FAIRFLOW: DEMYSTIFYING AND MITIGATING STEREOTYPE BIAS IN TEXT-TO-IMAGE DIFFUSION TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal diffusion transformers (MM-DiTs) underpin modern text-to-image systems, yet can reproduce demographic bias under neutral prompts. Existing debiasing methods were largely designed for U-Net backbones and offer limited guidance on where semantic attributes are controlled in joint transformer architectures. In this work, we trace stereotype bias through MM-DiTs and find that its causal influence is concentrated in a sparse set of *semantic control layers*. These layers exhibit strong visual and cross-modal signals, as revealed by our attention analysis. Crucially, our investigation into text-state trajectories uncovers a pivotal bias reinforcement loop, demonstrating how repeated text-image updates systematically amplify initial bias priors during the generation process. Based on these observations, we propose **FairFlow**, a parameter-preserving method that learns reusable attribute directions and injects them into visual features at the identified control layers during early denoising. Experiments on FLUX.1 and Stable Diffusion 3 across multiple demographic settings show that FairFlow improves the fairness–fidelity tradeoff and remains effective on complex scene prompts.

1 INTRODUCTION

Text-to-image generation is shifting from U-Net denoisers toward multimodal diffusion transformers (MM-DiTs). Models such as FLUX.1 (Black Forest Labs, 2024) and Stable Diffusion 3 (Esser et al., 2024) are increasingly available through commercial generation services (Replicate, 2026; fal.ai, 2026; Together AI, 2026; Stability AI, 2026).

However, deployment at this scale also broadens the impact of stereotype bias. Under neutral prompts, generators can amplify demographic associations in their training data and produce skewed depictions of gender and race (Naik & Nushi, 2023; Ananya, 2024). Such failures affect both represented groups and service providers. Google, for example, temporarily paused Gemini's human image generation after public criticism of inaccurate and biased depictions (The Guardian, 2024).

To address this problem, prior work has explored parameter-updating and parameter-preserving mitigations. Parameter updates can directly alter learned associations, but are costly to adapt to changing fairness targets or demographic ratios (Jung et al., 2025; Qi et al., 2025; Huang et al., 2025). Parameter-preserving methods instead intervene through prompts, text embeddings, or internal visual features. Prompt- and embedding-level methods are lightweight (Friedrich et al., 2023; Sakurai et al., 2025; Fu et al., 2025; Han et al., 2025; Kim et al., 2025), yet often have limited capacity to coordinate multiple demographic attributes, as illustrated in Figure 1. Image-side intervention instead acts directly on how attributes are visually realized while preserving the model and prompt. We therefore pursue this direction for more precise demographic control. Existing designs, however, cannot be transferred to MM-DiTs by simply porting U-Net-based paradigms. In U-Nets, bottleneck features and cross-attention blocks provide well-defined structural locations for semantic intervention (Shi et al., 2025; Li et al., 2024; Park et al., 2025). In contrast, MM-DiTs distribute multimodal computation across transformer blocks and repeatedly update text and image tokens through joint attention. They therefore offer no localized structural interface at which to intervene. This raises a key question: *where do demographic attributes become causally controllable in MM-DiTs, and can these locations support precise inference-time mitigation?*

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



Figure 1: Representative FLUX.1 generations from the vanilla model, FairDiffusion (a prompt-level text-side method) and FairImagen (an embedding-level text-side method), and FairFlow. FairFlow improves demographic diversity across occupations while retaining visual and occupational fidelity.

To answer this question, we first apply residual causal patching across MM-DiT layers and find that causal influence is concentrated in a sparse set of internal control layers. These layers organize visual context and write attribute information into the image stream, as revealed by attention analysis. More importantly, tracking text-state trajectories uncovers a cross-modal bias reinforcement loop: an initial bias prior shapes visual states, which feed back into the text stream and make subsequent multimodal states more bias-oriented. We further show that injecting anti-bias visual states at these layers redirects subsequent text states, disrupting this reinforcement loop. This establishes that image-side intervention can alter bias propagation, but leaves open how to achieve the same effect efficiently at inference time. FairFlow addresses this challenge by learning one reusable attribute direction per targeted layer and attribute, then applying it during early denoising according to the desired target ratios.

