Stay Fair! Ensuring Group Fairness in Diffusion Models Across Guidance Scales

Myeongsoo Kim[1] **Eunji Kim**[2,*] **Minwoo Chae**[1] **Sangwoo Mo**[1]
[1]POSTECH [2]Amazon

Abstract

Diffusion models steer conditional generation with a tunable guidance scale to trade off prompt alignment and diversity. However, existing debiasing techniques are optimized for a single scale, degrading fairness when users adjust this parameter. We trace this behavior to a previously overlooked source by decomposing total bias into two components: a model bias and a guidance bias. While prior work primarily targets the former, we show that the *guidance bias* varies with the guidance scale, persisting even in models that have already been debiased. To address this, we extend Strong Demographic Parity to guidance and derive a condition under which the target distribution retains its group ratio across guidance scales. Building on this condition, we propose StayFair, which designs fair guidance algorithms in both regimes. For classifier guidance, it equalizes the classifier's output distributions across groups; for classifier-free guidance, it shifts the null embedding by a prompt-dependent offset. Because StayFair modifies only the guidance step, it is orthogonal to model debiasing and can be layered onto existing fair diffusion models. Across class-conditional and text-to-image generation, StayFair decouples fairness from the guidance scale without sacrificing image quality. Our decomposition also enables a more detailed analysis of the sources of bias in diffusion models.

Project page: https://kim-myeong-soo.github.io/stayfair/

1 Introduction

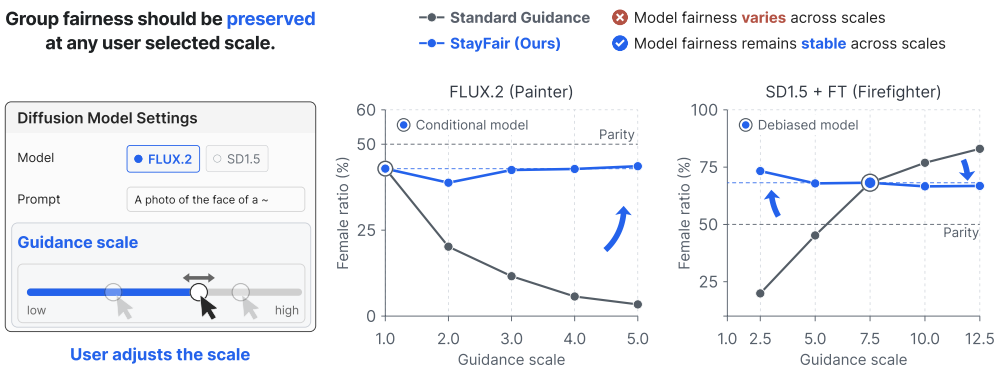
Recent advances in diffusion models have enabled high-quality conditional generation, particularly in text-to-image synthesis [1, 2]. Guidance [3, 4] is central to the widespread adoption of diffusion models, and its scale, chosen by the user, sets how strongly conditioning information influences generation. However, higher guidance scales amplify not only the condition but also its inherent social bias, increasing major-group prevalence in generated samples [5]. This shift can therefore undermine group fairness by moving the distribution away from the target distribution, such as the data distribution [6] or a uniform distribution [7], depending on the notion of fairness.

This raises the question: *Do existing models maintain their reported fairness when users change the guidance scale?* Prior work typically evaluates fairness at a fixed scale, leaving its preservation across scales unexamined. Social bias and its amplification in vanilla diffusion models are evaluated at a fixed guidance scale [8], while debiasing methods fit and evaluate models at a fixed scale [7]. Figure 1 illustrates the resulting concern: bias in the vanilla model increases with guidance scale, while the debiased model's group distribution drifts from its target. Thus, group fairness should not only hold at a single guidance scale but also remain robust as users adjust the scale to their preferences.

We formalize this requirement as *guidance scale-invariant group fairness*, extending group fair generation from a fixed scale to the range of scales users choose. To understand why it breaks, we propose a *bias decomposition* that separates total bias into a static *model bias* and a scale-dependent

*The work was done prior to joining Amazon.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Problem (b) Vanilla Model (c) Debiased Model

Figure 1: **StayFair maintains group fairness as users vary the guidance scale.** (a) Users adjust the guidance scale to generate outputs that match their preferences. Since fairness should hold regardless of users' scale choices within the specified range, we extend the group fairness requirement across that range of scales. (b,c) Prior work that examines bias or optimizes fairness at a fixed guidance scale is insufficient for fairness across scales: group ratios drift in both vanilla (FLUX.2 [10]) and debiased (SD1.5 + FT [7]) models. StayFair stabilizes group ratios across scales, mitigating average and worst total bias depending on the setting.

guidance bias. Unlike prior work targeting model bias at a fixed scale, we pursue *fair guidance* to mitigate this scale-dependent guidance bias, thereby improving average and worst total bias as guidance varies. To provide a rigorous foundation, we extend Strong Demographic Parity (SDP) [9] to guidance, identifying a condition under which the target distribution remains fair across guidance scales. Building on this principle, we introduce StayFair, which addresses guidance bias in both major guidance regimes: for classifier guidance (CG), it approximates this condition via a debiased classifier with group fairness constraints, while for classifier-free guidance (CFG), it uses an adaptive null prompt to rebalance bias without sacrificing prompt alignment, across model architectures.

We evaluate whether StayFair stabilizes group ratios across scales while preserving image quality, and whether this improves group fairness. We observe this behavior in class-conditional generation with CG on ADM-G [3] (Section 5.1) and text-to-image generation with CFG on SD1.5 [1], SD3 [2], and FLUX.2 [10] (Section 5.2). Moreover, StayFair mitigates average and worst total bias across guidance scales in vanilla models (Section 5.2.1), and worst total bias while maintaining average debiasing performance in debiased models (FT [7] and UCE [11]; Section 5.2.2). Modifying the guidance mechanism is thus an effective way to mitigate guidance bias, even compared with existing debiasing methods extended across guidance scales. We also examine how guidance bias contributes to asymmetric gender bias amplification [12] in SD3 and FLUX.2 (Section 5.3). StayFair further extends to multi-group attributes such as race and to unseen prompts (Section 5.4).

2 Related Work

Group fairness in diffusion models. Diffusion models exhibit biased distributions over sensitive attributes such as gender, race, age, and culture [13–16]. Recent text-to-image (T2I) diffusion models often amplify these biases, exaggerating the group imbalances in their training data [8, 17]. The underlying mechanisms remain poorly understood, as exemplified by the stronger amplification for male-associated occupations than for female-associated ones [12, 18]. StayFair decomposes total bias into model and guidance components, helping explain bias amplification.

Debiasing diffusion models. One line of work, *attribute-based methods*, conditions on sensitive attributes through guidance [19, 20], prompt embeddings [21], or latent manipulation [22] to control group distributions. However, they risk visual artifacts or unintended changes to non-target content [21, 23]. Another line, *train-time methods*, either fine-tunes the model with fairness-aware objectives [7, 24] or updates specific modules such as text embeddings [25, 26] or cross-attention layers [11]. These methods mitigate learned biases by updating model parameters, allowing generation from the original prompt without additional attribute conditioning. Both lines of work focus on

fairness at a fixed guidance scale, which does not guarantee fairness across scales. StayFair addresses this gap by reducing guidance bias.

Guidance in diffusion models. Guidance improves image quality and condition alignment at the cost of diversity [1, 2]. Common variants include CG [3, 29, 30], CFG [4, 31, 32], and autoguidance (AG) [33–35], differing in how the conditioning signal is constructed. Recent work [5, 36] reports that varying the guidance scale also shifts demographic composition, exposing a fairness dimension to this tradeoff. However, no prior method directly reduces guidance bias. StayFair debiases guidance to maintain fairness across user-controlled scales.

Group fairness in classifiers. Group fairness has been widely studied in computer vision [37–39]. Generally, it requires that some statistic of the predictor be matched across sensitive groups. Common criteria such as demographic parity [40] and equalized odds [41] are defined through group-wise averages of predictions. These criteria are typically addressed through constrained classification [42, 43] or group-distributionally robust training [44, 45]. A stronger notion, SDP [9], requires prediction distributions to match across groups. Prior work enforces SDP by minimizing discrepancies between group-conditional distributions, using Wasserstein distance [46], total variation [47], or optimal-transport objectives [48]. Building on this view, we characterize the classifier fairness required to preserve group ratios under guidance and extend SDP to guidance functions.

3 Decomposing Bias in Guided Diffusion

In this section, we first review diffusion models and guidance (Section 3.1). We extend group fair generation from a fixed guidance scale to a range of user-selected scales and introduce a decomposition of total bias into model and guidance components (Section 3.2). Finally, we outline two scenarios: guidance bias emerging in vanilla models and persisting in debiased models (Section 3.3).

3.1 Preliminary: diffusion models and guidance

Diffusion models [49, 50] generate clean data $X_0 \sim p_0(x)$ by reversing a forward corruption process $\{X_t\}$ that progressively perturbs X_0 with Gaussian noise. Following the EDM formulation [51], we consider $X_t = X_0 + \sigma_t \epsilon$ with $\epsilon \sim \mathcal{N}(0, I)$, where σ_t denotes the noise level at time t. This induces the marginal density of X_t as $p_t(x) = p_0(x) * \mathcal{N}(x; 0, \sigma_t^2 I)$. In the probability flow formulation, generation proceeds by denoising X_T into X_0 via

$$dX_t = -\dot{\sigma}_t \, \sigma_t \, \nabla \log p_t(X_t) \, dt. \tag{1}$$

We approximate the score using a noise-prediction network $\epsilon_\theta(x_t, t)$ [50]. We write the model score as $\nabla \log p_\theta(x_t) = -\epsilon_\theta(x_t, t)/\sigma_t$, suppressing t when it is clear from the noisy input x_t.

Guidance was introduced to sample from a target distribution $p_\theta^w(x_t \mid y) \propto p_\theta(x_t) p_\theta(y \mid x_t)^w$ at each timestep t, where the guidance scale w controls the strength of the condition y. It replaces the score in Eq. (1) with a guided score $\nabla \log p_\theta^w(x_t \mid y)$:

$$\nabla \log p_\theta^w(x_t \mid y) = \nabla \log p_\theta(x_t) + w \, \nabla \log f(x_t, t), \tag{2}$$

where $f(x_t, t)$ is a guidance function that typically approximates $p(y \mid x_t)$. We denote by $X_t \sim q_t^w$ the marginal density induced by integrating Eq. (1) with this guided score, and by $Q^w := q_0^w$ the actual generated distribution [52–54].

The regimes in this paper differ in how they realize f. *Classifier guidance* [3] trains the time-dependent classifier $f_\phi(x_t, t)$ with noisy input x_t and timestep t. *Classifier-free guidance* [4] takes an implicit classifier via Bayes' rule

$$\nabla \log p_\theta(y \mid x_t) = \nabla \log p_\theta(x_t \mid y) - \nabla \log p_\theta(x_t \mid \emptyset) \tag{3}$$

where $\emptyset$ denotes the null prompt, representing the unconditional model. *Autoguidance* [33] replaces $\nabla \log p_\theta(x_t \mid \emptyset)$ with $\nabla \log p_{\theta_{\text{weak}}}(x_t \mid y)$, using a weaker model under the same y.

The *group distribution* gives the probabilities of sensitive attribute groups $a \in \mathcal{A}$ in the generated distribution q_0^w. Let $g : \mathcal{X} \to \mathcal{A}$ map images to groups and $\mathbf{1}\{\cdot\}$ denote the indicator function. Then,

$$Q^w(a \mid y) := \int_{\mathcal{X}} \mathbf{1}\{g(x_0) = a\} \, q_0^w(x_0 \mid y) \, \mathrm{d}x_0.$$

3.2 Problem formulation

We consider generation under a condition y at a user-chosen guidance scale w, with images assigned to sensitive attribute groups $a \in \mathcal{A}$. Let $Q^w(a \mid y)$ denote the resulting group distribution and $T(a \mid y)$ the target, *e.g.*, uniform over $\mathcal{A}$ or the group distribution of the data. *Group fair generation* asks that $\mathrm{Bias}^w(a \mid y) := Q^w(a \mid y) - T(a \mid y) = 0$, typically at a single scale w.

Our goal is to achieve group fairness, $\mathrm{Bias}^w(a \mid y) = 0$, for any chosen scale $w \in [w_{\min}, w_{\max}]$. We thus introduce a *bias decomposition* into model bias $\mathrm{Bias}_{\mathrm{M}}(a \mid y)$ and guidance bias $\mathrm{Bias}^w_{\mathrm{G}}(a \mid y)$.

Bias decomposition

$$\underbrace{\mathrm{Bias}^w(a \mid y)}_{\text{Total bias}} = \underbrace{Q^w(a \mid y) - Q^{w_{\mathrm{ref}}}(a \mid y)}_{\text{Guidance bias: } \mathrm{Bias}^w_{\mathrm{G}}(a|y)} + \underbrace{Q^{w_{\mathrm{ref}}}(a \mid y) - T(a \mid y)}_{\text{Model bias: } \mathrm{Bias}_{\mathrm{M}}(a|y)}, \tag{4}$$

Our goal requires a fair model ($\mathrm{Bias}_{\mathrm{M}}$=0) matching the target distribution at w_{ref} and *fair guidance* satisfying $\mathrm{Bias}^w_{\mathrm{G}}$=0 for any scale $w \in [w_{\min}, w_{\max}]$ the user chooses.

Definition 1 (Fair guidance)**.** Guidance is *fair* if $\mathrm{Bias}^w_{\mathrm{G}}(a \mid y) = 0$ for all $w \in [w_{\min}, w_{\max}]$ and groups $a \in \mathcal{A}$, *i.e.*, the group distribution is invariant to w.

We pursue fair guidance that can be used together with the debiased diffusion models developed in prior work. Existing approaches focus on model bias, leaving guidance bias largely unaddressed. We therefore target the guidance function f in Eq. (2).

3.3 The overlooked impact of guidance bias

We reveal the overlooked consequences of $\mathrm{Bias}^w_{\mathrm{G}}$ in conventional diffusion sampling through two scenarios: its emergence in *vanilla models* and persistence in *debiased models* (Figure 1). First, a vanilla model follows its data distribution; applying guidance amplifies that distribution toward the majority group, more strongly as the scale grows. Second, even debiased models lose fairness as w deviates from w_{ref}, because their group distributions drift from the target without scale-aware guidance. Together, these observations indicate that *guidance scale-invariant* group fairness is not reached by model debiasing, even when it is fitted across a range of guidance scales (Table 2(b)); it requires a direct intervention on the guidance mechanism itself.

4 StayFair: Fair Diffusion Across Guidance Scales

To reduce guidance bias, we derive a sufficient condition for fairness across scales (Section 4.1), use it to design StayFair for CG and CFG (Section 4.2), and show how StayFair combines with existing diffusion models (Section 4.3).

4.1 Sufficient condition for fair guidance

We study the group-conditional distribution of the guidance function $f : \mathcal{X} \times [0, T] \to \mathbb{R}_+$. Since $p^w_t(\cdot \mid y) \propto p^0_t(\cdot \mid y) f(\cdot, t)^w$, group-dependent bias in f transfers to p^w_t and grows with w. We therefore extend SDP [9] to f.

Definition 2 (SDP of the guidance function)**.** The guidance function f satisfies Strong Demographic Parity if, for all $t \in [0, T]$ and $a, a' \in \mathcal{A}$,

$$\mathrm{Law}\big(f(X_t, t) \mid g(X_0){=}a,\, y\big) \;=\; \mathrm{Law}\big(f(X_t, t) \mid g(X_0){=}a',\, y\big). \tag{5}$$

Under SDP, the target distribution $p^w_t(\cdot \mid y)$ preserves the group ratios across guidance scales.

