation results of each DM.

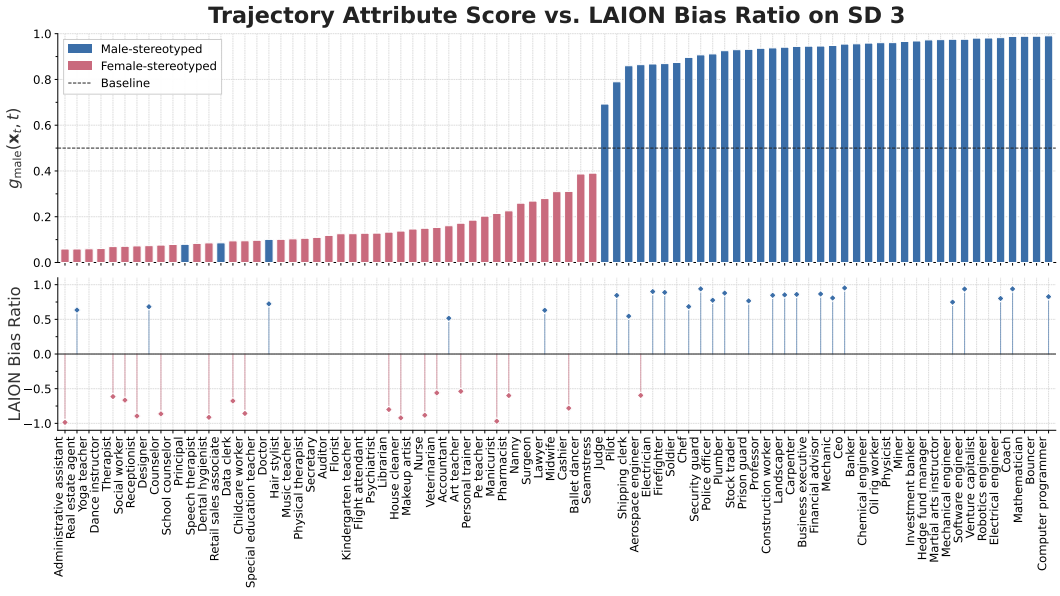



Figure 8: Comparison of trajectory attribute scores with the gender ratios of LAION samples on SD 3. Stereotype categories are defined based on the generation results of each DM.

# C Additional Details

## C.1 Implementation Details

In our main experiments, we evaluate fair text-to-image generation on 100 male- or female-stereotyped occupations from the extended WinoBias dataset (Fu et al., 2026). For all models, image generation is performed with 50 denoising steps. We use a classifier-free guidance scale $\gamma = 7.5$ for SD v1.5, SD 3, and PixArt-$\Sigma$, and $\gamma = 3.5$ for FLUX.1. Each experiment is conducted with a batch size of 16 and repeated three times using different random seeds.

For attribute prediction, we use lightweight pretrained classifiers based on FairFace (Karkkainen & Joo, 2021), since the classifiers are trained on balanced datasets to reduce potential bias in attribute prediction.

For our trajectory-level debiasing framework, we apply trajectory regulation within the normalized timestep interval $[0.10, 0.80]$ with a regulation strength $\lambda = 1.0$. The ablation studies of different intervention windows and attribute classifiers are evaluated in Appendix E.4.

### C.2 Background of Denoiser Architectures

Denoising networks are the core components of DMs, and their architectures have evolved substantially to improve generation quality, efficiency, and scalability (Ma et al., 2025).

Early diffusion models employed convolutional architectures for denoising, with DDPM (Ho et al., 2020) establishing the modern diffusion framework using a U-Net backbone (Ronneberger et al., 2015). The U-Net model adopts a hierarchical encoder-decoder structure, with the encoder progressively downsamples the input to capture features with broader spatial context, while the decoder restores spatial resolution through upsampling and skip connections. This structure yields a compact bottleneck layer that integrates high-level information, making it a natural intervention point for existing model-specific debiasing methods. Subsequent latent diffusion models (LDMs) (Rombach et al., 2022), such as Stable Diffusion v1.5, further improve computational efficiency by performing diffusion in a learned latent space and introduce cross-attention to incorporate conditioning signals such as text (Rombach et al., 2022).