Through extensive empirical evaluations across FLUX.1 and Stable Diffusion 3, we demonstrate that FairFlow improves demographic balance while preserving prompt fidelity and image quality, supports adjustable target group ratios, remains effective on complex scene prompts, and incurs negligible inference overhead. Our contributions can be summarized as follows:

- We causally localize stereotype bias influence in MM-DiTs to a sparse set of internal control layers and use attention analysis to characterize their functional roles in organizing visual context and binding attribute semantics.
- We uncover a cross-modal bias reinforcement loop—where bias-aligned visual states continuously feed back into the text stream—and demonstrate that sparse early visual interventions can effectively disrupt this process.
- We introduce a parameter-preserving visual steering approach that extracts fixed attribute directions for a localized layer set. By directly applying these directions during deployment, we effectively mitigate generation bias and enhance demographic fairness.
- Comprehensive evaluations across two MM-DiT architectures and multiple demographic attributes validate our method's highly effective performance in achieving fairness, preserving fidelity, handling complex scenes, disrupting the bias reinforcement loop, and maintaining efficiency.

2 Related Work

2.1 Bias Mitigation in Text-to-Image Models

Existing methods can be organized by whether backbone parameters remain frozen (Elsharif et al., 2025; Wan et al., 2024). Parameter-updating methods fine-tune or edit model parameters (Jung et al., 2025; Qi et al., 2025; Huang et al., 2025). They directly modify learned associations, but can change general generation behavior and require renewed adaptation when fairness targets change. Parameter-preserving methods leave the backbone intact and intervene during generation. They include text-side methods, which add fairness instructions or modify text representations (Friedrich et al., 2023; Sakurai et al., 2025; Fu et al., 2025; Han et al., 2025; Kim et al., 2025), and image-

side methods, which intervene directly in the visual generation process. Text-side interventions are lightweight, but act before the backbone's multimodal computation and leave internal bias formation largely uncontrolled. They can also depend on a particular text-encoder stack, limiting transfer across architectures. Image-side methods are therefore the closest existing line to directly controlling bias as it emerges during generation. Some extract and manipulate attribute directions in a U-Net bottleneck (Shi et al., 2025; Li et al., 2024), while others attach attribute-specific adapters to cross-attention blocks (Park et al., 2025; Teo et al., 2024). These interfaces are natural for U-Nets, but do not map directly to MM-DiTs, which lack a single bottleneck and replace isolated cross-attention with joint multimodal blocks. We therefore investigate where visual states in an MM-DiT causally control attribute semantics and use those locations for mitigation.

2.2 Mechanistic Analysis of Text-to-Image Models

Mechanistic studies use attention attribution, probing, and causal tracing to connect internal components to observable model behavior (Wang et al., 2022; Yao et al., 2024). In text-to-image models, prior work has localized visual attributes and factual knowledge across text encoders, U-Nets, and diffusion transformers (Basu et al., 2023; 2024; Lin et al., 2025). Layer-wise studies of DiTs further identify components important for instruction following or input-image editing (Avrahami et al., 2025; Li et al., 2025). Stable Flow, for example, identifies layers for training-free image editing, where the objective is to modify a source image while preserving its visual structure (Avrahami et al., 2025). Our analysis asks a related localization question for demographic semantics in text-to-image generation, where no source image is provided and the goal is to control the generated semantic distribution. We use residual causal patching to test whether a visual state at a given layer can transfer an attribute between otherwise matched generations, adapting activation-patching ideas from model interpretability (Wang et al., 2022) to the visual stream of an MM-DiT.

3 Preliminary: Rectified Flow and MM-DiT

Diffusion and flow-based text-to-image models generate images by progressively transforming a noise sample into data over multiple sampling steps. Rectified flow formulates this process through the interpolation $\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1$ and learns a text-conditioned velocity field

$$\mathcal{L}_{\mathrm{RF}} = \mathbb{E}\big[\|v_\theta(\mathbf{x}_t, t, c) - (\mathbf{x}_1 - \mathbf{x}_0)\|_2^2\big], \qquad \frac{d\mathbf{x}_t}{dt} = v_\theta(\mathbf{x}_t, t, c), \tag{1}$$

where $\mathbf{x}_0$ and $\mathbf{x}_1$ denote noise and data endpoints, t is the interpolation time, and c is the text condition. The learned field is numerically integrated over multiple sampling steps (Lipman et al., 2022; Liu et al., 2022). Implementations commonly reverse this time convention, describing denoising from 1.0 to 0.0. We use this decreasing scheduler time in the subsequent method and experiments.