Theorem 1 (SDP preserves fairness in the target distribution)**.** *Suppose that $f(\cdot, t)$ satisfies SDP (Definition 2) with respect to $p_t(\cdot \mid y)$. Then, for every $w \geq 0$, $t \in [0, T]$, and $a, a' \in \mathcal{A}$, we have*

$$p^w_t(a \mid y)/p^w_t(a' \mid y) \;=\; p^0_t(a \mid y)/p^0_t(a' \mid y). \tag{6}$$

Consequently, $p^w_t(\cdot \mid y)$ retains its group ratio for every guidance scale w.

Theorem 2 shows how this controls the actual process q^w_t in a Gaussian setting (proof in Appendix E).

Theorem 2 (SDP transfers to the actual distribution; informal)**.** *In a Gaussian setting with a log-affine endpoint potential, if f satisfies Definition 2, then $\mathrm{Bias}^w_{\mathrm{G}}(a \mid y) = 0$ for all $w, w_{\mathrm{ref}} \geq 0$ and $a \in \mathcal{A}$.*

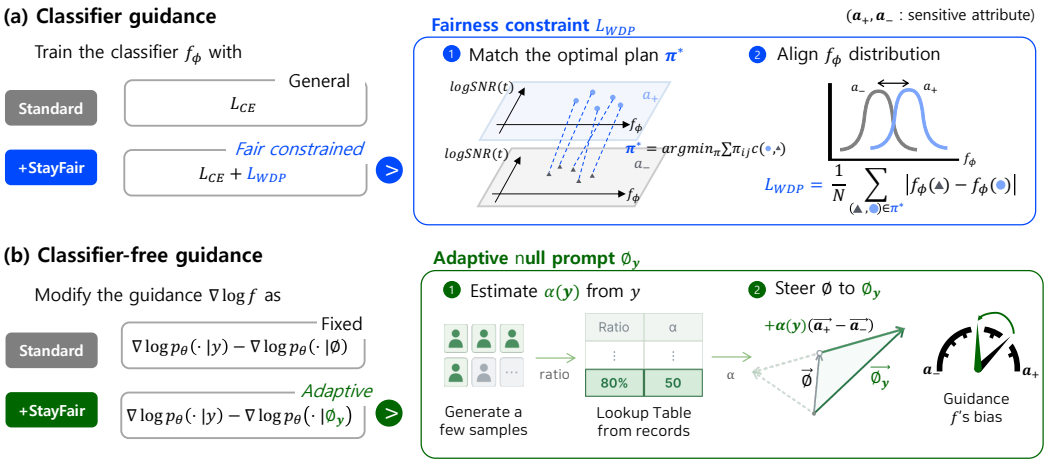


Figure 2: **Overview of StayFair.** In (a) *classifier guidance*, StayFair debiases the classifier f_ϕ by adding the fairness constraint $\mathcal{L}_{\text{WDP}}$. It matches samples between groups using a joint cost on noise level and f_ϕ value, aligning the groupwise f_ϕ distributions. In (b) *classifier-free guidance*, StayFair replaces the fixed null prompt $\emptyset$ with an adaptive $\emptyset_y$ by shifting its embedding along the attribute direction. $\alpha(y)$ is estimated from group ratios in a few generated images using searched α records.

4.2 Debiasing guidance in practice

Building on Theorem 1, StayFair approximately balances f in both guidance regimes. Figure 2 summarizes the CG and CFG instantiations, detailed below, and we then relate it to autoguidance.

Classifier guidance. In CG, the guidance function is an explicit classifier $f_\phi(\cdot, t)$, so we regularize its outputs toward SDP during training. We adopt Wasserstein Demographic Parity (WDP) [9], a tractable relaxation of SDP that measures the 1-Wasserstein distance between the $f_\phi(\cdot, t)$ distributions of two groups, *i.e.*, $\Delta\text{WDP}_t(a, a') = 0$ if and only if SDP holds at timestep t.

$$\Delta\text{WDP}_t(a, a') \;=\; \mathcal{W}_1\big(\text{Law}(f_\phi(X_t, t) \mid g(X_0){=}a, y),\, \text{Law}(f_\phi(X_t, t) \mid g(X_0){=}a', y)\big). \tag{7}$$

To enforce $\Delta\text{WDP}_t(a, a') = 0$, we minimize an empirical WDP loss between cross-group classifier outputs within a local timestep window to account for varying noise levels. For each group pair, we match outputs u_i and v_j using an optimal transport plan $\pi^* = \arg\min_{\pi \in \Pi} \sum_{i,j} \pi_{ij}\, c\big((u_i, t_i), (v_j, t'_j)\big)$, where Π is the set of balanced transport plans. The transport cost is noise-aware $c\big((u, t), (v, t')\big) = |u - v| + \gamma|\text{logSNR}(t) - \text{logSNR}(t')|$ with parameter γ. The empirical loss $\mathcal{L}_{\text{WDP}}(a, a') = \frac{1}{n} \sum_{i,j} \pi^*_{ij} |u_i - v_j|$ combines with the cross-entropy loss $\mathcal{L}_{\text{CE}}$ as

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \sum_{a,a' \in \mathcal{A}} \mathcal{L}_{\text{WDP}}(a, a'). \tag{8}$$

Classifier-free guidance. In CFG, $f(\cdot, t)$ is only implicitly defined (Eq. (3)), so direct regularization is infeasible. We instead aim to reduce group bias in f without enforcing SDP. We replace the fixed null prompt $\emptyset$ with an adaptive $\emptyset_y$ (Eq. (9)), whose embedding is shifted along attribute directions to correct the bias of $p_\theta(x_t \mid \emptyset_y)$ for prompt y (Eq. (10)).

$$\nabla \log f(x_t, t) := \nabla \log p_\theta(x_t|y) - \nabla \log p_\theta(x_t|\emptyset_y) \tag{9}$$

$$E(\emptyset_y) := E(\emptyset) + \sum_i \alpha_i(y) \frac{E(a_i^+) - E(a_i^-)}{\|E(a_i^+) - E(a_i^-)\|} \tag{10}$$

where E is the text encoder and $a_i^\pm$ are opposing attribute texts (*e.g.*, male/female); we omit i for a single attribute. Attribute directions follow embedding subtraction [55]. We choose the null prompt $\emptyset$ (*e.g.*, an empty string, a space, or a padding token) for each tokenizer so that special tokens occupy the same positions as in $a^\pm$. We use embedding-level control for architecture-agnostic application.

Adjusting α changes the guidance bias of $\nabla \log f$ in the opposite direction to the bias of $p_\theta(x_t \mid \emptyset_y)$, since the null-branch score $\nabla \log p_\theta(x_t \mid \emptyset_y)$ is subtracted in Eq. (9). Motivated by the observation that bias amplification is smaller when $p_\theta(a \mid y)$ and $p_\theta(a \mid \emptyset_y)$ have similar bias levels, we adjust the latter while retaining its role as the null branch in CFG.

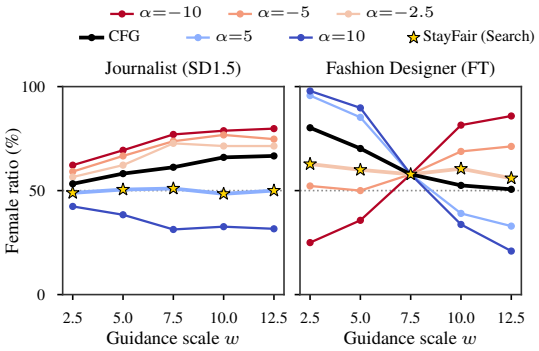


Figure 3: **Slope varies monotonically with α in both models, so we can find the flattest curve, which minimizes guidance bias.**

We now introduce two variants, StayFair (Search) and StayFair (Estimate), to select $\alpha(y)$ that minimizes guidance bias across scales. First, StayFair (Search) uses binary search to find $\alpha(y)$ that approximately equalizes group ratios at the lowest and highest scales. Since the curve's slope changes monotonically with α (Figure 3), the search selects the $\alpha(y)$ that yields the flattest curve. Second, StayFair (Estimate) assigns $\hat{\alpha}(y)$ from a precomputed lookup table indexed by the group ratios observed at the lowest and highest scales, leveraging the observation that similar initial proportions change similarly with scale (Figure 7). Procedures are in Appendix F.4; StayFair denotes StayFair (Search) unless stated otherwise.

Relation to autoguidance. StayFair lies between CFG and AG [33] in how it trades condition alignment against bias amplification. AG is the special case $\emptyset_y = y$ with a weaker model: it lowers guidance bias, but also gains less condition alignment. CFG is the case $\emptyset_y = \emptyset$: it gains alignment strongly, and carries a high guidance bias. StayFair instead makes $\emptyset_y$ independent of the condition's sensitive attribute and keeps CFG's alignment gain while reducing guidance bias toward AG's level. A comparison with PAG is in Appendix G (Table 10).

4.3 Combining with fair diffusion models

StayFair can be applied to a diffusion model and carries its group distribution across scales. Given the model $\epsilon_\theta^{\text{Fair}}$ at w_{ref} and a scale w, we predict $\epsilon_\theta^{\text{StayFair}}$ by adding fair guidance (Eq. (9)) for $w - w_{\text{ref}}$:

$$\epsilon_\theta^{\text{StayFair}}(x_t, t, y) = \underbrace{\epsilon_\theta^{\text{Fair}}(x_t, t, y)}_{\text{fair model at } w=w_{\text{ref}}} + (w - w_{\text{ref}}) \underbrace{\big(\epsilon_\theta(x_t, t, y) - \epsilon_\theta(x_t, t, \emptyset_y)\big)}_{\text{fair guidance}}. \tag{11}$$

The model $\epsilon_\theta^{\text{Fair}}$ has $\text{Bias}_{\text{M}}{=}0$ as its objective, since a vanilla model follows the data distribution at $w_{\text{ref}}{=}1$ (without guidance) and a debiased model follows the target distribution at its debiasing scale w_{ref}. Fair guidance has $\text{Bias}_{\text{G}}^{w}{=}0$ as its objective (Definition 1). Meeting both objectives gives $\text{Bias}^{w}(a \mid y) = 0$ for all w in the bias decomposition of Eq. (4).

Our main analysis evaluates fair guidance in standard sampling using both vanilla diffusion models and models debiased during training, without relying on attribute conditioning. Appendix D analyzes attribute-based methods Fair Diffusion [19] and Weak Guidance [5] across guidance scales, showing how adapting $\emptyset_y$ reduces their bias.

5 Experiments

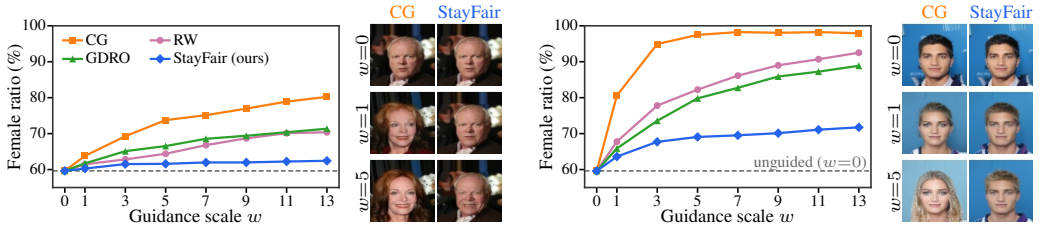
We evaluate StayFair in class-conditional generation with CG (Section 5.1) and in text-to-image generation with CFG (Section 5.2), covering vanilla (Section 5.2.1) and debiased generation (Section 5.2.2). We use our bias decomposition to analyze asymmetric gender bias amplification (Section 5.3). We also examine multi-group attributes and unseen prompts (Section 5.4). Our main comparisons assess whether StayFair stabilizes group ratios across scales while preserving condition alignment and image quality, and how reducing guidance bias affects average and worst total bias.

5.1 Class-conditional generation with CG

Setup. We apply CG to ADM-G [3] on CelebA [56] (64×64), using the unconditional model from Ning et al. [57]. We consider gender as the sensitive attribute and two conditions with different imbalance levels: *Smiling* (weak) and *Blond Hair* (strong). Gender labels are classified by using

Table 1: **StayFair reduces guidance bias while improving image quality in class-conditional generation with CG.** On CelebA, we report guidance bias (Range) and image quality using fair-intra versions of FID and $\text{FD}_{\text{DINOv2}}$. Both compare generated samples with a reference set matched to the target group ratios. Best in bold; our target metric shaded; $\pm$: std over six seeds.

	Smiling			Blond Hair		
Classifier	Range (%) ↓	FID ↓	$\text{FD}_{\text{DINOv2}}$ ↓	Range (%) ↓	FID ↓	$\text{FD}_{\text{DINOv2}}$ ↓
CG	20.6 $\pm$2.2	17.99	87.51	39.2 $\pm$3.5	22.70	151.04
RW	11.0 $\pm$3.4	16.96	83.39	32.9 $\pm$3.1	20.05	**136.55**
GDRO	11.9 $\pm$3.5	17.50	85.28	29.2 $\pm$2.8	22.29	152.70
StayFair (ours)	**3.5** $\pm$1.2	**16.66**	**82.56**	**12.4** $\pm$2.6	**19.68**	139.65



(a) Smiling (b) Blond Hair

Figure 4: **Effect of guidance scale under classifier guidance.** Female ratio (left) and samples (right) across w. As w increases, StayFair preserves minority groups while CG shifts toward the majority.

a pretrained classifier [58]. We sweep $w \in \{0, 1, 3, 5, 7, 9, 11, 13\}$, generating 256 images per condition and scale with six seeds for the gender ratio and 10,000 for image quality. Further implementation details are in Appendix F.

Baselines. We compare StayFair against a vanilla classifier and two debiased ones in the same CG pipeline: RW [45] (data-level) and GDRO [44] (optimization-level).

Metrics. Our target metric is the guidance bias, measured by the bias range ($\max - \min$ of the female ratio across w). Image quality is measured by *fair-intra* FID [59, 60] and $\text{FD}_{\text{DINOv2}}$ [61], using a reference set adjusted to the target distribution.

Experimental results. Table 1 shows that StayFair substantially reduces guidance bias in CG: the bias range drops from 20.6 to 3.5 on *Smiling* and from 39.2 to 12.4 on *Blond Hair*, while achieving the best fair-intra FID in both conditions and the best $\text{FD}_{\text{DINOv2}}$ on *Smiling*. While RW and GDRO also reduce the bias range, their gains remain suboptimal because they do not match guidance-score distributions across groups, whereas StayFair does.

Qualitative comparison. Figure 4 shows that StayFair maintains the unguided female ratio across w, whereas CG drifts toward the majority group. Compared with CG, StayFair's debiased classifier mitigates the collapse of minority-group modes, as suggested by the preserved visual context and minority gender of the unguided sample ($w = 0$).

5.2 Text-to-image generation with CFG

Setup. We measure gender–occupation bias at five guidance scales: for each of 33 occupations [8], we generate 100 images with three seeds from the prompt "`A photo of the face of a/an {occupation}`". The scales are evenly spaced: 2.5–12.5 on SD1.5 (default 7.5), 1.5–7.5 on SD3 (default 7.0), and 1.0–5.0 on FLUX.2 (default 4.0). Gender is labeled by zero-shot CLIP [62] after YOLO [63] person detection (Appendix F).

Metrics. Guidance bias is the bias range ($\max - \min$ female ratio across scales). We report average and worst total bias (the female ratio's absolute deviation from 50%) across scales. CLIP score [64] measures condition alignment; aesthetic score [65] and PickScore [66] measure image quality.

5.2.1 Vanilla generation

Baselines. We compare StayFair against CFG using vanilla SD1.5 [1], SD3-medium [2], and FLUX.2 [klein] 9B Base [10].

Table 2: **StayFair reduces guidance bias while maintaining image quality comparable to CFG.** For gender–occupation prompts, we report total bias (Avg., Worst), guidance bias (Range), and image quality over five guidance scales, with results averaged across occupations. StayFair reduces average and worst bias in vanilla generation (a), and lowers worst bias while maintaining average debiasing performance in debiased generation (b). StayFair achieves smaller bias ranges than both Multi baselines that extend debiasing across scales, highlighting the benefit of modifying the guidance mechanism. Best in bold; our target metric shaded; ±: std over three seeds.