Transformer-based denoisers later emerged as an alternative to U-Net architectures. DiT (Peebles & Xie, 2023) replaces the convolutional U-Net with a Transformer that operates on tokenized latent patches, allowing global interactions through self-attention and providing a highly scalable architecture. Its scaling behavior enables generation quality to improve systematically with increased model capacity and computation. Models such as PixArt-$\Sigma$ further build on the DiT framework with more efficient token processing and high-quality training data, enabling improved prompt alignment and high-resolution generation (Chen et al., 2024).

More recent text-to-image models further integrate multimodal processing within the denoiser. Stable Diffusion 3 adopts a Multimodal Diffusion Transformer (MMDiT) (Esser et al., 2024), which uses separate parameterizations for text and image streams while enabling joint attention between their tokens. This design allows bidirectional information exchange between the two modalities, improving text understanding and text-to-image interaction during generation (Esser et al., 2024).

In parallel, recent models such as FLUX (Labs et al., 2025) further develop Transformer-based generative models under the rectified-flow paradigm. Rather than using the conventional diffusion formulation, rectified flow learns a vector field along a straighter path between noise and data, while FLUX employs a hybrid architecture combining multimodal and single-stream Transformer blocks. This combination provides a scalable framework for high-quality text-to-image generation.

Despite these architectural advances, existing debiasing methods are often designed around specific denoiser structures, making their intervention mechanisms difficult to generalize across architectures. In contrast, our framework adopts a trajectory-oriented perspective and directly regulates the denoising trajectory, enabling architecture-agnostic fair generation across diverse denoiser architectures. We validate our approach on four representative diffusion models spanning U-Net, DiT, MMDiT, and rectified-flow Transformer architectures, demonstrating consistent effectiveness through balanced trajectory exploration.

### C.3 Details of Metrics

In our experiments, we evaluate the fairness, accuracy, and image quality of debiased generations. In this section, we detail the computation of these metrics.

**Fairness.** We evaluate fairness using two complementary metrics: Fairness Discrepancy (FD) (Parihar et al., 2024) and Fairness Score (Fu et al., 2026). FD measures the Euclidean distance between the generated attribute distribution and a target reference distribution, typically the uniform distribution. Given an attribute $a$, we use a pretrained classifier $C$ to obtain the softmax output $y = C_a(x)$

for each generated image $x$. The FD is then computed as

$$\mathrm{FD} = \left\| \bar{p} - \mathbb{E}_{x \sim p_\theta(x)}[y] \right\|_2 ,$$

where $\bar{p}$ is a uniform vector with the same dimension as $y$, and $p_\theta$ denotes the distribution of generated images. A lower FD indicates that the generated attribute distribution is closer to the target uniform distribution.

In contrast, Fairness Score measures the balance of group representation across generated images. The classifier detects and counts members of each group in each generated image, from which the group proportions $p = (p_1, \ldots, p_k)$ are obtained. Following previous research (Fu et al., 2026), the Fairness Score is defined as

$$\mathrm{FS} = 1 - \frac{\sum_i \left| p_i - \frac{1}{k} \right|}{2\left(1 - \frac{1}{k}\right)},$$

where $k$ is the number of groups. A score of 1 indicates equal representation across all groups, whereas a score of 0 indicates that all generated instances belong to a single group.

**Accuracy.** We evaluate the accuracy of debiased generation using two CLIP-based metrics, CLIP-T and CLIP-I, with ViT-L/14(Radford et al., 2021) as the CLIP model.

CLIP-T measures the semantic alignment between the generated image and its text prompt. Given the embedding of text prompt $e_{\text{text}}$ and the embedding of generated image $e_{\text{img}}^{\text{gen}}$, the CLIP-T score is computed as:

$$\text{CLIP-T} = \frac{e_{\text{text}} \cdot e_{\text{img}}^{\text{gen}}}{\|e_{\text{text}}\| \left\| e_{\text{img}}^{\text{gen}} \right\|}.$$

where a higher CLIP-T score indicates stronger alignment between the generated image and the text condition.