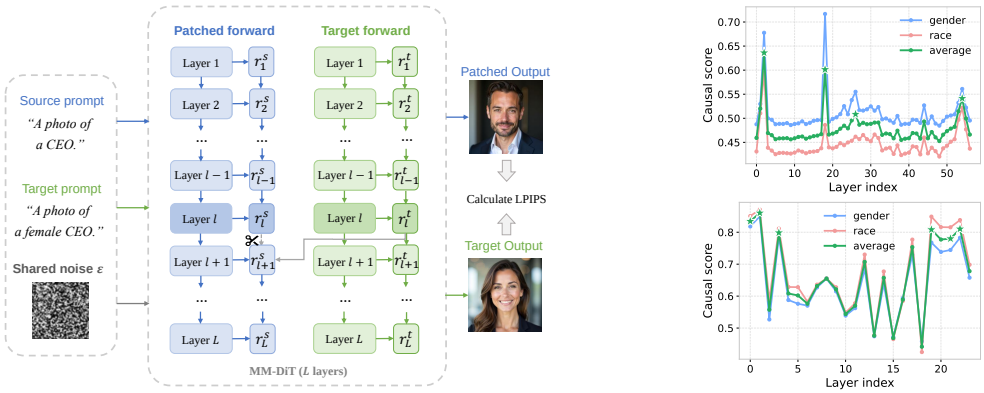
The velocity field is parameterized by a denoising backbone. U-Net backbones inject text conditions through cross-attention, whereas MM-DiTs process text and image tokens with transformer blocks and joint attention, allowing both modalities to be updated (Esser et al., 2024). A transformer block consists of normalization, multi-head attention, a feed-forward network, and residual connections. In an MM-DiT, modality-specific projections produce text and image queries, keys, and values before joint attention combines their information. A *stream* denotes the evolving sequence of features for one modality. SD3 maintains separate text and image streams within multimodal blocks, while FLUX.1 combines dual-stream blocks with single-stream blocks that concatenate both modalities into one sequence (Black Forest Labs, 2024). We use *layer* to denote one transformer block.

4 Tracing Stereotype Bias in MM-DiTs

In this section, we analyze where attribute information becomes causally important in MM-DiTs and how repeated text–image interaction reinforces bias. We first localize the critical control layers for intervention, then characterize their attention patterns and the evolution of internal text states.

4.1 Localizing Control Layers via Residual Causal Patching

Prior U-Net steering methods commonly intervene in bottleneck features, which provide a recognizable latent interface for semantic control (Kwon et al., 2022). MM-DiTs have no corresponding



(a) Residual causal patching (b) Layer-wise causal scores

Figure 2: Residual causal patching for localizing semantic control layers and the resulting causal scores. The upper and lower curves in (b) correspond to FLUX.1 and SD3, respectively.

single bottleneck, so the first step is to determine where a visual intervention can causally transfer an attribute between otherwise matched generations. We use layer-wise residual causal patching (Wang et al., 2022) for this purpose. The procedure asks whether replacing the visual residual at one layer changes the final attribute while holding the prompt pair and sampling noise fixed. Figure 2 summarizes the procedure and results.

Specifically, for source and target prompts that share an occupation but differ in one demographic attribute, we generate with the same seed and replace the source visual residual at layer l with the target residual during early denoising. Let $x^{(t)}$ be the target image and $x_l^{(s\leftarrow t)}$ the patched source image. We define

$$C(l) = 1 - \mathrm{LPIPS}\left(x_l^{(s\leftarrow t)}, x^{(t)}\right), \tag{2}$$

where the Learned Perceptual Image Patch Similarity (LPIPS) metric measures perceptual distance (Zhang et al., 2018). A high score means that patching the visual state at layer l makes the final source generation more similar to the target.

We construct gender- and race-swapped prompt pairs, compute $C(l)$ for each occupation–seed pair, and average across pairs and attributes. Figure 2(b) shows sparse, non-uniform profiles on both models. For FLUX.1, we select dominant local maxima at layers 2, 18, 26, and 54. SD3 has broader peaks, so we select the top-scoring layers 0, 1, 3, 19, 21, and 22. We refer to these locations as *semantic control layers*, or *control layers*. Their high causal-transfer scores identify effective sites for changing attribute semantics through visual features and provide a reusable layer set for subsequent attribute-direction learning. This localization also motivates exploring whether the same layers can support broader semantic control. We return to this possibility with an exploratory style-steering case in Appendix M.

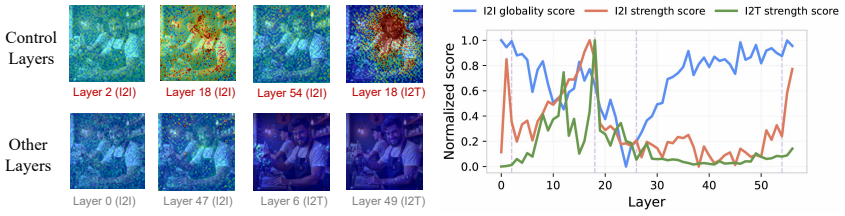
4.2 Attention Patterns at Control Layers

After locating the control layers, we characterize visual interaction and attribute write-in across depth. For image-to-image (I2I) attention, let $A_{h,q,k}^{(l)}$ denote the attention from image query q to image key k in head h, with value $V_{h,k}^{(l)}$. We define the contribution of source token k across image queries as