(a) Vanilla generation		Bias (%) ↓			Quality ↑		
Model	Guidance	Avg.	Worst	Range	CLIP	Aesth.	Pick
SD1.5	CFG	28.0±0.3	32.5±0.5	10.7±0.7	27.61±0.02	**6.70**±0.01	**19.99**±0.02
	StayFair (ours)	**27.6**±0.6	**30.8**±0.5	**7.2**±0.3	**27.62**±0.02	**6.70**±0.01	19.96±0.01
SD3	CFG	39.7±0.5	43.0±0.4	9.1±0.3	27.36±0.01	**6.62**±0.00	**20.29**±0.00
	StayFair (ours)	**33.9**±0.5	**36.6**±0.5	**6.3**±0.2	**27.38**±0.02	6.61±0.00	20.23±0.00
FLUX.2	CFG	39.4±0.2	44.3±0.2	15.9±0.6	**28.18**±0.02	**6.93**±0.00	21.15±0.01
	StayFair (ours)	**32.7**±0.3	**36.0**±0.1	**8.4**±0.9	28.17±0.02	**6.93**±0.00	**21.18**±0.01

(b) Debiased generation			Bias (%) ↓			Quality ↑		
Method	Scales	Guidance	Avg.	Worst	Range	CLIP	Aesth.	Pick
UCE	Single	CFG	18.0±0.7	23.4±1.4	13.2±0.7	27.28±0.01	6.70±0.01	**19.96**±0.01
	Multi (random)	CFG	19.1±0.8	24.2±1.2	13.7±1.1	**27.38**±0.02	6.71±0.01	19.89±0.01
	Multi (per-scale)	CFG	**17.7**±0.6	29.4±0.7	30.6±0.4	27.32±0.02	**6.71**±0.00	19.92±0.01
	Single	**StayFair (ours)**	18.5±0.6	**22.2**±1.1	**7.6**±0.7	27.28±0.01	6.70±0.01	**19.96**±0.01
FT	Single	CFG	14.1±0.3	21.3±0.6	19.3±0.8	27.54±0.03	6.77±0.01	20.03±0.02
	Multi (random)	CFG	16.5±0.9	21.3±0.9	12.0±0.6	**27.63**±0.02	**6.79**±0.01	**20.07**±0.01
	Multi (per-scale)	CFG	**13.6**±0.4	22.0±0.6	17.9±0.7	27.62±0.03	6.77±0.01	20.03±0.02
	Single	**StayFair (ours)**	13.8±0.2	**18.2**±0.4	**9.4**±0.4	27.53±0.02	6.77±0.01	20.03±0.02

Experimental results. StayFair reduces bias ranges, lowering average and worst bias across models while maintaining image quality comparable to CFG. On FLUX.2, StayFair reduces the bias range from 15.9 to 8.4 and average and worst bias by 6.7 and 8.3 percentage points (pp), respectively (Table 2(a)). SD3 and FLUX.2 show larger bias reductions than SD1.5 because their evaluation includes lower scales, where guidance bias primarily arises. Figure 5 shows that StayFair keeps female ratios on FLUX.2 stable across scales. It thus maintains the model distribution, avoiding CFG's extreme bias in some occupations even within the recommended scale range ($w = 3.0$–5.0).

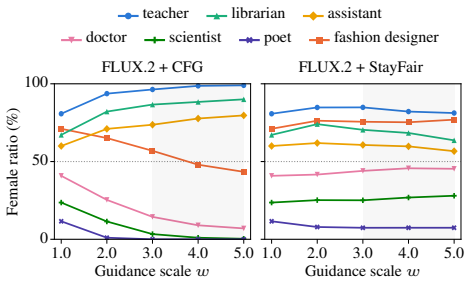


Figure 5: **Unlike CFG, StayFair maintains female ratios across guidance scales.** Further results are in Appendix G.

Qualitative comparison. Figure 6 shows that StayFair preserves the minority gender across guidance scales while retaining visual context similar to CFG; more examples are in Appendix H.

5.2.2 Debiased generation

Baselines. We use SD1.5 debiased with FT [7] or UCE [11] at w_{ref}=7.5 (Single) and compare StayFair against CFG. We also test whether extending existing debiasing methods across guidance scales can resolve guidance bias. Multi (random) samples w during debiasing, while Multi (per-scale) trains a separate model for each w (Appendix B).

Experimental results. First, StayFair reduces bias ranges with UCE and FT, improving worst bias while maintaining their average debiasing performance. Table 2(b) shows bias ranges falling from 13.2 to 7.6 with UCE and from 19.3 to 9.4 with FT, with worst bias reduced by 1.2 and 3.1 pp, respectively, and image quality comparable to CFG. Second, StayFair achieves lower bias ranges than both Multi variants: modifying guidance better preserves group distributions across scales than extending debiasing. Appendix B analyzes failure cases of both Multi variants.

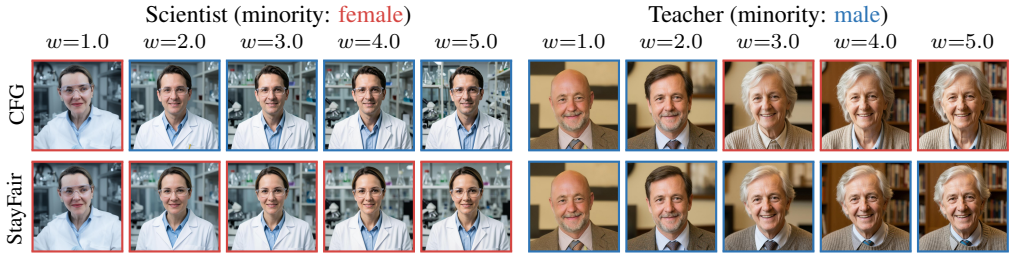
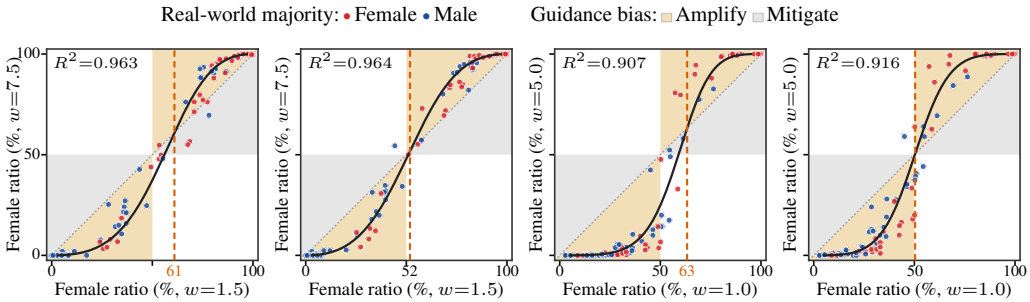


Figure 6: **StayFair preserves minority gender across guidance scales.** Border: female, male.



(a) SD3, CFG (b) SD3, $\alpha=-50$ (c) FLUX.2, CFG (d) FLUX.2, $\alpha=-2000$

Figure 7: **Guidance amplifies gender bias asymmetrically in T2I diffusion models.** Each point is an occupation–template pair for the 25 occupations with BLS statistics (Table 8). Solid curves are logit-space fits; vertical dashed lines mark crossings with $y = x$. In (a,c), deviations from $y = x$ indicate stronger amplification of male than female bias. In (b,d), asymmetry is reduced after shifting $\emptyset$ toward male to obtain $\emptyset_y$ (Eq. (10)).

5.3 Understanding bias amplification through bias decomposition

Setup. Using our bias decomposition, we examine how guidance bias contributes to T2I models overrepresenting male majorities more often than female ones relative to workforce ratios [12, 67]. Real-world majority labels follow U.S. workforce statistics [68]. Following Section 5.2, we assess CFG's effect by comparing female ratios at low and high scales over 4 prompt templates and the 25 occupations with available statistics (100 pairs; Table 8).

Experimental results. Guidance bias amplifies the real-world majority gender more often and more strongly in male-majority than in female-majority occupations. In Figures 7(a) and 7(c), guidance amplifies the majority gender in 75.0% of male-majority pairs but only 70.8% of female-majority pairs on SD3, and in 96.2% vs. 39.6% on FLUX.2. The amplification is also stronger, with a mean increase of 8.4 vs. 5.2 pp on SD3 and 18.2 vs. 13.1 pp on FLUX.2 among amplified pairs.

Our analysis indicates that female bias in the unconditional distribution $p_\theta(\cdot \mid \emptyset)$ biases guidance toward men (Eq. (9)), leading to stronger male than female amplification. In Figures 7(a) and 7(c), the crossings with $y = x$ exceed 50% (61% on SD3 and 63% on FLUX.2), reflecting female bias in $p_\theta(\cdot \mid \emptyset)$. In Figures 7(b) and 7(d), reducing this female bias mitigates asymmetric amplification. We reduce female bias in $p_\theta(\cdot \mid \emptyset)$ by shifting $\emptyset$ toward male via $\emptyset_y$ (Eq. (10)). The crossings shift to 52% on SD3 and 50% on FLUX.2, while the amplification gap across each crossing narrows ($3.8 \to 0.3$ pp on SD3 and $8.2 \to 4.7$ pp on FLUX.2). These results suggest that asymmetric amplification by guidance helps explain why T2I models overrepresent male majorities more often than female ones.

5.4 Additional analysis

Extension to multi-group attributes. Table 3 shows that StayFair reduces guidance bias and total bias across four racial groups. We set the attribute direction in Eq. (10) per occupation using the groups with the largest rise and fall in group ratio under CFG. These results support extending our null-prompt adjustment beyond binary attributes (Appendix A.3).

Table 3: **Results on race (SD3).**

Guidance	Bias (%) ↓ Avg.	Worst	Range
CFG	48.3±0.4	53.0±0.2	20.2±0.9
StayFair	**45.1**±0.1	**49.9**±0.4	**16.0**±0.3

Evaluating StayFair (Estimate) on unseen prompts. Table 4 shows that StayFair (Estimate) achieves intermediate performance between CFG and StayFair (Search) in the unseen prompt setting. To evaluate unseen prompts, we exclude all records for the test occupation–template pair when constructing the lookup table. We bootstrap group ratios from five-image samples at the lowest and highest scales to evaluate expected performance (Appendix C). These results suggest that our lookup-based approach can mitigate guidance bias on unseen prompts with a sampling budget of 10 images.

Table 4: **Results on unseen prompt (SD3).**

	Bias (%) ↓		
Guidance	Avg.	Worst	Range
CFG	38.8$_{\pm 0.1}$	42.2$_{\pm 0.2}$	9.9$_{\pm 0.3}$
StayFair (Estimate)	35.8$_{\pm 0.1}$	39.1$_{\pm 0.1}$	8.6$_{\pm 0.2}$
StayFair (Search)	34.4$_{\pm 0.2}$	37.5$_{\pm 0.3}$	7.8$_{\pm 0.4}$

6 Conclusion

We introduce guidance scale-invariant group fairness and show that total bias splits into model bias and guidance bias. StayFair mitigates guidance bias in CG and CFG, preserving image quality and group ratios across scales, and complements existing model debiasing. Our decomposition also explains how guidance contributes to asymmetric gender bias amplification in T2I diffusion models.

Limitations. First, StayFair addresses guidance bias, but achieving guidance scale-invariant group fairness also requires reducing model bias. Jointly debiasing the model and its guidance is a natural next step. Second, a performance gap remains between StayFair (Estimate) and StayFair (Search) on unseen prompts. Closing this gap requires more accurate estimation, and real-time use requires greater efficiency. Finally, our evaluation focuses on binary gender with a supplementary analysis of race, and extending StayFair to multi-group and intersectional settings remains future work.

Broader impact. CFG is widely used in diffusion research and applications such as synthetic-data generation and model distillation. Our work highlights the role of guidance in the bias of diffusion models by isolating it within total bias. We hope this finding helps practitioners use diffusion models more reliably and responsibly.

Acknowledgments and Disclosure of Funding

This work was partly supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (Nos. RS-2026-25608412, RS-2026-25607210) and the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2026-25524173, Ultra-Long-Term Hierarchical Memory and Reasoning Architecture for Next-Generation Omnimodal Agents).

References

[1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.

[2] 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 *ICML*, pages 12606–12633, 2024.

[3] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *NeurIPS*, volume 34, pages 8780–8794, 2021.

[4] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS Workshop on Deep Generative Models and Downstream Applications*, 2021.

[5] 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 *CVPR*, pages 13361–13370, 2025.

[6] Yeongmin Kim, Byeonghu Na, Minsang Park, JoonHo Jang, Dongjun Kim, Wanmo Kang, and Il-Chul Moon. Training unbiased diffusion models from biased dataset. In *ICLR*, 2024.

[7] Xudong Shen, Chao Du, Tianyu Pang, Min Lin, Yongkang Wong, and Mohan Kankanhalli. Finetuning text-to-image diffusion models for fairness. In *ICLR*, 2024.

[8] Preethi Seshadri, Sameer Singh, and Yanai Elazar. The bias amplification paradox in text-to-image generation. In *NAACL*, pages 6367–6384, 2024.

[9] Ray Jiang, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. Wasserstein fair classification. In *UAI*, pages 862–872, 2020.

[10] Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

[11] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *WACV*, pages 5111–5120, 2024.

[12] Leander Girrbach, Stephan Alaniz, Genevieve Smith, and Zeynep Akata. A large scale analysis of gender biases in text-to-image generative models. *arXiv preprint arXiv:2503.23398*, 2025.

[13] Jaemin Cho, Abhay Zala, and Mohit Bansal. DALL-Eval: Probing the reasoning skills and social biases of text-to-image generation models. In *ICCV*, pages 3043–3054, 2023.

[14] Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Evaluating societal representations in diffusion models. In *NeurIPS*, volume 36, pages 56338–56351, 2023.

[15] Charlotte Bird, Eddie Ungless, and Atoosa Kasirzadeh. Typology of risks of generative text-to-image models. In *AAAI/ACM AIES*, pages 396–410, 2023.

[16] Julia Barroso da Silveira and Ellen Alves Lima. Racial biases in AIs and Gemini's inability to write narratives about Black people. *Emerging Media*, 2(2):277–287, 2024.

[17] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *FAccT*, pages 1493–1504, 2023.

[18] Yankun Wu, Yuta Nakashima, and Noa Garcia. Stable Diffusion exposed: Gender bias from prompt to image. In *AAAI/ACM AIES*, 2024.

[19] 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.

[20] Rishubh Parihar, Abhijnya Bhat, Abhipsa Basu, Saswat Mallick, Jogendra Nath Kundu, and R. Venkatesh Babu. Balancing act: Distribution-guided debiasing in diffusion models. In *CVPR*, pages 6668–6678, 2024.

[21] Christopher T H Teo, Milad Abdollahzadeh, Xinda Ma, and Ngai-Man Cheung. FairQueue: Rethinking prompt learning for fair text-to-image generation. In *NeurIPS*, volume 37, pages 22878–22926, 2024.

[22] Hang Li, Chengzhi Shen, Philip Torr, Volker Tresp, and Jindong Gu. Self-discovering interpretable diffusion latent directions for responsible text-to-image generation. In *CVPR*, pages 12006–12016, 2024.

[23] 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 *ICCV*, pages 17567–17576, 2025.

[24] Eunji Kim, Siwon Kim, Chaehun Shin, and Sungroh Yoon. De-stereotyping text-to-image models through prompt tuning. In *ICML Workshop on Challenges in Deployable Generative AI*, 2023.

[25] Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*, 2023.

[26] Boyu Han, Qianqian Xu, Shilong Bao, Zhiyong Yang, Kangli Zi, and Qingming Huang. LightFair: Towards an efficient alternative for fair T2I diffusion via debiasing pre-trained text encoders. In *NeurIPS*, volume 38, pages 22671–22724, 2025.