**Image quality.** We use MUSIQ (Ke et al., 2021) to measure the perceptual quality of generated images, which is based on a pretrained multi-scale image quality model that aggregates visual features using Transformer encoders and outputs a scalar quality score. A higher MUSIQ score indicates better perceptual quality, reflecting lower distortion and fewer generation artifacts, as well as better agreement with human subjective quality assessment.

# D Qualitative Results

We visualize the gender and racial debiasing results of our framework in Figures 9 and 10. The qualitative results show that our method effectively broadens attribute coverage in generated samples, while preserving image quality and semantic coherence.

# E More Experimental Results

## E.1 Generation of Intersectional Bias

Besides debiasing individual bias such as gender or race, real-world biases are often multifaceted and may arise from their intersections. In this section, we evaluate our debiasing framework on the simultaneous regulation of gender and racial biases. Specifically, we consider the intersectional attribute space $\mathcal{A}_{\text{gender}\times\text{race}} = \mathcal{A}_{\text{gender}} \times \mathcal{A}_{\text{race}}$, which covers all combinations of gender and race attributes. This setting imposes a more demanding fairness requirement by considering a larger number of attribute groups and their intersecting biases.

Table 3 reports the fairness, accuracy, and image quality results for intersectional debiasing, compared with architecture-agnostic baselines. When simultaneously regulating multiple attribute groups, our method consistently improves fairness while preserving image quality. In contrast, some baselines, such as Weak Guidance, achieve fairness improvements at the cost of substantial image quality degradation. These results demonstrate the potential of trajectory-level debiasing for fair generation under more complex and realistic attribute combinations.




Figure 9: Qualitative results of gender debiasing across four diffusion models for the prompt *A photo of the face of a firefighter*. Colored borders indicate male and female attributes, respectively.

### E.2 Generation of Non-sensitive Prompts

Beyond fairness on prompts containing sensitive concepts, a practical debiasing framework should also preserve image quality and prompt alignment on general, non-sensitive prompts. Yet this aspect remains largely overlooked in existing debiasing work, where model-specific interventions can come at the cost of general generation quality (Kim et al., 2025). We therefore evaluate our framework on non-sensitive prompts to examine whether debiasing introduces unintended degradation in general image generation.

We use 50 non-sensitive prompts without human figures, covering diverse categories such as animals, landscapes, and objects. We compare our results with those of the corresponding base models