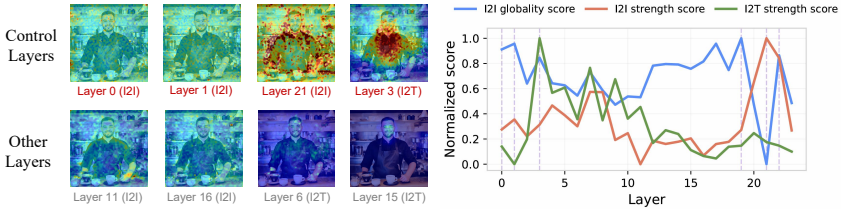
$$M_{\mathrm{I2I}}^{(l)}(k) = \sqrt{\sum_h \left\| \sum_q A_{h,q,k}^{(l)} V_{h,k}^{(l)} \right\|_2^2}. \tag{3}$$

Let $P^{(l)}(k) = M_{\mathrm{I2I}}^{(l)}(k) / \sum_j M_{\mathrm{I2I}}^{(l)}(j)$. For N image tokens, its spatial globality and average strength are

$$G_{\mathrm{I2I}}^{(l)} = \frac{-\sum_k P^{(l)}(k) \log P^{(l)}(k)}{\log N}, \qquad S_{\mathrm{I2I}}^{(l)} = \frac{1}{N} \sum_k M_{\mathrm{I2I}}^{(l)}(k). \tag{4}$$



(a) FLUX.1



(b) SD3

Figure 3: Attention patterns across layers. The heatmaps are computed from $M_{\text{I2I}}^{(l)}$ and $E^{(l)}$, while the curves report normalized I2I globality $G_{\text{I2I}}^{(l)}$, I2I strength $S_{\text{I2I}}^{(l)}$, and the average of $E^{(l)}(i)$. Purple dashed lines indicate the control layers identified by our causal analysis.

For image-to-text (I2T) attention, let $k^\star$ be an explicit attribute token, $A_{h,i,k^\star}^{(l)}$ denote attention from image query i to $k^\star$, and $V_{h,k^\star}^{(l)}$ its value. The target-token write-in map is

$$E^{(l)}(i) = \sqrt{\sum_h \left\| A_{h,i,k^\star}^{(l)} V_{h,k^\star}^{(l)} \right\|_2^2}, \tag{5}$$

which measures the attribute contribution written to each image position.

Figure 3 combines depth-wise attention metrics with representative $M_{\text{I2I}}^{(l)}$ and $E^{(l)}$ heatmaps. Across both architectures, early control layers exhibit broadly distributed visual contributions, mid-depth control layers strongly write prompt attributes into subject regions, and late control layers restore strong image self-attention with broad spatial participation. In each model, the layer with the highest mean $E^{(l)}$ is also identified by residual causal patching, directly linking target-token write-in to causal influence. The displayed non-control layers instead show weaker attribute write-in or spatially concentrated visual contributions. FLUX.1 layer 26 is a local exception, suggesting that residual and feed-forward transformations can complement the dominant attention pattern.

4.3 Cross-Modal Bias Reinforcement

The preceding analysis identifies visual locations that can transfer attribute semantics. We next test whether those visual states also feed back into the text stream, which would make bias formation a recurrent cross-modal process rather than a one-time text-to-image transfer during iterative denoising. Taking SD3 as a representative MM-DiT, we compare text trajectories from matched empty-attribute, bias, and anti-bias occupational prompts. To test whether visual context causally shifts the text trajectory toward the bias or anti-bias reference, we apply visual residual causal patching. At the control layers, we inject visual residuals from a matched bias or anti-bias run into the forward pass under analysis while preserving its text input and sampling seed.

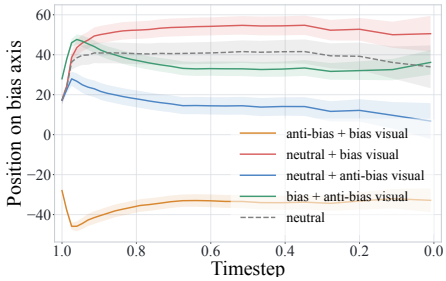
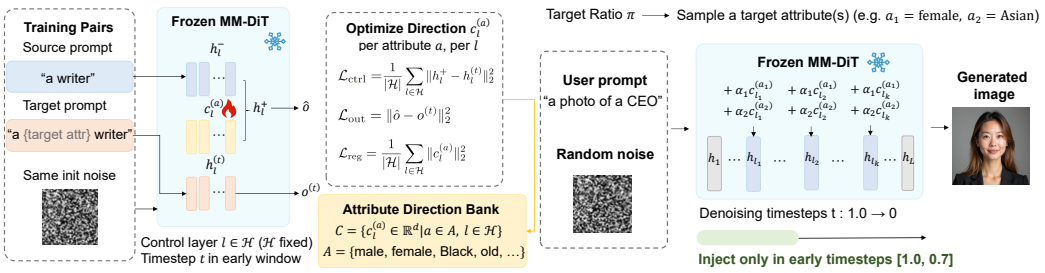


Figure 4: Visual-to-text feedback in SD3 measured by midpoint-centered bias-axis projection P_t.