[27] Sangwoo Mo, Chiheon Kim, Sungwoong Kim, Minsu Cho, and Jinwoo Shin. Mining GOLD samples for conditional GANs. In *NeurIPS*, 2019.

[28] Subin Kim, Sangwoo Mo, Mamshad Nayeem Rizve, Yiran Xu, Difan Liu, Jinwoo Shin, and Tobias Hinz. Rethinking prompt design for inference-time scaling in text-to-visual generation. In *CVPR*, 2026.

[29] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *CVPRW*, pages 843–852, 2023.

[30] Haotian Ye, Haowei Lin, Jiaqi Han, Minkai Xu, Sheng Liu, Yitao Liang, Jianzhu Ma, James Zou, and Stefano Ermon. TFG: Unified training-free guidance for diffusion models. In *NeurIPS*, volume 37, pages 22370–22417, 2024.

[31] Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. In *NeurIPS*, 2024.

[32] Seyedmorteza Sadat, Otmar Hilliges, and Romann M. Weber. Eliminating oversaturation and artifacts of high guidance scales in diffusion models. In *ICLR*, 2025.

[33] Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. In *NeurIPS*, volume 37, pages 52996–53021, 2024.

[34] Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. In *ICCV*, pages 7462–7471, 2023.

[35] Donghoon Ahn, Hyoungwon Cho, Jaewon Min, Wooseok Jang, Jungwoo Kim, SeonHwa Kim, Hyun Hee Park, Kyong Hwan Jin, and Seungryong Kim. Self-rectifying diffusion sampling with perturbed-attention guidance. In *ECCV*, pages 1–17, 2024.

[36] Nathan Roos, Ekaterina Iakovleva, Ani Gjergji, Vito Paolo Pastore, and Enzo Tartaglione. How I met your bias: Investigating bias amplification in diffusion models. In *WACV*, pages 5374–5383, 2026.

[37] Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM computing surveys (CSUR)*, 55(3):1–44, 2022.

[38] Younghyun Kim, Sangwoo Mo, Minkyu Kim, Kyungmin Lee, Jaeho Lee, and Jinwoo Shin. Discovering and mitigating visual biases through keyword explanation. In *CVPR*, 2024.

[39] Jiyun Bae, Hyunjong Ok, Sangwoo Mo, and Jaeho Lee. Understanding the effects of distractors on reasoning vision-language models. *arXiv preprint arXiv:2511.21397*, 2025.

[40] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *ITCS*, 2012.

[41] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *NeurIPS*, 2016.

[42] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *AISTATS*, 2017.

[43] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *ICML*, 2018.

[44] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *ICLR*, 2020.

[45] Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *CLeaR*, 2022.

[46] Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Fair regression with Wasserstein barycenters. In *NeurIPS*, volume 33, pages 7321–7331, 2020.

[47] Farhad Farokhi. Optimal pre-processing to achieve fairness and its relationship with total variation barycenter. *arXiv preprint arXiv:2101.06811*, 2021.

[48] Kunwoong Kim, Insung Kong, Jongjin Lee, Minwoo Chae, Sangchul Park, and Yongdai Kim. Fairness through matching. *Transactions on Machine Learning Research*, 2025.

[49] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, volume 33, pages 6840–6851, 2020.

[50] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021.

[51] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *NeurIPS*, volume 35, pages 26565–26577, 2022.

[52] Candi Zheng and Yuan Lan. Characteristic guidance: Non-linear correction for diffusion model at large guidance scale. In *ICML*, pages 61386–61412, 2024.

[53] Gen Li and Yuchen Jiao. Provable efficiency of guidance in diffusion models for general data distribution. In *ICML*, pages 35034–35046, 2025.

[54] Iskander Azangulov, Peter Potaptchik, Qinyu Li, Eddie Aamari, George Deligiannidis, and Judith Rousseau. Adaptive diffusion guidance via stochastic optimal control. In *AISTATS*, 2026.

[55] Guillaume Couairon, Matthijs Douze, Matthieu Cord, and Holger Schwenk. Embedding arithmetic of multimodal queries for image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4950–4958, June 2022.

[56] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738, 2015.

[57] Mang Ning, Enver Sangineto, Angelo Porrello, Simone Calderara, and Rita Cucchiara. Input perturbation reduces exposure bias in diffusion models. In *ICML*, pages 26245–26265, 2023.

[58] Kimmo Karkkainen and Jungseock Joo. FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *WACV*, pages 1548–1558, 2021.

[59] 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 *NeurIPS*, pages 6626–6637, 2017.

[60] Junhyun Nam, Sangwoo Mo, Jaeho Lee, and Jinwoo Shin. Breaking the spurious causality of conditional generation via fairness intervention with corrective sampling. *Transactions on Machine Learning Research*, 2023.

[61] George Stein, Jesse C. Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Leigh Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L. Caterini, J. Eric T. Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In *NeurIPS*, 2023.

[62] 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 *ICML*, pages 8748–8763, 2021.

[63] Ultralytics. Ultralytics YOLO11. URL https://docs.ultralytics.com/models/yolo11/.

[64] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *EMNLP*, pages 7514–7528, 2021.

[65] rsinema. Aesthetic scorer. URL https://huggingface.co/rsinema/aesthetic-scorer.

[66] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-Pic: An open dataset of user preferences for text-to-image generation. In *NeurIPS*, volume 36, pages 36652–36663, 2023.

[67] Shesh Narayan Gupta and Nik Bear Brown. Newer Is Not Fairer: Gender stereotyping in text-to-image ai across model generations. *arXiv preprint arXiv:2609.18007*, 2026.

[68] U.S. Bureau of Labor Statistics. Employed people by detailed occupation, sex, race, and Hispanic or Latino ethnicity. Current Population Survey, 2025 annual averages, Table 11, 2025. URL https://www.bls.gov/cps/data/aa2025/cpsa2025.pdf.

[69] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *CVPR*, pages 11472–11481, 2022.

[70] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Appendix

A Guidance Bias and Bias Amplification **16**
A.1 Toy example: Colored MNIST 16
A.2 Gender 17
A.3 Race 19

B Extending Debiasing Methods Across the Guidance Scale **20**

C StayFair (Estimate) on Unseen Prompts **21**

D Combining StayFair with Attribute-Based Methods **22**

E Theoretical Analysis **23**
E.1 Target distribution 23
E.2 Actual distribution 23

F Experimental Details **25**
F.1 Class-conditional generation 25
F.2 Text-to-image generation 25
F.3 Baseline implementations 26
F.4 StayFair implementation 26

G Additional Quantitative Results **28**
G.1 Additional per-occupation results 28

H Additional Qualitative Results **33**
H.1 Text-to-image generation examples 33
H.2 Class-conditional generation examples 35

I Limitations and Broader Impact **36**

J Licenses for Existing Assets **37**

A Guidance Bias and Bias Amplification

We extend the amplification analysis of Section 5.3. A Colored MNIST toy isolates the effect; we then show how it appears in text-to-image generation across models, occupations, and prompts.

A.1 Toy example: Colored MNIST

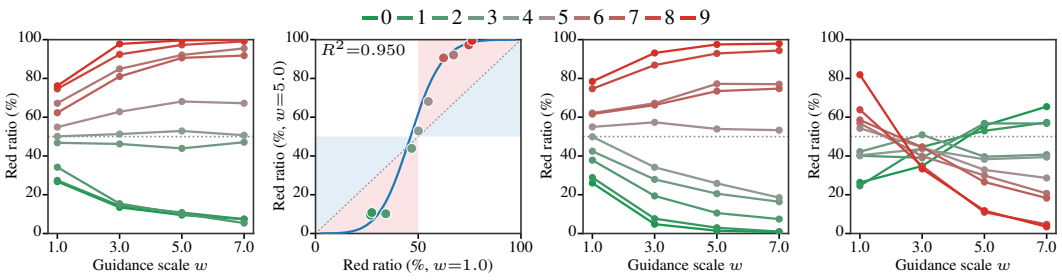
We use the toy example to observe the whole scale range: the text-to-image sweeps drop their lowest scales, where image quality degrades, and that is where amplification concentrates.

Setup. We construct a Colored MNIST dataset ($\mathcal{D}_{\text{org}}$) where digits represent occupations and colors (red or green) represent gender, with the red ratio rising linearly from 20% to 80% across digits so that every class has a known color bias. Using ADM [3], we train a class-conditional diffusion model and time-dependent classifiers on $\mathcal{D}_{\text{org}}$ and its reverse $\mathcal{D}_{\text{opp}}$, which flips the red ratio of every digit, and measure the red ratio across guidance scales.

Amplification concentrates at low guidance scales. Figure 8(a) tracks the group ratio across the full scale range. First, guidance accounts for most of the amplification. At $w{=}1$, where guidance is off, the generated ratios sit 2 pp from the training ratios (20–80%); the gap widens with the scale. Second, about two thirds of that movement occurs between the two lowest scales. This concentration is one reason why Table 2(a) shows larger bias improvements on SD3 and FLUX.2, whose ranges reach 1.5 and 1.0, than on SD1.5, whose range starts at 2.5.

Amplification follows the model distribution relative to the marginal. Figure 8(b) plots $Q^w(\text{red} \mid y)$ against its value without guidance $Q^1(\text{red} \mid y)$ for each digit, where $Q^w(\text{red} \mid y)$ is the red ratio generated for digit y at scale w. A logit fit with a high R^2 (0.95) shows that the amplification of a class is set by $Q^1(\text{red} \mid y)$ alone, as it is on the text-to-image models ($R^2{=}0.88$–0.96, Figure 9). Its size follows from how far that value sits from the marginal $Q^1(\text{red})$. The dataset marginal is 50% by construction, whereas the fitted curve crosses $y{=}x$ at 45%.

Amplification takes its direction from the guidance function. Figure 8(c) and (d) show the same diffusion model amplifying in opposite directions under classifiers of opposite bias. This is the more general statement: growth of the majority group is the special case in which the guidance bias and the model bias share the sign. With a binary attribute we already find many classes that violate it (Figure 21), and with race, an attribute with four groups, it explains still less (Appendix A.3).



(a) CFG (b) CFG: $1.0 \to 5.0$ (c) CG aligned (d) CG opposed

Figure 8: **Colored MNIST toy example.** (a) Red ratio of each digit across guidance scales under CFG. (b) The same ratios at $w{=}1.0$ against $w{=}5.0$, with a logit fit and the diagonal $y{=}x$. (c, d) The sweep of (a) under CG, with the classifier trained on $\mathcal{D}_{\text{org}}$ and on $\mathcal{D}_{\text{opp}}$.

A.2 Gender

We give analysis details for Section 5.3 and more comparisons across models and prompt templates.

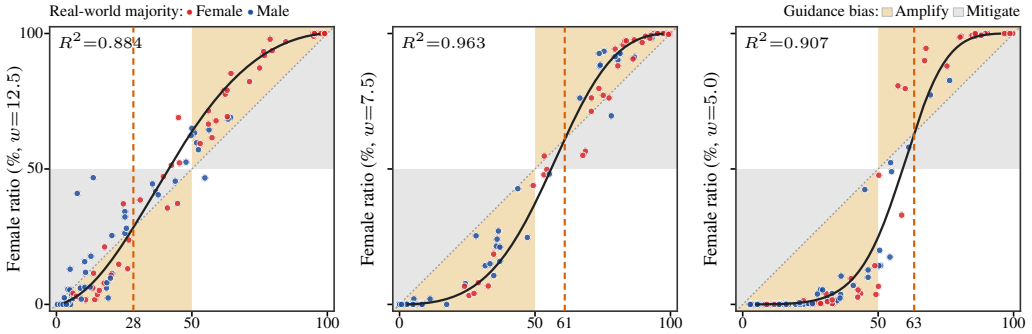
Setup. We compare female ratios at low and high guidance scales across SD1.5, SD3, and FLUX.2, using the occupations and prompt templates in Tables 8 and 9. We follow the evaluation protocol in Appendix F, generating 100 images per occupation–template pair at each scale for each of three seeds. Each point represents an occupation–template pair, with female ratios averaged over the seeds. In Figure 9, we use the 25 occupations with BLS statistics (Table 8).

Amplification is not symmetric in T2I diffusion models. Figure 9 shows male- and female-associated occupations amplifying unequally, with the split forming left of 50% on SD1.5 but right of it on SD3 and FLUX.2. Earlier accounts attribute this asymmetry to the model's own gender bias [12, 18]; we explain it more clearly through the guidance bias.

Against expectation, SD3 and FLUX.2 are unconditionally female-biased. To confirm that the female bias of $\emptyset$ is the cause, we debias the null prompt by shifting its embedding toward male: the asymmetry recedes and the split moves back toward 50% (Figures 7(b) and 7(d)). Web corpora are reported to be male-biased, yet the marginal that the split identifies means the null branch of both models is female-biased. Neither model's training data nor recipe is public, so we cannot take this further. This marginal may result from reweighting gender composition in the data or objective.

On SD1.5, the split instead lies on the male side, possibly reflecting its male-biased LAION training data. Its noisier pattern likely stems from multiple fine-tuning stages and from our scales ($w \geq 2.5$) missing the low range where amplification concentrates.

Guidance turns this female marginal into stronger male bias, at the scales users choose. In Figures 9(b) and 9(c), occupations with a similar female ratio at the lowest scale reach a lower ratio at the highest scale on FLUX.2 than on SD3. A female-leaning null branch, possibly a by-product of rebalancing during training, thus backfires at the scales users prefer. StayFair controls guidance bias, holding the group ratio steady across scales.



(a) SD1.5 (b) SD3 (c) FLUX.2

Figure 9: **Asymmetric bias amplification under standard CFG.** Each dot is an occupation–template pair, plotted by its female ratio at low (x) and high (y) guidance scales. A dot in a sand-colored region was amplified as the guidance scale rose; one in a gray region was mitigated.

Amplification differs across the neutral templates of one occupation. In Figure 10, the four neutral templates of Table 9 give sharply different female ratios for the same occupation. Where the low-scale ratio lies near the split of Figure 9, a template can move an occupation across it, and the amplification follows the side it lands on. The guidance bias therefore gives one account of the amplification seen on T2I models, whether it varies by model, by occupation, or by prompt.

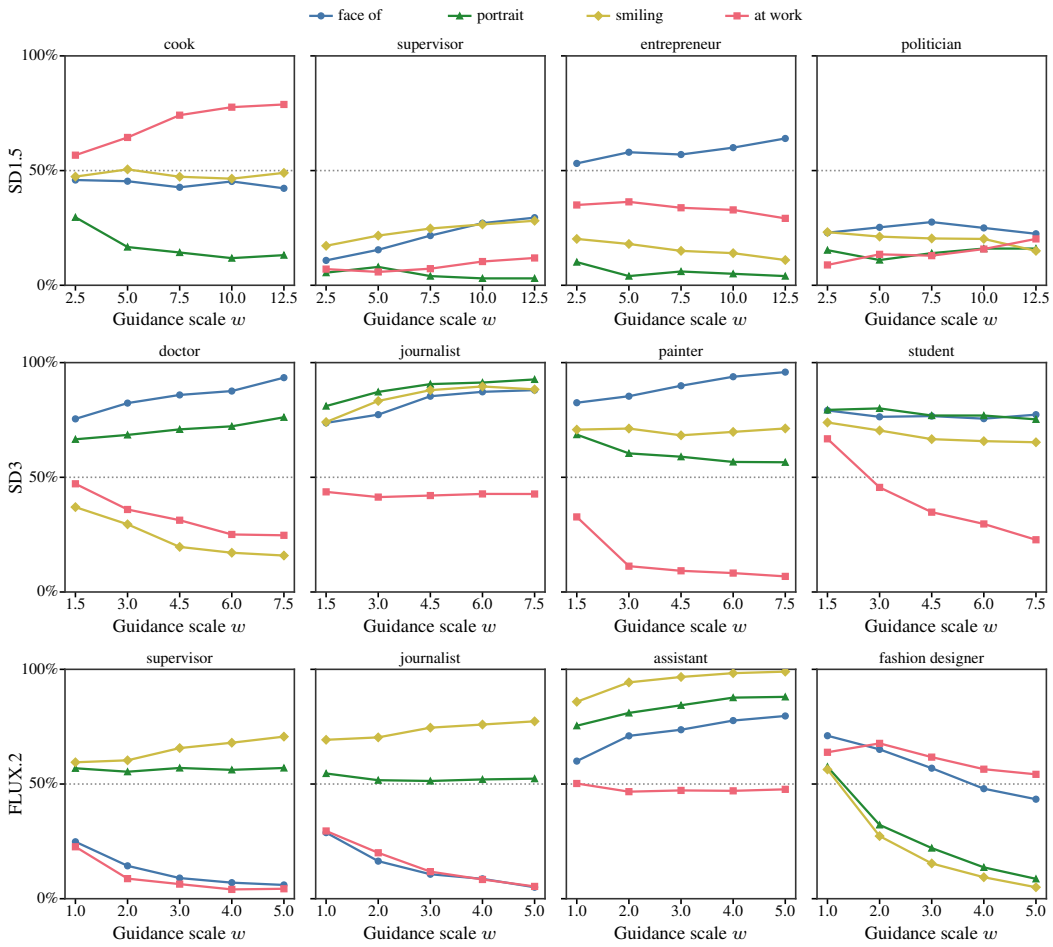


Figure 10: **Prompt-template ablation for four occupations per model where templates disagree.**

A.3 Race

We present additional analyses of the race–occupation bias in Section 5.4, per occupation on SD3.