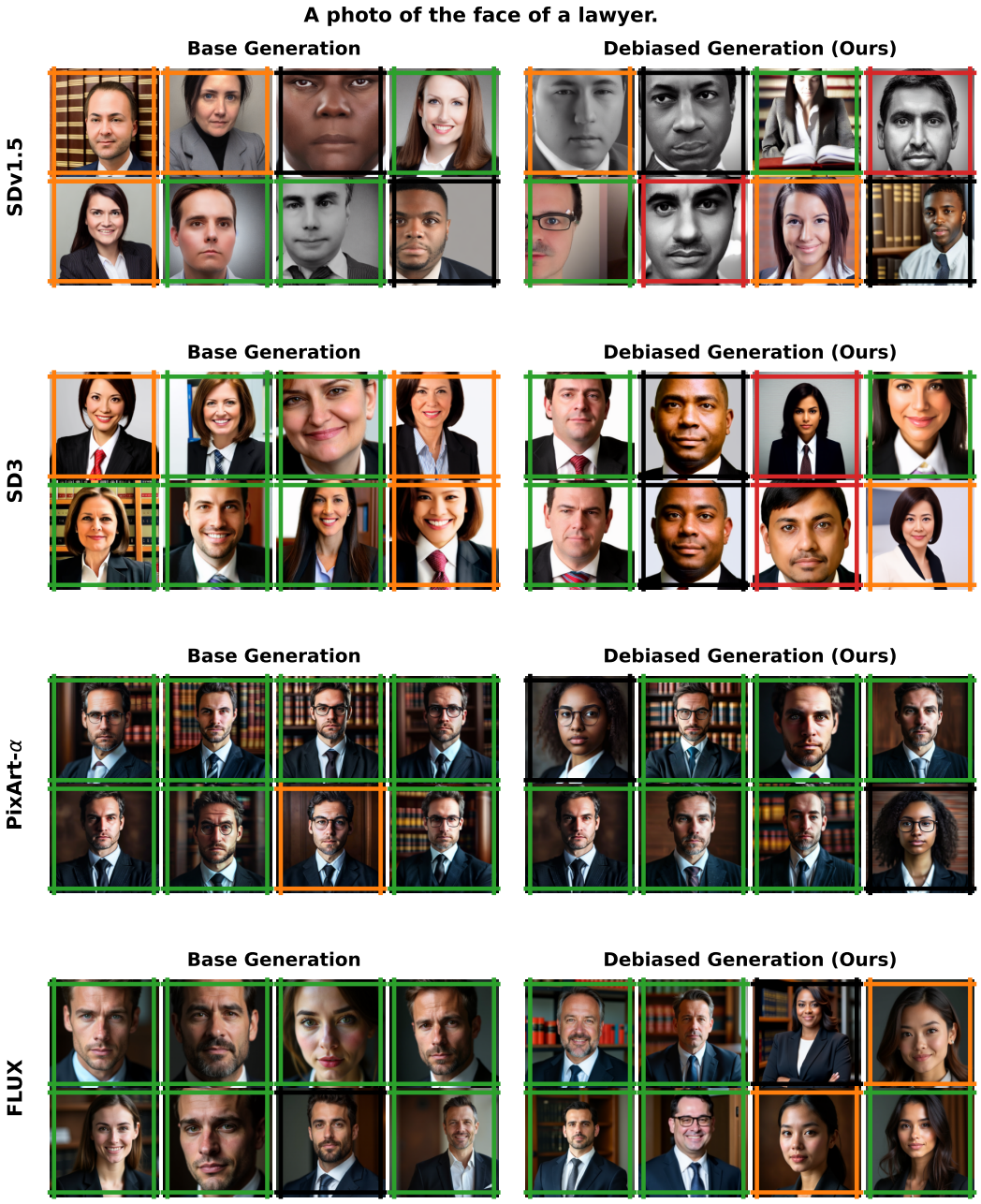



Figure 10: Qualitative results of racial debiasing across four diffusion models for the prompt *A photo of the face of a lawyer*. Colored borders indicate WMELH, Asian, Black, and Indian, respectively.

using three metrics. In addition to CLIP-T for prompt alignment and MUSIQ for image quality, as discussed in Section 4.1 and Appendix C.3, we use CLIP-I to measure the similarity between the embeddings of images generated by the base and debiased models, reflecting the consistency of the generated content.

Table 4 presents the results of generation on non-sensitive prompts. Our framework achieves image quality and semantic alignment comparable to the base models, while maintaining high consistency with the original generations. These results suggest that trajectory-level debiasing can improve fairness without substantially affecting generation on general, non-sensitive prompts.

Table 3: Results of fairness, accuracy, and quality on four DMs under gender $\times$ race setting.