Let z_t denote the text representation of a patched or unmodified forward pass at timestep t. We define the midpoint c_t between the bias and anti-bias references and the unit direction d_t from the



(a) Attribute direction optimization **(b)** Sparse inference-time injection

Figure 5: Overview of FairFlow. FairFlow precomputes attribute directions, injects them only at the semantic control layers identified by residual causal patching, restricts intervention to an early denoising window, and evaluates the resulting generations with fairness and quality metrics. The design preserves model parameters and supports adjustable target ratios.

anti-bias reference to the bias reference:

$$c_t = \frac{z_t^{\text{bias}} + z_t^{\text{anti}}}{2}, \qquad d_t = \frac{z_t^{\text{bias}} - z_t^{\text{anti}}}{\|z_t^{\text{bias}} - z_t^{\text{anti}}\|_2}. \tag{6}$$

We then measure the signed displacement of z_t along this axis:

$$P_t(z) = \langle z_t - c_t, d_t \rangle. \tag{7}$$

Positive values place the state on the bias side of the reference midpoint, whereas negative values place it on the anti-bias side. Zero indicates no displacement along the axis relative to the midpoint.

Figure 4 provides causal evidence that visual states influence the subsequent text trajectory. Injecting bias-associated visual residuals shifts the text states toward the bias side, whereas anti-bias residuals redirect them toward the anti-bias side. These shifts do not collapse the text states onto the corresponding reference trajectories, indicating that visual feedback modulates rather than replaces prompt-conditioned semantics. In the unmodified forward pass, the empty-attribute trajectory remains on the bias side, showing that this feedback path is bias-aligned during ordinary generation. Together with text-to-visual conditioning in joint attention, these results reveal a reinforcement loop in which biased text states shape visual features that feed back to sustain and amplify the bias tendency. Complementary cosine-similarity results for both causal patching and FairFlow are reported in Appendix C.

5 FairFlow: Sparse Steering at Control Layers

The preceding analysis identifies sparse visual states that control attribute semantics and participate in cross-modal bias reinforcement loop. FairFlow turns them into an inference-time interface without updating MM-DiT parameters. It learns *attribute directions* offline and injects them at the control layers during early denoising. Each attribute direction contains one vector per selected layer, is shared across samples and timesteps, and composes additively with other directions under independently adjustable strengths. These properties parallel semantic directions in U-Net h-space (Kwon et al., 2022), but arise here at causally localized MM-DiT layers. Figure 5 summarizes the workflow.

5.1 Attribute Direction Optimization

Let $\mathcal{H}$ denote the control-layer set. For each attribute a, we form source–target prompt pairs that share an occupation, with only the target explicitly specifying a. Each update samples one seed and timestep and runs both prompts from the same noisy state, isolating the attribute difference from sampling noise. With the MM-DiT frozen, FairFlow learns one visual-space vector $c_l^{(a)}$ per $l \in \mathcal{H}$. Let $h_l^{(t)}$ denote the target visual state and h_l^- the steered-source state immediately before injection, including the effects of earlier interventions. The injected state is $h_l^+ = h_l^- + c_l^{(a)}$; the target and steered-source predictions are $o^{(t)}$ and $\hat{o}$. The same vectors are learned across independently sampled timesteps without backpropagating through the full denoising trajectory.

Control-layer alignment. The main objective matches the steered and target visual states at the causally localized layers:

$$\mathcal{L}_{\text{ctrl}} = \mathbb{E}_{p,z,t}\left[\frac{1}{|\mathcal{H}|}\sum_{l\in\mathcal{H}}\|h_l^+ - h_l^{(t)}\|_2^2\right]. \tag{8}$$

The expectation covers prompt pairs p, seeds z, and sampled timesteps t.

Denoising-output alignment. An auxiliary term aligns their denoising predictions:

$$\mathcal{L}_{\text{out}} = \mathbb{E}_{p,z,t}\left[\|\hat{o} - o^{(t)}\|_2^2\right]. \tag{9}$$

Here o is the predicted velocity.