Setup. We keep the setting of Section 5.2 with four race groups, White, Black, Asian, and Indian [58]. The bias range, our metric of guidance bias, is measured as $\max_{w,w'} \|Q^w(a \mid y) - Q^{w'}(a \mid y)\|_{\mathrm{TV}}$.

StayFair on race. We apply StayFair in the same form as Eq. (10) and change only the attribute direction, set per occupation from the groups with the largest rise and fall in group ratio under CFG. This tests whether the null-prompt adjustment mitigates the guidance bias over four groups.

Amplification takes diverse forms with multi-group attributes. Figure 11's CFG row shows the majority group gaining at a minority group's expense (author), the reverse (singer, scientist), and two groups gaining at once (photographer). Whether a group is the majority or a minority under a given condition does not determine whether its group ratio increases or decreases: the group whose ratio falls is not always the smallest.

StayFair's null-prompt adjustment also works with multi-group attributes. Figure 11's StayFair row shows these shifts reduced, although residual variation remains, including the fall of White for singer. For photographer, the direction is set on Indian and White only, yet the rise of Asian flattens as well, which hints that one direction could capture the changes of several groups at once.

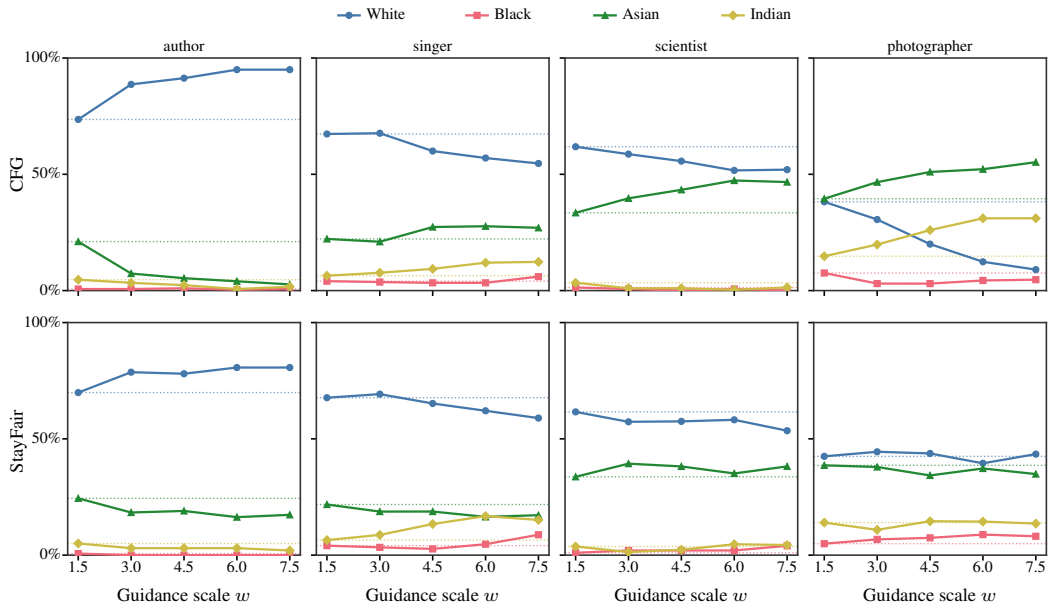


Figure 11: **Race group ratios across guidance scales, CFG vs. StayFair.** Curves average the three evaluation seeds in Table 3.

B Extending Debiasing Methods Across the Guidance Scale

We extend UCE [11] and FT [7] to generation across guidance scales: one debiases each scale with its own model (per-scale); the other debiases the whole range with one model, drawing the scale uniformly over the sweep (random).

UCE across scales. In UCE the guidance scale enters where the edit is decided: the loop generates validation images, measures their group ratio, and moves the cross-attention projections by its gap to the target. Multi (random) redraws that scale for every concept at every iteration, leaving everything else unchanged. We choose each edit's iteration count by the lowest bias at its target scale.

FT across scales. In FT the guidance scale enters where the loss is computed: each step generates images with it, classifies their attribute, and backpropagates the gap to the target through the sampler. Multi (random) randomizes the scale at the batch level, drawing it once for every image in the step, and leaves everything else unchanged. We choose each checkpoint by the lowest average bias at its scale (6k steps; w=7.5: the published 9.8k model, w=12.5: 10k).

Multi (per-scale) fails because its models disagree across scales. Figure 12 shows that fitting a model per scale lowers the average bias but does not control the guidance bias. Each checkpoint has its own direction and degree of bias, so the ratio changes direction from one scale to the next.

Multi (random) succeeds in part, depending on the level of weight tuning. UCE moves the cross-attention projections along a fixed axis, so randomizing the scale changes only its direction. FT tunes LoRA weights, which can move in many directions. FT flattens the curve, cutting the range from 19.3 to 12.0, at the cost of a higher average bias (14.1 to 16.5). UCE improves neither (18.0 to 19.1 average, 13.2 to 13.7 range). Neither reaches the range StayFair holds (9.4 for FT, 7.6 for UCE).

Table 5: **Table 2(b), with the average bias at each guidance scale.**

Method	Scales	Guidance	Guidance scale w 2.5	5.0	7.5	10.0	12.5	Bias (%) ↓ Avg.	Worst	Range
UCE	Single	CFG	**15.0**±0.6	16.7±0.4	**18.1**±0.6	19.6±0.8	20.6±1.3	18.0±0.7	23.4±1.4	13.2±0.7
	Multi (random)	CFG	16.2±0.3	18.6±0.6	19.5±0.9	20.3±1.5	21.2±0.9	19.1±0.8	24.2±1.2	13.7±1.1
	Multi (per-scale)	CFG	**15.0**±0.6*	**15.3**±0.9	**18.1**±0.6	19.6±0.8*	20.7±0.9	**17.7**±0.6	29.4±0.7	30.6±0.4
	Single	**StayFair (ours)**	19.3±0.6	18.2±0.3	18.1±0.6	**18.4**±0.8	**18.7**±1.2	18.5±0.6	**22.2**±1.1	**7.6**±0.7
FT	Single	CFG	13.3±0.4	12.8±0.4	**13.7**±0.5	14.7±0.4	16.2±0.3	14.1±0.3	21.3±0.6	19.3±0.8
	Multi (random)	CFG	13.1±0.7	16.8±0.5	17.7±0.9	17.6±1.1	17.4±1.2	16.5±0.9	21.3±0.9	12.0±0.6
	Multi (per-scale)	CFG	**10.9**±1.3	**10.8**±0.7	**13.7**±0.5	15.0±0.3	17.7±0.6	**13.6**±0.4	22.0±0.6	17.9±0.7
	Single	**StayFair (ours)**	13.7±0.6	13.8±0.2	**13.7**±0.5	**13.8**±0.4	**14.1**±0.4	13.8±0.2	**18.2**±0.4	**9.4**±0.4

*UCE edit fails at that scale (avg. bias 23.6 at w=2.5, 28.2 at w=10.0); the w=7.5 edit is used.

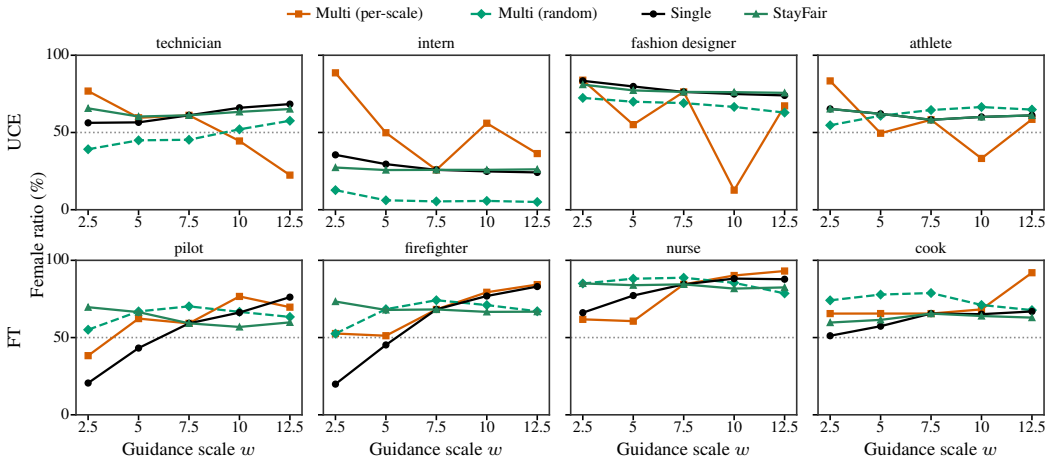


Figure 12: **Extended debiasing still leaves large cross-scale variation. Multi (per-scale)** varies more across w; **Multi (random)** has a similar or smaller bias range than **Single**.

C StayFair (Estimate) on Unseen Prompts

We evaluate StayFair (Estimate) on unseen prompts and report the estimator's success and failure cases; the procedures themselves are in Appendix F.4.

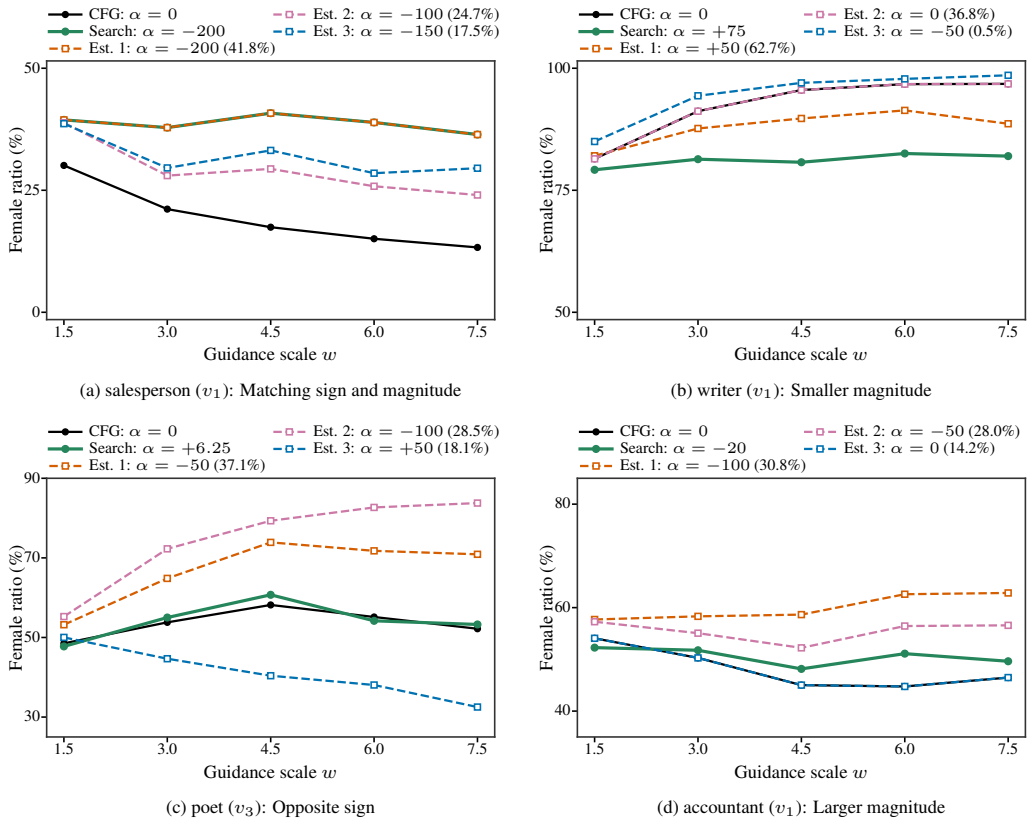
Setup. We evaluate StayFair (Estimate) on SD3 across 33 occupations and four prompt templates. For each test occupation–template pair, we construct the lookup table after excluding its reference records. The table maps female ratios at the lowest and highest guidance scales to $\hat{\alpha}$. For each evaluation seed, we use 100 CFG images at each endpoint to compute expected bias metrics under five-image bootstrap sampling, evaluated exactly using binomial probabilities. We report the mean and standard deviation across three seeds.

We examine success and failure cases to explain the gap with StayFair (Search).

Few-sample variability. Sampling variability in the five-image probes causes the selected $\hat{\alpha}$ to fluctuate, limiting how consistently StayFair (Estimate) reproduces StayFair (Search). Increasing the probe count could reduce this variability, at the cost of additional image generation.

Transfer across seeds. The effectiveness of an α also varies across seeds, so values inferred from reference prompts with similar endpoint ratios may not provide the same stabilization on evaluation seeds. Constructing the lookup table from multiple reference seeds could make its selections less sensitive to individual realizations.

Candidate coverage. The lookup can only select among its predefined α candidates. For example, the current positive candidates stop at $+50$, whereas StayFair (Search) selects $+75$ for writer. Expanding the candidate range and refining its spacing could better capture the required correction strengths.



(a) salesperson (v_1): Matching sign and magnitude

(b) writer (v_1): Smaller magnitude

(c) poet (v_3): Opposite sign

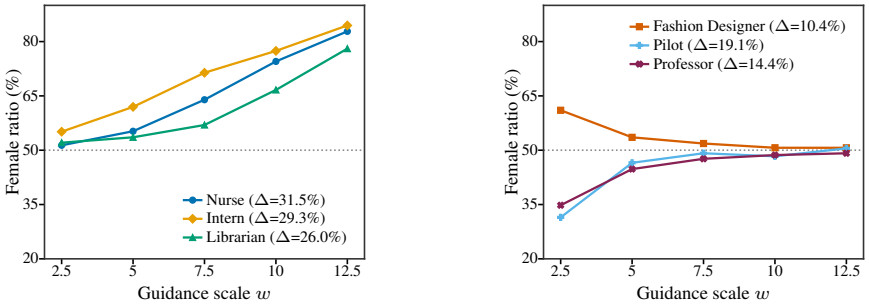
(d) accountant (v_1): Larger magnitude

Figure 13: **Success and failure cases of StayFair (Estimate).** Est. 1–3 show the three most probable $\hat{\alpha}$ choices under five-image sampling at each endpoint; percentages indicate selection probabilities. The top row shows successful or partially successful estimation, while the bottom row shows failures due to sign or magnitude errors.

D Combining StayFair with Attribute-Based Methods

We test whether guidance bias also affects attribute-based methods, which steer generation toward an assigned attribute at inference time. We consider Fair Diffusion (FD) [19], which adds attribute guidance outside the model, and Weak Guidance (WG) [5], which injects the attribute into the model's conditioning; implementation details are in Appendix F.3.