| Method | FD ↓ | Fairness ↑ | CLIP-T ↑ | MUSIQ ↑ |
|---|---|---|---|---|
| **SD v1.5** | | | | |
| Base | $0.374_{\pm 0.032}$ | $0.274_{\pm 0.028}$ | $0.722_{\pm 0.001}$ | $0.725_{\pm 0.002}$ |
| Controlled | $0.294_{\pm 0.010}$ | $0.320_{\pm 0.062}$ | $0.652_{\pm 0.009}$ | $\mathbf{0.730}_{\pm 0.010}$ |
| Weak Guid. | $0.218_{\pm 0.006}$ | $0.511_{\pm 0.152}$ | $0.608_{\pm 0.003}$ | $0.319_{\pm 0.018}$ |
| FairImagen | $0.229_{\pm 0.009}$ | $0.531_{\pm 0.168}$ | $0.639_{\pm 0.005}$ | $0.616_{\pm 0.012}$ |
| *Ours* | $\mathbf{0.200}_{\pm 0.007}$ | $\mathbf{0.602}_{\pm 0.205}$ | $\mathbf{0.686}_{\pm 0.003}$ | $0.688_{\pm 0.004}$ |
| **SD 3** | | | | |
| Base | $0.489_{\pm 0.025}$ | $0.221_{\pm 0.172}$ | $0.723_{\pm 0.001}$ | $0.744_{\pm 0.012}$ |
| Controlled | $0.357_{\pm 0.068}$ | $0.275_{\pm 0.078}$ | $0.702_{\pm 0.007}$ | $0.735_{\pm 0.015}$ |
| Weak Guid. | $0.329_{\pm 0.016}$ | $0.428_{\pm 0.042}$ | $0.698_{\pm 0.005}$ | $\mathbf{0.741}_{\pm 0.003}$ |
| FairImagen | $0.297_{\pm 0.011}$ | $0.456_{\pm 0.118}$ | $0.694_{\pm 0.028}$ | $0.648_{\pm 0.005}$ |
| *Ours* | $\mathbf{0.237}_{\pm 0.009}$ | $\mathbf{0.566}_{\pm 0.142}$ | $\mathbf{0.707}_{\pm 0.001}$ | $0.712_{\pm 0.003}$ |
| **PixArt-$\Sigma$** | | | | |
| Base | $0.442_{\pm 0.006}$ | $0.182_{\pm 0.191}$ | $0.733_{\pm 0.001}$ | $0.720_{\pm 0.004}$ |
| Controlled | $0.379_{\pm 0.003}$ | $0.219_{\pm 0.065}$ | $0.710_{\pm 0.005}$ | $\mathbf{0.708}_{\pm 0.006}$ |
| Weak Guid. | $0.280_{\pm 0.009}$ | $0.423_{\pm 0.038}$ | $0.648_{\pm 0.004}$ | $0.375_{\pm 0.007}$ |
| FairImagen | $\mathbf{0.270}_{\pm 0.004}$ | $0.433_{\pm 0.185}$ | $0.700_{\pm 0.000}$ | $0.636_{\pm 0.006}$ |
| *Ours* | $0.288_{\pm 0.008}$ | $\mathbf{0.472}_{\pm 0.188}$ | $\mathbf{0.726}_{\pm 0.022}$ | $0.704_{\pm 0.021}$ |
| **FLUX.1** | | | | |
| Base | $0.471_{\pm 0.014}$ | $0.221_{\pm 0.153}$ | $0.702_{\pm 0.001}$ | $0.709_{\pm 0.006}$ |
| Controlled | $0.383_{\pm 0.007}$ | $0.294_{\pm 0.056}$ | $0.696_{\pm 0.001}$ | $\mathbf{0.733}_{\pm 0.002}$ |
| Weak Guid. | $0.417_{\pm 0.004}$ | $0.344_{\pm 0.189}$ | $\mathbf{0.702}_{\pm 0.000}$ | $0.711_{\pm 0.006}$ |
| FairImagen | $0.287_{\pm 0.004}$ | $0.370_{\pm 0.189}$ | $0.686_{\pm 0.000}$ | $0.616_{\pm 0.006}$ |
| *Ours* | $\mathbf{0.286}_{\pm 0.021}$ | $\mathbf{0.466}_{\pm 0.125}$ | $0.695_{\pm 0.004}$ | $0.684_{\pm 0.006}$ |

Table 4: Results of generation on non-sensitive prompts on four DMs.

| | SD v1.5 | | | SD 3 | | |
|---|---|---|---|---|---|---|
| **Method** | **CLIP-I ↑** | **CLIP-T ↑** | **MUSIQ ↑** | **CLIP-I ↑** | **CLIP-T ↑** | **MUSIQ ↑** |
| Base | N/A | $0.746_{\pm 0.002}$ | $0.714_{\pm 0.004}$ | N/A | $0.695_{\pm 0.003}$ | $0.706_{\pm 0.012}$ |
| *Ours* | $0.932_{\pm 0.001}$ | $0.738_{\pm 0.004}$ | $0.705_{\pm 0.004}$ | $0.915_{\pm 0.002}$ | $0.686_{\pm 0.006}$ | $0.716_{\pm 0.014}$ |
| | **PixArt-$\Sigma$** | | | **FLUX.1** | | |
| Base | N/A | $0.747_{\pm 0.002}$ | $0.709_{\pm 0.003}$ | N/A | $0.708_{\pm 0.001}$ | $0.725_{\pm 0.003}$ |
| *Ours* | $0.967_{\pm 0.002}$ | $0.707_{\pm 0.006}$ | $0.711_{\pm 0.003}$ | $0.951_{\pm 0.001}$ | $0.708_{\pm 0.004}$ | $0.753_{\pm 0.002}$ |

### E.3 Cost Analysis

For real-world deployment, including online text-to-image generation services, debiasing should avoid prohibitive inference-time and memory overhead. To assess the practicality of our trajectory-level debiasing framework, we compare the inference time per image (s) and peak memory usage (GB) between base generation and debiased generation.