Direction regularization. We penalize direction magnitude to discourage overly large interventions and preserve perceptual quality:

$$\mathcal{L}_{\text{reg}} = \frac{1}{|\mathcal{H}|}\sum_{l\in\mathcal{H}}\|c_l^{(a)}\|_2^2. \tag{10}$$

The full objective is

$$\mathcal{L} = \lambda_{\text{ctrl}}\mathcal{L}_{\text{ctrl}} + \lambda_{\text{out}}\mathcal{L}_{\text{out}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}, \tag{11}$$

where the three λ terms are hyperparameters. Control-layer alignment provides the primary signal, while output alignment complements this signal and regularization helps preserve perceptual quality.

5.2 Sparse Inference-Time Injection

FairFlow intervenes only at $l \in \mathcal{H}$ during an early denoising window. For a selected attribute set $\mathcal{A}$, it composes the corresponding attribute directions with independently adjustable strengths α_a and restores the original feature norm:

$$h_l^+ = h_l^- + \sum_{a\in\mathcal{A}}\alpha_a c_l^{(a)}, \qquad \hat{h}_l = \frac{\|h_l^-\|_2}{\|h_l^+\|_2}h_l^+. \tag{12}$$

SD3 uses classifier-free guidance (CFG), so we inject the same direction into both branches. Conditional-only injection would be amplified by the guidance scale. FLUX.1 uses guidance modulation and one generation branch, where we inject directly.

6 Experiments

6.1 Experimental Setup

Models. We evaluate FLUX.1-dev and Stable Diffusion 3 Medium. Software, optimization, and inference configurations are provided in Appendix A.

Baselines. We compare with the vanilla model and four representative methods: FairDiffusion (Friedrich et al., 2023) uses attribute-conditioned prediction differences, FairImagen (Fu et al., 2025) projects text toward a neutral representation, WeakGuidance (Kim et al., 2025) injects target-attribute embeddings, and BiasNeuron (Qi et al., 2025) fine-tunes bias-sensitive neurons.

Datasets. Following FairImagen (Fu et al., 2025; Zhao et al., 2018), we use 30 occupations for localization and attribute-direction optimization and a disjoint set of 100 occupations for testing, with the template "`a photo of a face of a {occupation}`". We use 6 seeds per prompt and 50 synthesized complex-scene prompts. The evaluation on bias-irrelevant COCO prompts is described in Appendix H.

Metrics. DeepFace (Taigman et al., 2014) predicts perceived gender, race, and age. For each occupation, its predictions yield group proportions p, scored over k groups following FairImagen:

$$F = 1 - \frac{\sum_i |p_i - 1/k|}{2(1-1/k)}. \tag{13}$$

Table 1: Debiasing performance on FLUX.1.

Method	Gender				Race				Gender×Race				Average			
	Fairness	CLIP	MUSIQ	Tradeoff	Fairness	CLIP	MUSIQ	Tradeoff	Fairness	CLIP	MUSIQ	Tradeoff	Fairness	CLIP	MUSIQ	Tradeoff
Vanilla	0.084	**0.712**	0.735	0.155	0.164	**0.712**	0.735	0.282	0.135	**0.712**	0.735	0.238	0.128	**0.712**	0.735	0.227
FairDiffusion	0.244	0.685	**0.740**	0.391	0.464	0.683	**0.742**	0.631	0.323	0.673	**0.742**	0.485	0.344	0.680	**0.741**	0.509
WeakGuidance	0.486	0.671	0.727	0.647	0.254	0.677	0.735	0.403	0.286	0.651	0.728	0.440	0.342	0.666	0.730	0.505
FairImagen	0.236	0.706	0.728	0.381	0.224	0.705	0.727	0.365	0.210	0.705	0.728	0.347	0.223	0.705	0.728	0.364
BiasNeuron	0.276	0.670	0.726	0.429	0.245	0.670	0.726	0.391	0.222	0.707	0.738	0.363	0.248	0.682	0.730	0.395
FairFlow	**0.600**	0.704	0.725	**0.746**	**0.489**	0.703	0.728	**0.654**	**0.556**	0.687	0.710	**0.706**	**0.548**	0.698	0.721	**0.703**

Table 2: Debiasing performance on SD3.

Method	Gender				Race				Gender×Race				Average			
	Fairness	CLIP	MUSIQ	Tradeoff	Fairness	CLIP	MUSIQ	Tradeoff	Fairness	CLIP	MUSIQ	Tradeoff	Fairness	CLIP	MUSIQ	Tradeoff
Vanilla	0.113	0.707	**0.749**	0.204	0.176	0.707	**0.749**	0.300	0.156	0.707	**0.749**	0.270	0.149	0.707	**0.749**	0.259
FairDiffusion	**0.648**	0.583	0.484	0.687	0.468	0.578	0.475	0.567	0.452	0.560	0.504	0.558	0.523	0.574	0.488	0.608
WeakGuidance	0.300	0.697	0.748	0.461	0.174	0.701	0.747	0.296	**0.508**	0.688	0.739	**0.669**	0.327	0.695	0.745	0.492
FairImagen	0.173	0.682	0.719	0.293	0.218	0.686	0.717	0.356	0.174	0.684	0.717	0.295	0.188	0.684	0.718	0.315
BiasNeuron	0.128	**0.710**	0.738	0.227	0.160	**0.709**	0.738	0.276	0.124	**0.710**	0.737	0.220	0.137	**0.710**	0.738	0.241
FairFlow	0.637	0.698	0.744	**0.775**	**0.741**	0.699	0.731	**0.845**	0.490	0.678	0.708	0.647	**0.623**	0.692	0.728	**0.760**