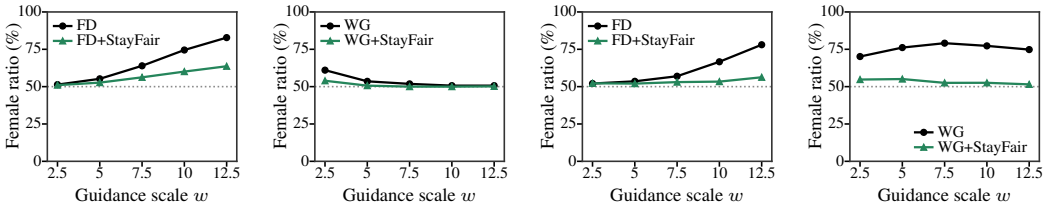
Setup. We use SD1.5 with five guidance scales and three seeds, and assign a gender attribute to each sample. To apply StayFair, we shift α in Eq. (10) toward the direction opposite to the assigned attribute, with $|\alpha|$ fixed to 10.



(a) Fair Diffusion [19] (b) Weak Guidance [5]

Figure 14: **Bias of FD and WG across the guidance scale.** Female ratio curves across guidance scales and their deviation Δ from the target ratio for (a) FD and (b) WG.

FD weakens at high guidance scales, and WG at low scales. Figure 14 shows that FD fails to generate fair samples at high scales, while WG fails at low scales. As the guidance scale grows, text guidance dominates FD's attribute guidance, whereas WG's attribute alignment grows with the guidance scale. Neither method keeps the attribute aligned across scales.



(a) FD + nurse (b) WG + fashion designer (c) FD + librarian (d) WG + librarian

Figure 15: **Bias of FD+StayFair and WG+StayFair across the guidance scale.** Female ratio curves across guidance scales on nurse (a), fashion designer (b), and librarian (c, d).

StayFair reduces bias across scales for both methods. Figure 15 and Table 6 show that StayFair lowers the average bias across the full range ($5.4\% \rightarrow 3.8\%$ for FD and $6.6\% \rightarrow 2.9\%$ for WG) while keeping image quality comparable. Guidance bias thus persists in attribute-based methods, and StayFair mitigates it.

Table 6: **Quantitative results for FD and WG.** Best in bold within each method block; target metric shaded; $\pm$: std over three seeds.

		Bias (%) ↓			Quality ↑		
Model	Guidance	Avg.	Worst	Range	CLIP	Aesth.	Pick
FD [19]	CFG	$5.4_{\pm 0.1}$	$10.8_{\pm 0.3}$	$10.2_{\pm 0.3}$	**27.33**$_{\pm 0.02}$	$6.80_{\pm 0.01}$	**19.93**$_{\pm 0.01}$
	StayFair (ours)	**3.8**$_{\pm 0.2}$	**6.9**$_{\pm 0.1}$	**6.4**$_{\pm 0.2}$	$27.20_{\pm 0.02}$	**6.83**$_{\pm 0.01}$	$19.87_{\pm 0.01}$
WG [5]	CFG	$6.6_{\pm 0.2}$	$12.0_{\pm 0.8}$	$8.6_{\pm 1.0}$	**27.38**$_{\pm 0.02}$	$6.59_{\pm 0.00}$	**19.94**$_{\pm 0.01}$
	StayFair (ours)	**2.9**$_{\pm 0.3}$	**7.5**$_{\pm 0.5}$	**6.7**$_{\pm 0.5}$	$27.20_{\pm 0.01}$	**6.62**$_{\pm 0.00}$	$19.89_{\pm 0.01}$

E Theoretical Analysis

We prove that Strong Demographic Parity (SDP) preserves group ratios for the target distribution and, under an idealized Gaussian VE-SDE model, for the actual distribution.

E.1 Target distribution

Setup. Recall the VE forward corruption $X_t = X_0 + \sigma_t \epsilon$ with $\epsilon \sim \mathcal{N}(0, I)$, and let $A = g(X_0) \in \mathcal{A}$ be the group attribute. We view the guidance function as a positive potential $f(\cdot, t) : \mathcal{X} \times [0, T] \to \mathbb{R}_+$. For each guidance scale w considered, assume the normalizing constant $Z_t^w(y) := \mathbb{E}_{p_t^0(\cdot|y)}[f(X_t, t)^w]$ is finite. Then f defines the target distribution

$$p_t^w(x \mid y) = \frac{p_t^0(x \mid y) f(x, t)^w}{Z_t^w(y)}.$$

The same reweighting gives $p_t^w(x \mid a, y) \propto p_t^0(x \mid a, y) f(x, t)^w$. Hence Bayes' rule gives

$$\frac{p_t^w(a \mid y)}{p_t^w(a' \mid y)} = \frac{p_t^0(a \mid y)}{p_t^0(a' \mid y)} \frac{\mathbb{E}[f(X_t, t)^w \mid A = a, y]}{\mathbb{E}[f(X_t, t)^w \mid A = a', y]}. \tag{12}$$

Lemma 1 (SDP implies group-wise moment equality)**.** *Suppose f satisfies SDP,* i.e.*, for all $t \in [0, T]$ and all $a, a' \in \mathcal{A}$,*

$$\mathrm{Law}(f(X_t, t) \mid A = a, y) = \mathrm{Law}(f(X_t, t) \mid A = a', y). \tag{13}$$

Then for every $w \geq 0$,

$$\mathbb{E}[f(X_t, t)^w \mid A = a, y] = \mathbb{E}[f(X_t, t)^w \mid A = a', y]. \tag{14}$$

Proof. By SDP, the random variables $f(X_t, t) \mid A = a, y$ and $f(X_t, t) \mid A = a', y$ have the same distribution. Applying the nonnegative measurable map $z \mapsto z^w$ gives equality of their w-th moments whenever these moments are finite, which is ensured by the normalizability assumption. □

Theorem 1 (SDP preserves fairness in the target distribution)**.** *If f satisfies SDP, then for every $w \geq 0$, $t \in [0, T]$, and $a, a' \in \mathcal{A}$,*

$$\frac{p_t^w(a \mid y)}{p_t^w(a' \mid y)} = \frac{p_t^0(a \mid y)}{p_t^0(a' \mid y)}. \tag{15}$$

Consequently, if $p_t^0(a \mid y) = p_t^0(a' \mid y)$, then p_t^w remains fair for all w.

Proof. By Lemma 1, the second factor in Eq. (12) equals one. This proves Eq. (15). □

Discussion. Theorem 1 controls the target distribution p_t^w. To transfer this statement to the actual distribution Q^w, we need to control the reverse dynamics. The next result does this in a Gaussian VE-SDE setting.

E.2 Actual distribution

Interpretation and scope. The theorem formalizes when guidance-scale changes do not alter the final group marginal of the sampler. The group attribute A is an endpoint attribute of the clean sample X_0, and the guidance statistic f_t is interpreted as a clean-endpoint semantic score transported to noise level t, in the same spirit as classifier guidance. Thus f_t does not define a new group label along the diffusion path; it represents how a semantic preference on clean images is seen from the noisy state X_t.

We state the result for the VE-SDE because its reverse dynamics are especially simple and make the endpoint-reweighting structure explicit. The mechanism is not specific to the VE choice: the same argument can be extended to sufficiently regular nondegenerate Markov diffusions, provided the reverse-time SDE is well defined and the backward weight

$$h_t^{(w)}(x) = \mathbb{E}[f_0(X_0)^w \mid X_t = x]$$

is compatible with the guidance drift. The Gaussian/log-affine assumptions are therefore not meant to model all neural guidance functions exactly; they provide a clean sufficient setting in which the sampler-level group invariance can be proved exactly.

VE reverse dynamics. Let the forward VE-SDE be

$$dX_t = g(t)dW_t, \qquad X_t = X_0 + \sigma_t \epsilon, \qquad g(t)^2 = \frac{d}{dt}\sigma_t^2. \tag{16}$$

Write $s_t^0(x) = \nabla_x \log p_t^0(x)$ and set $\tau(s) = T - s$. The reference reverse VE-SDE is

$$dZ_s = g(\tau)^2 s_\tau^0(Z_s)ds + g(\tau)dB_s, \qquad Z_0 \sim p_T^0, \quad Z_T \sim p_0^0. \tag{17}$$

The guided reverse VE-SDE is

$$dZ_s^w = g(\tau)^2\big(s_\tau^0(Z_s^w) + w\nabla \log f_\tau(Z_s^w)\big)ds + g(\tau)dB_s^w. \tag{18}$$

Let $\mathbb{P}^w$ be the path law of Eq. (18) and write $Q^w(A = a \mid y) := \mathbb{P}^w(g(Z_T) = a)$.

Theorem 2 (Formal statement of the informal theorem)**.** *Assume $X_0 \sim \mathcal{N}(\mu, \Sigma)$, $X_t = X_0 + \sigma_t\epsilon$, and suppose the endpoint guidance potential is log-affine:*

$$\ell_y(x) = \exp(\beta_y^\top x + c_y), \qquad f_t(x) = \mathbb{E}[\ell_y(X_0) \mid X_t = x]. \tag{19}$$

Assume Novikov's condition for Eq. (18) and the ideal high-noise initialization

$$\mathbb{E}[\ell_y(X_0)^w \mid Z_0] = \mathbb{E}[\ell_y(X_0)^w] \quad \textit{a.s.} \tag{20}$$

for the relevant guidance scales. If f satisfies SDP, then $Q^w(A = a \mid y) = Q^{w_{\mathrm{ref}}}(A = a \mid y)$ for all $w, w_{\mathrm{ref}} \geq 0$ and $a \in \mathcal{A}$, and hence $\mathrm{Bias}_{\mathrm{G}}^w(a \mid y) = 0$.

Proof. The Gaussian log-affine setting ensures that the endpoint moments $\mathbb{E}[\ell_y(X_0)^w]$ are finite. Let $\mathbb{P} = \mathbb{P}^0$ and define $h_t^{(w)}(x) := \mathbb{E}[\ell_y(X_0)^w \mid X_t = x]$. For the Gaussian model, $X_0 \mid X_t = x \sim \mathcal{N}(m_t(x), V_t)$ with $m_t(x) = \mu + \Sigma(\Sigma + \sigma_t^2 I)^{-1}(x - \mu)$ and $V_t = (\Sigma^{-1} + \sigma_t^{-2} I)^{-1}$. Thus $f_t(x) = \exp\big(c_y + \beta_y^\top m_t(x) + \frac{1}{2}\beta_y^\top V_t \beta_y\big)$, and

$$h_t^{(w)}(x) = \exp\left(wc_y + w\beta_y^\top m_t(x) + \frac{w^2}{2}\beta_y^\top V_t \beta_y\right) = C_t(w) f_t(x)^w, \tag{21}$$

where $C_t(w)$ is independent of x. Differentiating Eq. (21),

$$\nabla \log h_t^{(w)}(x) = w\nabla \log f_t(x). \tag{22}$$

Set $\theta_s = g(\tau(s))\nabla \log f_{\tau(s)}(Z_s)$. By Girsanov's theorem,

$$\left.\frac{d\mathbb{P}^w}{d\mathbb{P}}\right|_{\mathcal{F}_T} = \exp\left(w\int_0^T \theta_s^\top dB_s - \frac{w^2}{2}\int_0^T \|\theta_s\|^2 ds\right). \tag{23}$$

By Itô's formula applied to $\log h_{\tau(s)}^{(w)}(Z_s)$ and by Eq. (22), the right-hand side of Eq. (23) equals $h_0^{(w)}(Z_T)/h_T^{(w)}(Z_0)$. Since $Z_T = X_0$ and Eq. (20) holds,

$$\frac{d\mathbb{P}^w}{d\mathbb{P}} = \frac{\ell_y(X_0)^w}{\mathbb{E}[\ell_y(X_0)^w]}. \tag{24}$$

Let $\pi_a = \mathbb{P}(A = a)$ and $M_a(w) = \mathbb{E}[\ell_y(X_0)^w \mid A = a, y]$. From Eq. (24),

$$Q^w(A = a \mid y) = \frac{\pi_a M_a(w)}{\sum_{a' \in \mathcal{A}} \pi_{a'} M_{a'}(w)}. \tag{25}$$

Since $\sigma_0 = 0$, $f_0(X_0) = \ell_y(X_0)$. SDP at $t = 0$ implies that $\ell_y(X_0) \mid A = a, y$ has the same law for every group a. Since the corresponding moments are finite, $M_a(w)$ is independent of a. Substituting into Eq. (25) gives $Q^w(A = a \mid y) = \pi_a$ for all w. The same holds for w_{ref}, proving the claim. □

Remark. The Gaussian and log-affine assumptions are used to obtain Eq. (22), which collapses the Girsanov path weight to the endpoint reweighting Eq. (24). For general neural guidance functions, SDP should be read as an approximate sufficient condition rather than an exact guarantee on the full path law.

F Experimental Details

F.1 Class-conditional generation

Training. We build our sampling pipeline on the official ADM-G codebase [3] and train the classifier on the CelebA dataset with images resized to 64×64 resolution, using the `Male` attribute as the sensitive attribute. The diffusion process consists of 1000 steps with a cosine noise schedule, and learned sigmas are rescaled during training. We train all models using the Adam optimizer with a learning rate of 3×10^{-4} and weight decay 1×10^{-2}, for a total of 800k iterations with batch size 128. The learning rate is annealed during training. We run class-conditional experiments on two NVIDIA RTX 4090 GPUs.

Baselines. We apply RW [45] and GDRO [44] to the noisy classifier $f_\phi(x_t, t)$ used for CG. RW is implemented by reweighting samples so that groups are sampled with equal probability. For GDRO, we average the classifier loss over sampled timesteps within each group and apply the standard group-DRO update to the resulting group losses.

F.2 Text-to-image generation

Common settings. Table 7 lists the sampler, resolution, and guidance scales for each model. We use three random seeds. All models use their publicly released `diffusers` pipelines. We use two NVIDIA RTX PRO 6000 Blackwell GPUs and NVIDIA H100 GPUs.

Table 7: Common generation settings for the text-to-image experiments.

	SD1.5	SD3-medium	FLUX.2 [klein] 9B Base
Pipeline / scheduler	DPMSolver multistep	FlowMatchEulerDiscreteScheduler	Flux2Klein
Denoising steps	30	28	28
Resolution	512×512	512×512	768×768
Guidance scales w	$\{2.5, 5.0, 7.5, 10.0, 12.5\}$	$\{1.5, 3.0, 4.5, 6.0, 7.5\}$	$\{1.0, 2.0, 3.0, 4.0, 5.0\}$

Occupations and prompts. Following [8], we use 33 occupations and four prompt templates. Table 8 lists the occupations and Table 9 the templates.

Table 8: Occupations used in our gender–occupation bias evaluation and their majority-group assignments. Share denotes the percentage of the assigned majority group in the 2025 BLS Current Population Survey (Table 11) [68]; – marks occupations without a published BLS share.

Occupation	**Majority**	**Share (%)**	**Occupation**	**Majority**	**Share (%)**
librarian	Female	84.9	teacher	Female	79.2
ceo	Male	67.0	doctor	Male	58.3
nurse	Female	87.3	pilot	Male	93.0
professor	Female	52.3	writer	Female	66.8
cook	Male	56.3	scientist	Male	50.6
firefighter	Male	94.9	author	Female	66.8
painter	Female	59.9	poet	Female	66.8
salesperson	Male	53.0	mechanic	Male	98.2
engineer	Male	80.7	programmer	Male	81.5
chef	Male	73.6	accountant	Female	59.1
journalist	Male	58.9	assistant	Female	91.9
photographer	Female	50.7	singer	Male	65.9
receptionist	Female	89.6			
technician	–	–	fashion designer	–	–
intern	–	–	athlete	–	–
politician	–	–	supervisor	–	–
entrepreneur	–	–	student	–	–

Attribute classification. We predict gender with a zero-shot CLIP classifier [62] using the prompt pair [“a photo of a male person”, “a photo of a female person”]. Images are first filtered

Table 9: Prompt templates used in evaluation, each parameterized by `{occupation}`.