Table 5 reports the comparison. Our method increases inference time by approximately $6\times$ and peak memory usage by approximately $1.5\times$ relative to base generation. The additional overhead mainly comes from computing trajectory attribute scores with a pretrained classifier and optimizing intermediate latent states during denoising. Despite this overhead, the proposed framework requires no additional model training or modification of model parameters, and the entire debiasing process can be performed during GPU-based inference.

Although this additional computation enables fairer and quality-preserved generation, reducing the inference overhead remains an important direction for improving the efficiency of trajectory-level

Table 5: Results of inference time and peak memory on four DMs.

| Method | SD v1.5 | | SD 3 | |
|---|---|---|---|---|
| | **Time (s)** | **Peak Memory (GB)** | **Time (s)** | **Peak Memory (GB)** |
| Base | $6.210_{\pm 0.327}$ | $4.400_{\pm 0.000}$ | $8.590_{\pm 0.319}$ | $8.060_{\pm 0.000}$ |
| *Ours* | $36.400_{\pm 0.216}$ | $6.450_{\pm 0.000}$ | $43.070_{\pm 0.014}$ | $18.450_{\pm 0.002}$ |
| | **PixArt-Σ** | | **FLUX.1** | |
| Base | $25.040_{\pm 0.116}$ | $14.500_{\pm 0.001}$ | $40.460_{\pm 1.750}$ | $18.910_{\pm 0.000}$ |
| *Ours* | $148.050_{\pm 0.031}$ | $27.220_{\pm 0.000}$ | $191.100_{\pm 1.414}$ | $24.260_{\pm 0.005}$ |

* Measured on NVIDIA RTX 4090 GPU.

debiasing. Future work could improve efficiency through multi-GPU parallelization to increase aggregate throughput, adaptive early stopping to reduce per-sample latency, and training lightweight estimators to for more efficient trajectory attribute estimation.

### E.4 Ablation Studies

To evaluate the robustness of our trajectory-level debiasing framework under different settings, we conduct ablation studies on intervention steps, prompt variants, and classifiers. We experiment on SD 3 as the representative DM, with the implementation details and results presented below.

**Ablation on intervention steps.** Our framework regulates intermediate latent states within an intervention window $\mathcal{T}_{\text{int}}$, encouraging the denoising trajectory toward balanced attribute exploration. We vary the start of the intervention window over $[0.00, 0.05, 0.10, 0.15, 0.20]$ and the end over $[0.30, 0.40, 0.60, 0.80, 0.90]$, measured as normalized fractions of the total denoising steps. As shown in Figure 11, our method consistently achieves a Fairness Score above 0.50 across different intervention windows, demonstrating its robustness to the choice of intervention steps. Interestingly, the highest fairness is generally achieved when the intervention covers the middle denoising stages (e.g., 0.10-0.80), suggesting that these stages play an important role in developing core semantics.