Gender uses $k = 2$ with male/female labels, while race uses $k = 3$ with White/Black/Asian labels. $F = 1$ denotes balance and $F = 0$ a single group. Fidelity uses CLIP ViT-B/16 text–image alignment and MUSIQ perceptual quality (Radford et al., 2021; Ke et al., 2021). We also report

$$U = \sqrt{\frac{\text{CLIP}}{\text{CLIP}_{\text{vanilla}}} \cdot \frac{\text{MUSIQ}}{\text{MUSIQ}_{\text{vanilla}}}}, \qquad T = \frac{2FU}{F+U}, \tag{14}$$

where U is fidelity normalized to the vanilla model and T is their harmonic-mean fairness–fidelity tradeoff. Higher values are better.

6.2 Debiasing Performance

Tables 1 and 2 compare gender, race, and intersectional gender×race settings. Appendix E provides additional age-debiasing results.

Across both architectures, FairFlow achieves the best average fairness and fairness–fidelity tradeoff, with gains distributed across attributes rather than concentrated in one setting. In contrast, the baselines either preserve fidelity with limited or attribute-specific mitigation, or improve balance at a substantial utility cost.

FLUX.1. FairFlow ranks first in fairness and tradeoff for all three targets, with its clearest advantage under intersectional control. The strongest baselines vary by attribute: WeakGuidance favors gender and FairDiffusion favors race, while FairImagen and BiasNeuron preserve fidelity but yield smaller fairness gains. FairFlow instead combines consistent mitigation with near-vanilla fidelity.

SD3. FairFlow provides the strongest overall balance on SD3, achieving the highest average fairness and tradeoff while retaining substantially more utility than aggressive prompt-level intervention. Its advantage is most pronounced for race, where it leads both fairness and tradeoff, and it also obtains the best gender tradeoff. FairDiffusion reaches slightly higher gender fairness but incurs pronounced CLIP and MUSIQ degradation. WeakGuidance performs best in the intersectional setting, yet its mitigation is uneven across attributes. FairFlow remains competitive in that setting and is the only method that combines strong gains across all three targets into the best aggregate result.

6.3 Complex Scene Prompts

Figure 6 evaluates whether FairFlow generalizes to context-rich scenes. Across both architectures, FairFlow consistently improves fairness without substantially compromising fidelity. On FLUX.1, it achieves the highest gender and race fairness, while FairDiffusion attains higher intersectional

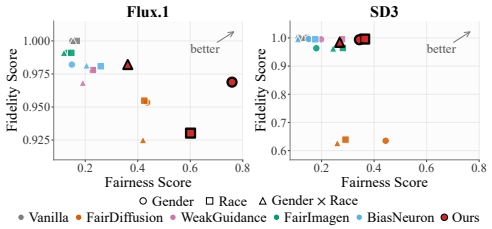


Figure 6: Complex-scene tradeoffs. Colors indicate methods; shapes indicate gender, race, or intersectional debiasing.

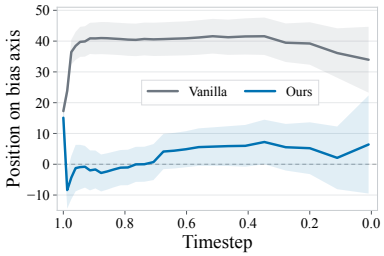


Figure 7: FairFlow shifts SD3 text states toward the anti-bias side.

fairness with a substantially larger fidelity loss. On SD3, FairFlow leads race fairness and achieves the second-highest gender fairness while retaining near-vanilla fidelity. Although WeakGuidance performs best for intersectionality, its improvements are less consistent across the other attributes. Detailed results are provided in Appendix G.

6.4 Disrupting Cross-Modal Bias Reinforcement

Figure 7 compares the bias-axis positions of the unmodified and FairFlow-steered text trajectories using the definition in Section 4.3. The unmodified trajectory remains on the bias side, whereas FairFlow produces a sustained shift toward the anti-bias side. This separation emerges early and persists after the intervention window, showing that visual-only steering affects subsequent text states. Together with the causal patching results, the pattern supports disruption of bias-oriented cross-modal feedback and connects the intervention mechanism to the demographic improvements in generated images.