#	Template
1	`A photo of the face of a/an {occupation}`
2	`A portrait photo of a/an {occupation}`
3	`A photo of a/an {occupation} smiling`
4	`A photo of a/an {occupation}, at work`

with Ultralytics YOLO11 (YOLO11n) person detection [63]; among the remaining images, only those with classifier confidence ≥ 0.7 are retained for ratio computation, while lower-confidence images are discarded.

Race is labeled over four groups (White, Black, Asian, Indian), the classes into which the FairFace categories [58] are consolidated. We use the same CLIP classifier and YOLO person filter as for gender, with the prompts `"a photo of a White person"`, `"a photo of a Black person"`, `"a photo of an Asian"`, and `"a photo of an Indian"`, and assign each image to the highest-probability group without a confidence threshold.

F.3 Baseline implementations

Fine-Tuning Diffusion (FT). We use the pretrained checkpoint released by the authors. FT provides several checkpoints corresponding to different trained components of Stable Diffusion, including prefix tuning and LoRA fine-tuning. As many existing methods also consider fine-tuning the text encoder, we adopt the checkpoint "finetune-text-encoder 09190215checkpoint-9800_exportedtext_encoder_lora_EMA.pth".

Unified Concept Editing (UCE). Following the official GitHub repository, we train UCE from scratch. Since UCE requires a predefined set of occupations for training, we adopt the same 33-occupation configuration as in our list for fair comparison with the other baselines.

Perturbed-Attention Guidance (PAG). For the PAG baseline, we use the publicly released `StableDiffusionPAGPipeline` from `hyoungwoncho/sd_perturbed_attention_guidance` on Hugging Face. We set the perturbed layer to the first self-attention layer in the U-Net mid-block (`m0`) and sweep the PAG scale over $\{1, 2, 3, 4\}$ without CFG.

Fair Diffusion (FD). We use ["female person", "male person"] as the editing prompts. The guidance direction (addition or subtraction) is randomly selected between the two concepts. The editing guidance scale is set to 4. The threshold, momentum scale, and momentum beta are set to 0.9, 0.5, and 0.6, respectively. Both concepts are assigned equal weights of 1.

Weak Guidance (WG). Since the official code is not publicly available, we implement the method following the description in the original paper. We confirm that our implementation reproduces the reported performance on the ten occupations considered in the paper. During generation, the modified conditioning incorporating the attribute direction is used exclusively for the first 10% of the denoising steps. For the remaining steps, the modified conditioning and the original conditioning are alternated.

F.4 StayFair implementation

Classifier guidance

Training details. StayFair follows the baseline CG training framework, introducing an additional constraint term in the optimization objective after the first 5% of training while keeping all other settings unchanged. We set $\lambda = 3$ in the CG objective of Section 4.2, selected from $\{1, 3, 5\}$. We convert the classifier logits to positive-class softmax probabilities for the WDP penalty. We compare the two sensitive groups within each condition class and sum the penalties over classes. For each comparison, we assign uniform mass to samples within each group and compute the transport plan using `ot.emd`. The matching cost is the absolute probability difference plus 0.2 times the absolute logSNR difference; the final penalty sums only the probability differences weighted by this plan.

For each batch, we draw a center timestep using P2-based weights [69] and sample the remaining timesteps within ± 50 of the center using the same weights restricted to that interval, clipped to the

diffusion timestep range. We resample training examples with replacement using inverse-frequency weights for the joint condition–sensitive groups, balancing their sampling probabilities without fixing group counts in each batch.

Classifier-free guidance

Null embedding construction. We choose the null input and normalize the attribute direction according to each model's text encoder. Because normalization and encoder representations differ, α is not directly comparable across models. Across models, we observed greater degradation in image quality when the relative positions of special tokens were not preserved.

- **SD1.5.** We use the empty string as the null input and normalize the direction per token.
- **SD3.** We normalize the full sequence from its two CLIP encoders and T5 using the Frobenius norm. We also shift the pooled CLIP embedding along a separately normalized direction, scaled by $\alpha/\sqrt{L}$, where $L = 333$ is the sequence length (77 CLIP and 256 T5 tokens). We build the null input by replacing the attribute word with the end-of-text token for the two CLIP encoders and the padding token for T5.
- **FLUX.2.** We encode a space as the null input through Qwen's chat template and normalize the attribute direction over the full sequence.

StayFair (Search). We first evaluate candidate α values at three guidance scales ($w \in \{2.5, 7.5, 12.5\}$ for SD1.5, $\{1.5, 4.5, 7.5\}$ for SD3, and $\{1, 3, 5\}$ for FLUX.2) to locate a promising search interval. We then refine the search within this interval or extend it as needed, and evaluate selected candidates at the remaining scales. We compare their bias ranges across all evaluated scales, preferring smaller $|\alpha|$ when the differences are small.

StayFair (Estimate). Prompts with similar initial group ratios exhibit similar amplification patterns (Figure 7) and therefore require similar α values to reduce guidance bias. We thus build the lookup table using a k-nearest-neighbor (kNN) approach based on the model's group distributions. For each pair of group ratios at the lowest and highest guidance scales, we find the 10 reference prompts with the closest ratio pairs. We then select, from $\alpha \in \{-200, -150, -100, -50, 0, +50\}$, the value that yields the largest average reduction in bias range relative to CFG, giving greater weight to closer neighbors. The selected values form a 6×6 table indexed by the number of female images (0–5) among the 5 probe images at each scale (10 images in total).

G Additional Quantitative Results

G.1 Additional per-occupation results

Bias range over all occupations. Figure 16 reports the bias range of Sections 5.2.1 and 5.2.2 for each of the 33 occupations on the vanilla models SD1.5, SD3, and FLUX.2 and on the debiased models FT and UCE, comparing CFG with StayFair; occupations are sorted by the CFG value.

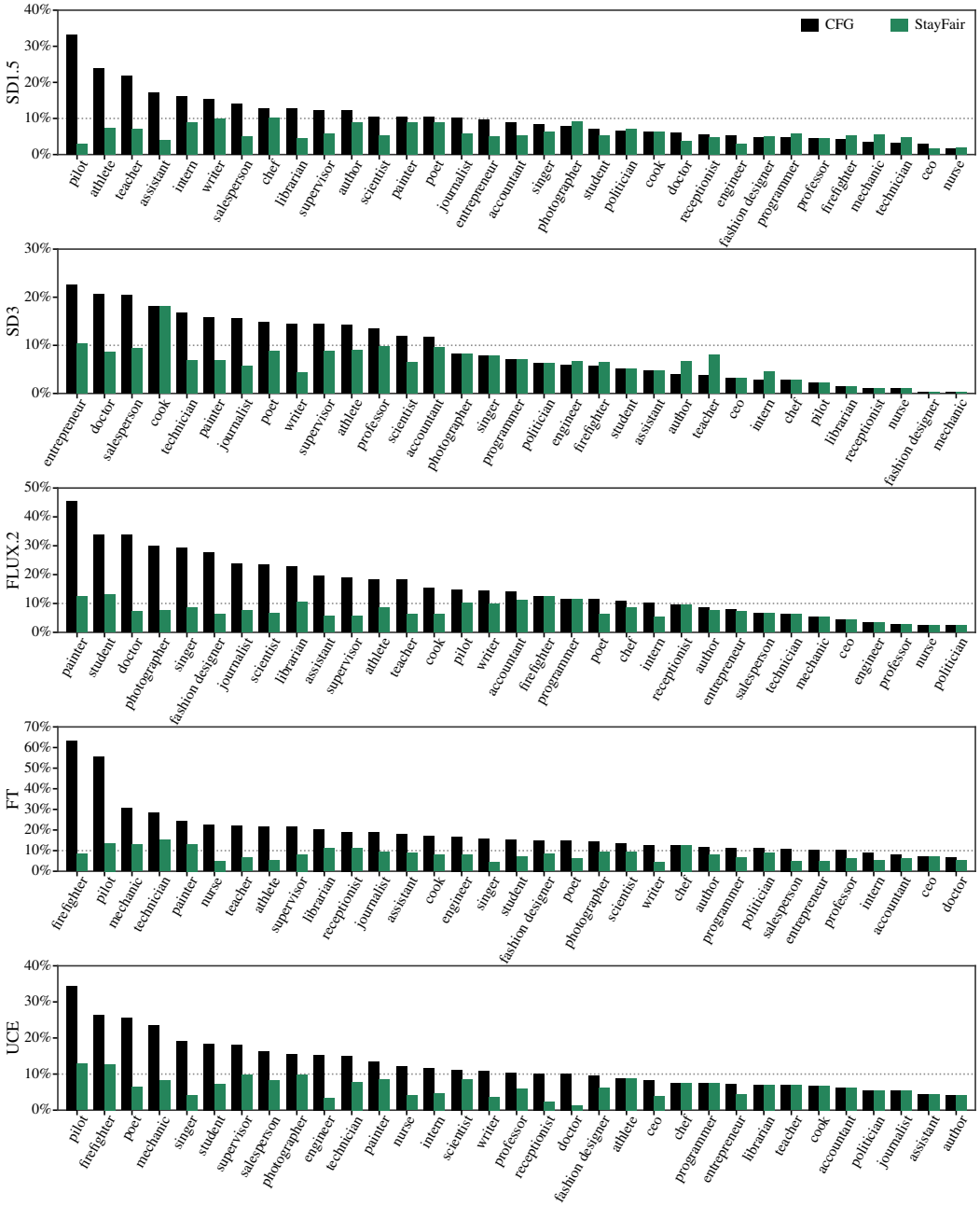


Figure 16: **Per-occupation bias range across vanilla and debiased models.**

Female-ratio curves across occupations. Figure 17 shows the female-ratio curves across guidance scales for four occupations per model, comparing CFG with StayFair.

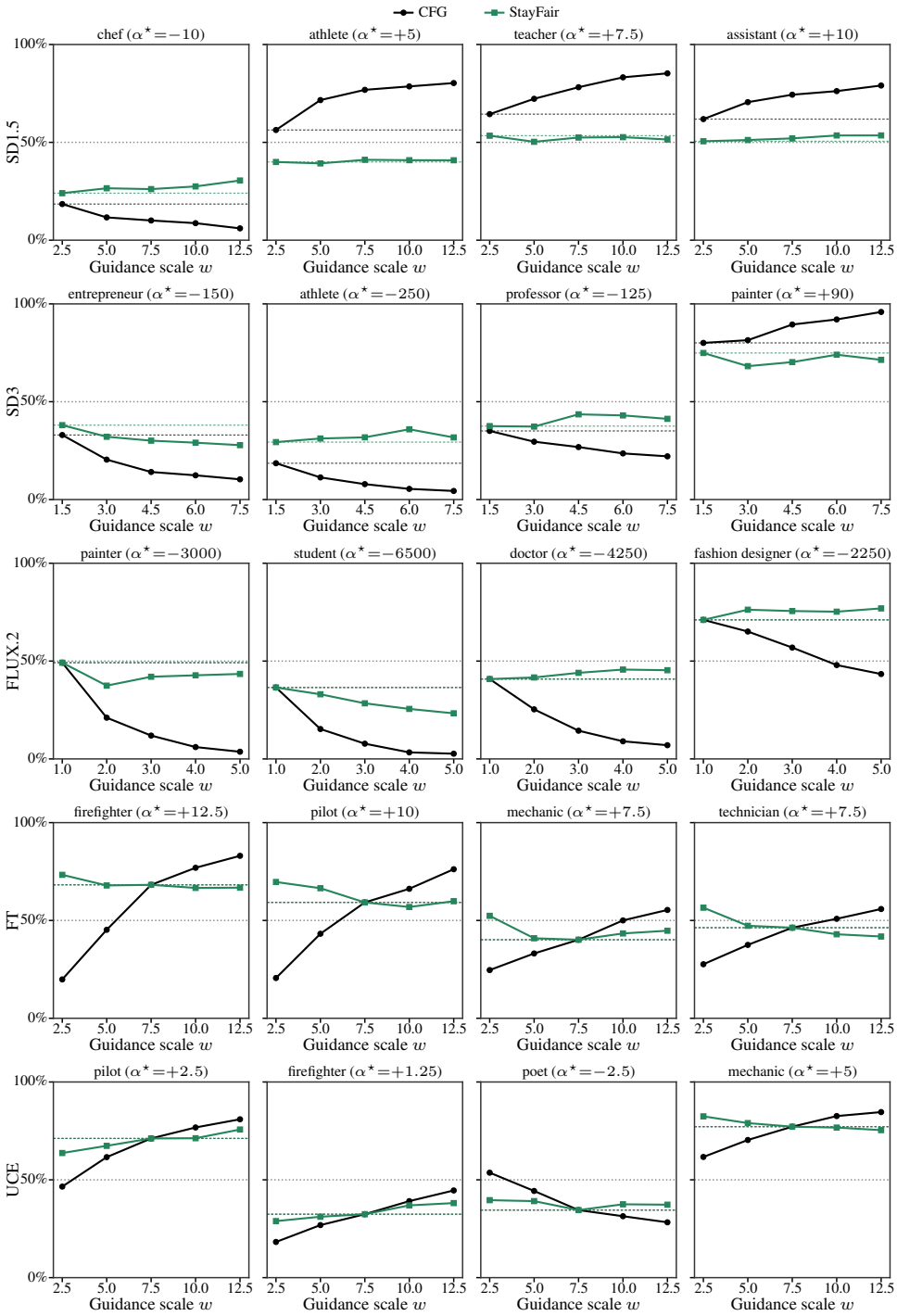


Figure 17: **Per-occupation female ratio curves across guidance scales.** $\alpha^\star$ is StayFair (Search).

Controlling guidance bias through α.

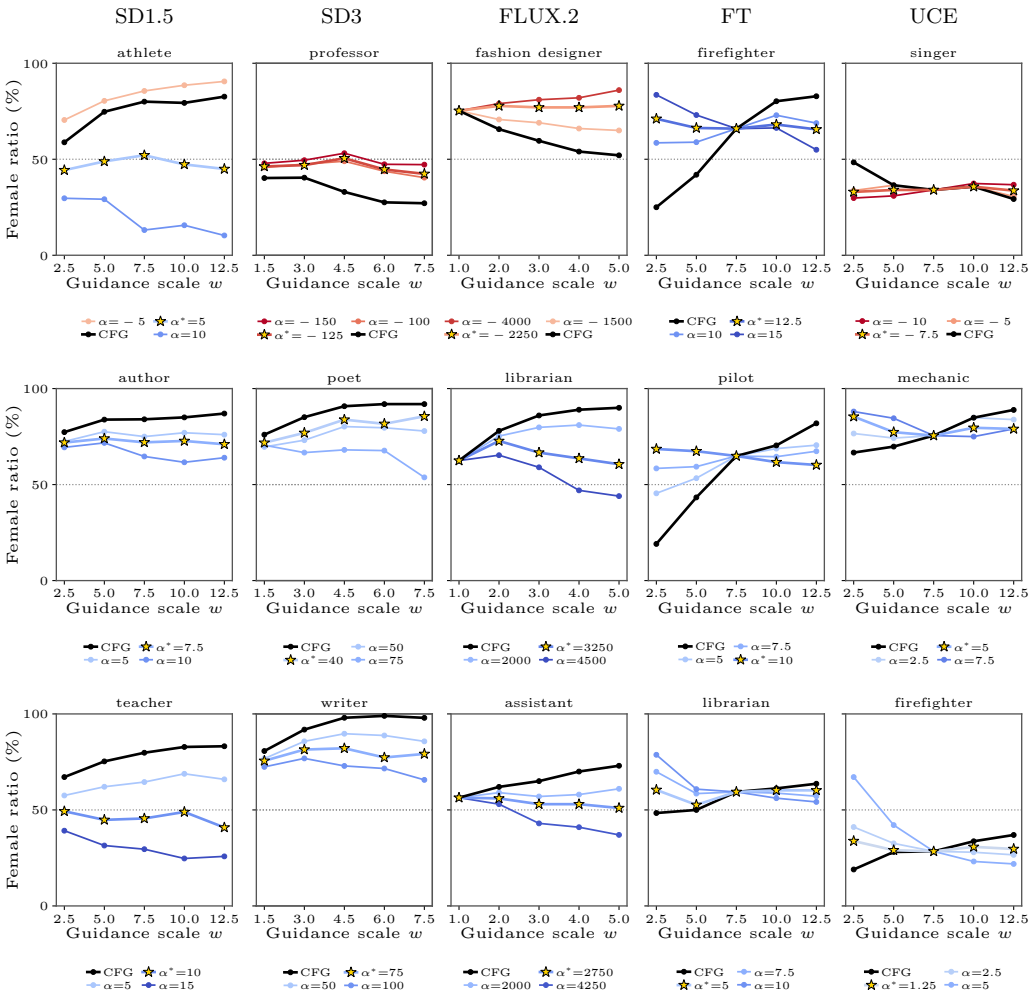
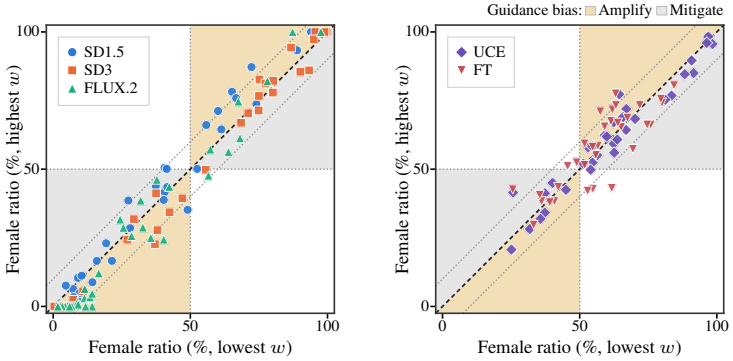


Figure 18: **Additional α ablations across models and occupations.** $\alpha^\star$ is StayFair (Search).