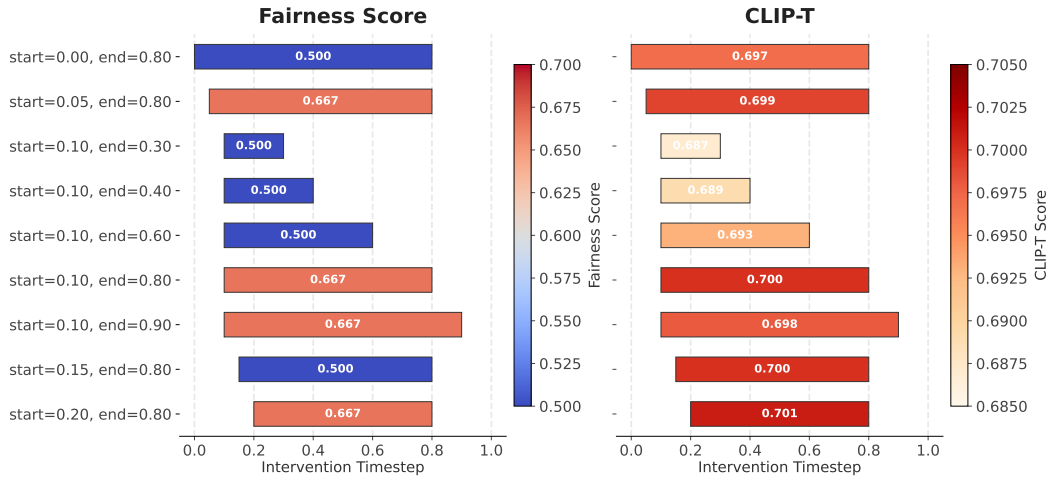



Figure 11: Results of ablation study on intervention steps.

**Ablation on prompt variants.** To examine the consistency of our debiasing method under different prompt formulations, we evaluate three prompt variants: *A photo of a/an* {*occupation*}, *A/an* {*occupation*}, and *A photo of a professional* {*occupation*}. Figure 12 presents the fairness and prompt-alignment results across these variants. Our method maintains a consistent debiasing effect across all prompt formulations, achieving a fairness score above 0.50 in each case. At the same time, CLIP-T remains above 0.68 with only minor variation, indicating that the debiased generations preserve the semantics of the input prompts.

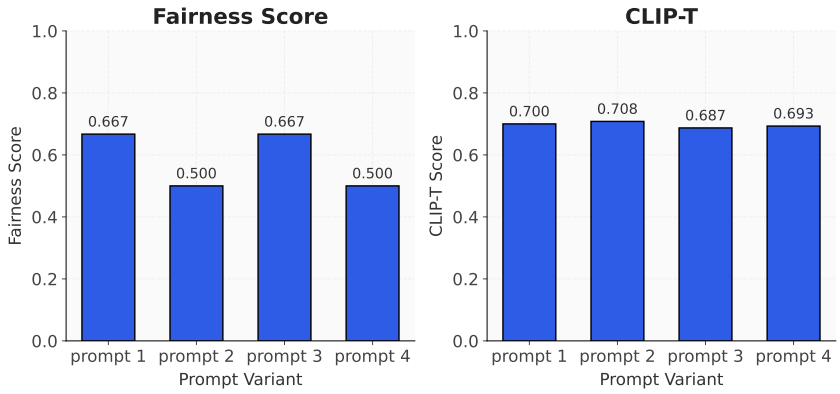



Figure 12: Results of ablation study on prompt variants.

**Ablation on classifiers.** To compute trajectory attribute scores, we use a pre-trained classifier to estimate attributes from the Tweedie-denoised estimates. To examine whether our debiasing results depend on a specific classifier, we conduct an ablation study with different classifier variants. In addition to the gender classifier described in Section C.1, we evaluate two other lightweight gender classifiers with high accuracy, obtained from their official repositories. Figure 13 presents the fairness and prompt-alignment results across classifier variants. The debiasing performance remains stable across all classifiers, with the second classifier achieving an even higher fairness score. The CLIP-T scores are also consistent across variants, confirming the robustness of our trajectory-level debiasing to classifier choice.

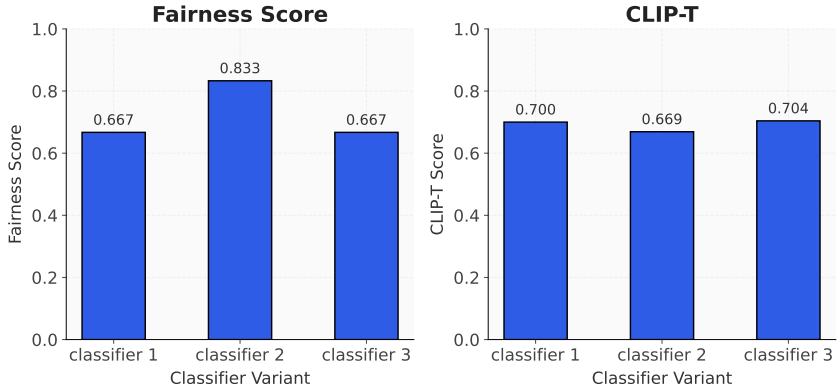



Figure 13: Results of ablation study on classifier variants.