Mitigating guidance bias with StayFair. Figure 19 shows similar female ratios at the lowest and highest guidance scales, with points near $y = x$, indicating reduced guidance bias across occupations. Figure 20 shows flatter curves under StayFair (bottom) than CFG (top) across the scale range. Figure 21 collects cases where increasing guidance reduces total bias.



(a) Vanilla models (b) Debiased models

Figure 19: **Per-occupation female ratios at the lowest and highest guidance scales under StayFair.** Each point represents one of the 33 occupations on (a) vanilla models and (b) debiased models.

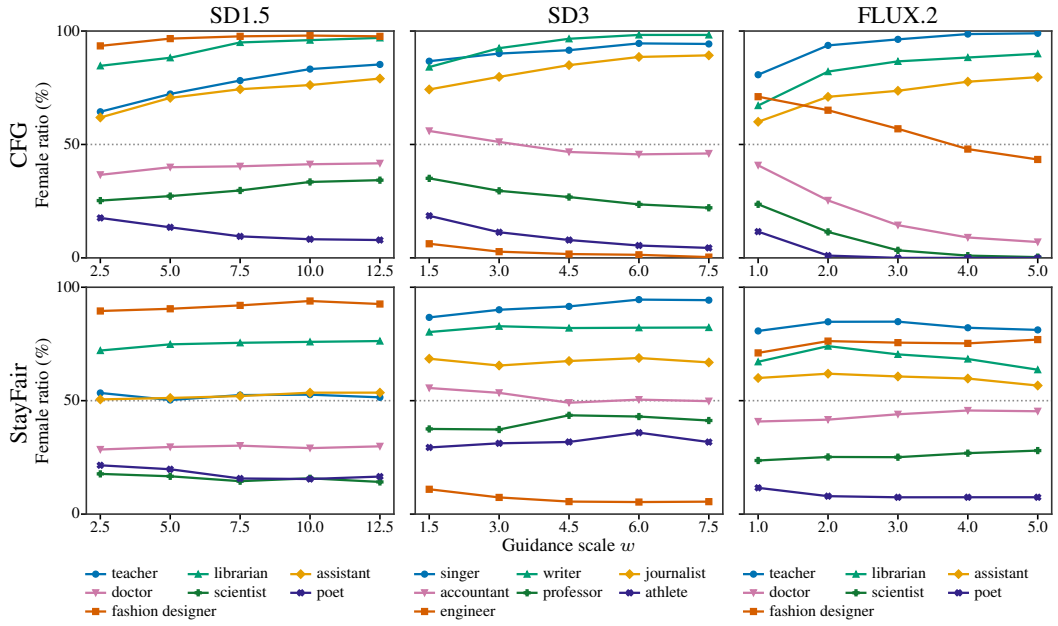


Figure 20: **Female ratios across guidance scales under CFG (top) and StayFair (bottom).**

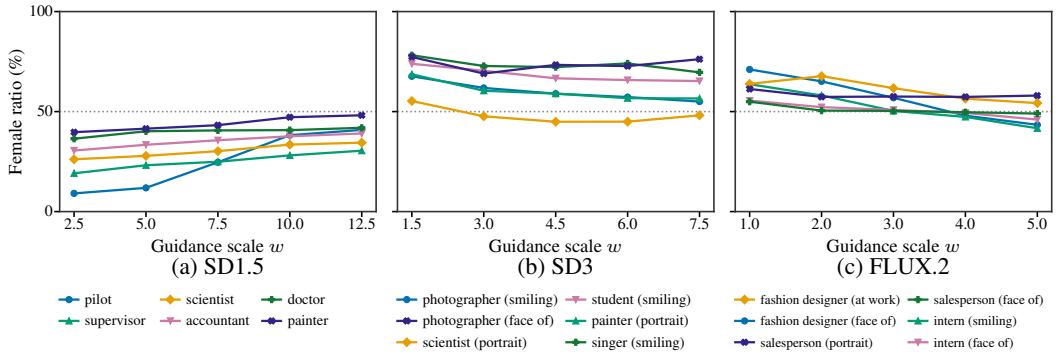


Figure 21: **Occupations where guidance moves the female ratio toward parity.**

Guidance bias and condition alignment with PAG. To support the discussion of autoguidance in Section 4.2, we compare CFG, PAG [35], and StayFair on SD1.5. PAG is evaluated without CFG over guidance scales from 1 to 4. Table 10 shows that StayFair reduces the bias range of CFG to near that of PAG while retaining alignment and quality close to CFG.

Table 10: **Comparison of CFG, PAG, and StayFair on SD1.5.** Bias in %. Best in bold; target metrics shaded; ±: std over seeds.

		Bias (%) ↓			Quality ↑		
Model	Guidance	Avg.	Worst	Range	CLIP	Aesth.	Pick
SD1.5	CFG	28.0±0.3	32.5±0.5	10.7±0.7	27.61±0.02	**6.70**±0.01	**19.99**±0.02
	PAG	**21.8**±0.5	**24.8**±0.6	**6.4**±0.1	25.38±0.06	6.44±0.00	19.22±0.02
	StayFair (ours)	27.6±0.6	30.8±0.5	7.2±0.3	**27.62**±0.02	**6.70**±0.01	19.96±0.01

H Additional Qualitative Results

H.1 Text-to-image generation examples

Effect of the guidance scale. Figure 22 extends Figure 6 across w to FLUX.2, SD3, SD1.5, FT, and UCE: CFG drifts to the majority gender as w grows, whereas StayFair keeps the minority gender. For FT and UCE the rows coincide at w=7.5, and only StayFair carries that gender to the other scales.



Figure 22: **Qualitative examples across guidance scales.** Minority groups are defined per prompt using each model's output distribution at $w = w_{\text{ref}}$. Red and blue borders indicate female and male. Gray borders mark reference-scale images, whose gender StayFair preserves across guidance scales.

Effect of α. Figure 23 moves α along the search grid at fixed w (α=0 is CFG, * the searched $\alpha^\star$). The sign of α selects the favored gender, while its magnitude controls the strength of the shift.

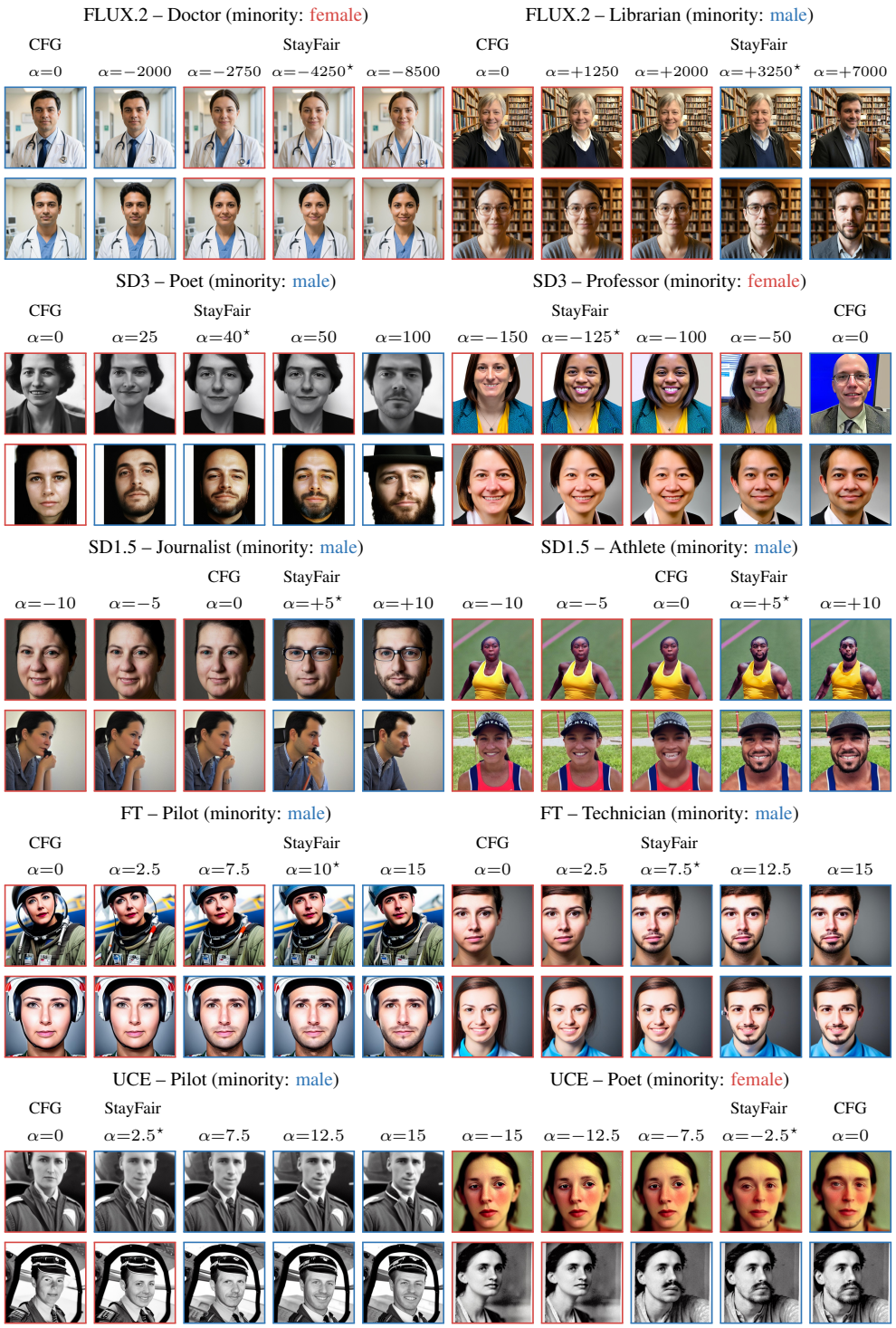


Figure 23: **Qualitative examples across α.** Same seed across α at a fixed guidance scale (w=5.0 on FLUX.2, w=7.5 on SD3, w=12.5 on SD1.5, FT, and UCE); columns list each occupation's α grid and * marks the searched $\alpha^\star$, which is StayFair (Search). Border: female, male, gray = low classifier confidence.

H.2 Class-conditional generation examples

Figure 24 compares CG, RW, and StayFair on CelebA with *Blond Hair* and *Smiling* conditions, supplementing Figure 4. Under CG, increasing w shifts generated images toward the majority group (*e.g.*, female for *Blond Hair*). RW reduces this tendency but does not fully suppress it at high scales.



Figure 24: **Qualitative examples on class-conditional generation.** Four samples per condition (*Blond Hair*, top; *Smiling*, bottom) across guidance scales; rows are CG, RW, and StayFair.

I Limitations and Broader Impact

We extend the limitations and broader impact discussed in Section 6.

Limitations. Extending StayFair to practical multi-group and intersectional settings requires further research on how to select attribute directions.

StayFair (Estimate) relies on many model-specific search results to select prompt-dependent corrections (Appendix C). Collecting these results requires repeated image generation, adding to the cost of extending the estimator to new models and more diverse prompts.

Our theoretical guarantee is established in an idealized Gaussian VE-SDE setting with a log-affine endpoint potential (Appendix E.2). Extending the analysis to more realistic settings with nonlinear neural guidance remains future work.

Broader impact. If α is misestimated, StayFair may shift generation toward one group and worsen the bias it aims to reduce (Appendix C). Our evaluation relies on binary gender labels from an automatic classifier, which may misrepresent individuals and does not capture non-binary identities.

J Licenses for Existing Assets

We list below all third-party assets used in our experiments, together with their licenses. All assets are used in accordance with their respective terms of use, and the original creators are credited at first use in the main paper.

Diffusion Models

- **Stable Diffusion 1.5** [1]: CreativeML Open RAIL-M License.
- **Stable Diffusion 3 (medium, Diffusers release)** [2]: Stability AI Non-Commercial Research Community License.
- **ADM-G (guided-diffusion)** [3]: MIT License.
- **Unconditional ADM checkpoint** [57]: MIT License (from the DDPM-IP codebase).
- **FLUX.2 [klein] 9B Base**: FLUX Non-Commercial License (FLUX NCL).

Software Libraries

- **Diffusers**: Apache License 2.0.

Baseline Methods

- **FT (Finetune-Fair-Diffusion)** [7]: MIT License.
- **UCE (Unified Concept Editing)** [11]: MIT License.
- **FD (Fair Diffusion)** [19]: Apache License 2.0.
- **PAG (Perturbed-Attention Guidance)** [35]: MIT License.
- **GDRO (Group DRO)** [44]: MIT License.

Classifiers and Detectors

- **CLIP** [62]: MIT License.
- **Ultralytics YOLO11 (YOLO11n)** [63]: AGPL-3.0 License.

Evaluation Metrics and Feature Extractors

- **CLIPScore** [64]: MIT License.
- **Aesthetic Scorer (rsinema/aesthetic-scorer)** [65]: MIT License.
- **PickScore** [66]: MIT License.
- **DINOv2** [70]: Apache License 2.0.

Datasets

- **CelebA** [56]: Released for non-commercial research purposes only, under the terms specified by the dataset providers (MMLAB, CUHK).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims in the abstract and introduction are supported by the theory in Section 4 and Appendix E and the experiments in Section 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in Section 6 and Appendix I.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Assumptions are stated in Section 4, and full proofs are in Appendix E.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Experimental settings are described in Section 5 and Appendix F.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code for the text-to-image experiments is available at https://github.com/Kim-Myeong-Soo/stay-fair; all datasets and models we use are publicly available.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: See Section 5 and Appendix F.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Standard deviations across seeds are reported in the tables.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Compute resources are reported in Appendices F.1 and F.2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics, and our research conforms with it.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Broader impacts, including negative ones, are discussed in Section 6 and Appendix I.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: We release no new models or datasets with a high risk of misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Existing assets are cited, and their licenses are listed in Appendix J.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Our code for the text-to-image experiments is released with documentation at https://github.com/Kim-Myeong-Soo/stay-fair.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs were used only for writing and editing.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.