## F Case Study: Fair Dermatological Image Generation

Beyond the experiments on stereotyped occupations (Section 4.2), we further investigate whether our trajectory-level debiasing framework can support fairer medical image generation, with dermatology as a representative application. Medical AI systems increasingly rely on large-scale image datasets for disease diagnosis, yet demographic imbalances in these datasets can limit their generalization across patient populations (Yang et al., 2024). In dermatology, in particular, darker skin tones have been underrepresented in publicly available skin lesion datasets, which has been associated with disparities in the performance of dermatological AI systems across skin tones (Daneshjou et al., 2022; Groh et al., 2021). This motivates the use of generative models as a data-centric approach for improving demographic diversity. Rather than requiring the costly collection and annotation of additional clinical images, generative models can provide complementary samples for underrepresented attribute groups. Existing approaches address this challenge by fine-tuning diffusion models with fairness-aware objectives, such as re-sampling strategies (Zhang et al., 2024) and explicit conditioning (Munia & Imran, 2025), to generate more representative dermatological

Table 6: Results on fair dermatological image generation.

| Method | FD ↓ | Fairness ↑ | CLIP-T ↑ | MS-SSIM ↓ | DSPR(%) ↑ |
|---|---|---|---|---|---|
| DermDiff | $0.192_{\pm 0.012}$ | $0.765_{\pm 0.022}$ | $0.257_{\pm 0.006}$ | $0.829_{\pm 0.009}$ | $\mathbf{100.00\%}_{\pm 0.00\%}$ |
| *Ours* | $\mathbf{0.165}_{\pm 0.008}$ | $\mathbf{0.797}_{\pm 0.018}$ | $\mathbf{0.272}_{\pm 0.005}$ | $\mathbf{0.745}_{\pm 0.007}$ | $93.33\%_{\pm 0.20\%}$ |

datasets. However, these approaches require dedicated training on attribute-balanced medical data, which limits their applicability when such data are unavailable.

In contrast, our framework provides a training-free alternative that regulates attribute exploration directly during the denoising process, enabling an existing dermatological diffusion model to generate more balanced samples at inference time. In this case study, we focus on skin cancer images and consider three skin-tone groups: light, brown, and dark. To adapt our framework to this setting, we build on a diffusion model specifically designed for dermatological image generation, which is based on SD v1.4 Rombach et al. (2022). We use a CLIP-based zero-shot classifier to estimate skin-tone attribute scores from intermediate latent states (Eq. 7). Following our trajectory-level formulation, we allocate a balanced exploration budget across the three skin-tone attributes and regulate the denoising trajectories during inference to encourage more balanced coverage of these attribute modes. Importantly, the underlying diffusion model is kept unchanged, and our method is applied solely at inference time on top of the dermatology-specific model.

We compare our method with DermDiff (Munia & Imran, 2025), which conditions diffusion generation on explicit skin-tone attributes and requires dedicated model training for dermatological data augmentation. We evaluate both generation quality and attribute balance using **FD** and **Fairness Score** across the three skin-tone groups. We additionally report **CLIP-T** to evaluate text-image alignment and **MS-SSIM** to characterize the structural diversity of the generated images. A lower MS-SSIM indicates less structural similarity across generated images, reflecting greater structural diversity among the generated samples. Besides, to examine whether the generated samples preserve the intended disease semantics, we report the **disease semantic preservation rate (DSPR)**, defined as the proportion of generated images classified as skin cancer by a pretrained dermatological diagnosis model.

Table 6 presents the comparison results on fair dermatological image generation. Our framework transfers effectively to the medical imaging setting, achieving more balanced coverage of skin colors even without explicitly specifying these attributes in the prompts. Compared with existing baselines, our method improves the fairness score while requiring only a lightweight classifier to guide the regulation of denoising latents. Moreover, regulating the denoising trajectories leads to greater structural diversity while preserving disease-related semantics, which is critical for downstream medical applications.




Figure 14: Qualitative results of skin color debiasing on dedicated dermatological model for the prompt *Malignant skin lesion.* Colored borders indicate light, brown, and dark skin colors, respectively.