6.5 Inference-Time Efficiency

Table 3 compares additional runtime on SD3 for methods that modify inference. FairFlow adds 0.056s, approximately one third of WeakGuidance's overhead and far below the seconds added by FairDiffusion and FairImagen. This low cost follows from injecting precomputed attribute directions at sparse control layers during early denoising. The same directions are reused across samples, so their offline optimization cost is amortized across deployment requests. These measurements show that the fairness gains can be obtained with little additional inference work.

Table 3: SD3 inference overhead.

Method	Overhead (s)
FairDiffusion	6.002
WeakGuidance	0.166
FairImagen	3.240
FairFlow	**0.056**

7 Discussion

Our analysis localizes attribute transfer at layer granularity, leaving open which attention heads, token interactions, or feed-forward subspaces implement this behavior. Decomposing these internal components could enable more precise and selective control. FairFlow uses fixed attribute directions throughout the early denoising window to support efficient reuse. Future work could explore context-aware composition while retaining this simplicity. Beyond internal representation steering, the properties of rectified-flow velocity fields and denoising trajectories provide a complementary perspective for understanding and mitigating how bias develops during generation.

8 Conclusion

We identify sparse semantic control layers and a cross-modal bias reinforcement loop that contribute to stereotype bias in MM-DiTs. Guided by these findings, FairFlow applies reusable attribute directions at the localized layers during early denoising. Experiments on FLUX.1 and SD3 demonstrate improved fairness–fidelity tradeoffs with negligible inference overhead, establishing a practical mechanism-guided approach to bias mitigation.

AI USE STATEMENT

We used GPT-5.6 Sol for grammar correction and light style polishing and Gemini-3.1-Pro to synthesize the complex-scene T2I prompts. We did not use generative AI tools to produce unverified experimental results, claims, or citations. All AI-assisted edits were reviewed by the authors, and the authors take responsibility for the final content of the paper.

ETHICS STATEMENT

This work aims to reduce stereotype bias in generated images. Its simplified demographic categories and automated attribute estimates are experimental proxies, not exhaustive accounts of identity or ground-truth social judgments. FairFlow targets underspecified prompts and should not override explicit or fact-sensitive attributes. Released artifacts contain no private personal data and are intended for reproducible defensive research.

REPRODUCIBILITY STATEMENT

We provide the code needed to reproduce our experiments in the supplementary material. The main text describes the method and evaluation protocol, while Appendix A documents implementation settings and model links. Additional experimental setups and results are reported in the remaining appendices.

REFERENCES

Ananya. Ai image generators often give racist and sexist results: Can they be fixed?, March 2024. URL https://www.nature.com/articles/d41586-024-00674-9. Nature News Feature. Accessed April 17, 2026.

Omri Avrahami, Or Patashnik, Ohad Fried, Egor Nemchinov, Kfir Aberman, Dani Lischinski, and Daniel Cohen-Or. Stable flow: Vital layers for training-free image editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7877–7888, 2025.

Samyadeep Basu, Nanxuan Zhao, Vlad I Morariu, Soheil Feizi, and Varun Manjunatha. Localizing and editing knowledge in text-to-image generative models. In *The Twelfth International Conference on Learning Representations*, 2023.

Samyadeep Basu, Keivan Rezaei, Priyatham Kattakinda, Vlad I Morariu, Nanxuan Zhao, Ryan A Rossi, Varun Manjunatha, and Soheil Feizi. On mechanistic knowledge localization in text-to-image generative models. In *Forty-first International Conference on Machine Learning*, 2024.

Black Forest Labs. Announcing Black Forest Labs. https://blackforestlabs.ai/announcing-black-forest-labs/, 2024.

Wala Elsharif, Mahmood Alzubaidi, and Marco Agus. Cultural bias in text-to-image models: a systematic review of bias identification, evaluation and mitigation strategies. *IEEE Access*, 2025.

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

fal.ai. Generative media platform for developers. https://fal.ai/, 2026.

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

Zihao Fu, Ryan Brown, Shun Shao, Kai Rawal, Eoin Delaney, and Chris Russell. Fairimagen: Post-processing for bias mitigation in text-to-image models. *arXiv preprint arXiv:2510.21363*, 2025.

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. *arXiv preprint arXiv:2509.23639*, 2025.

Lin-Chun Huang, Ching Chieh Tsao, Fang-Yi Su, and Jung-Hsien Chiang. Debiasing diffusion model: Enhancing fairness through latent representation learning in stable diffusion model. *arXiv preprint arXiv:2503.12536*, 2025.

Sangwon Jung, Alex Oesterling, Claudio Mayrink Verdun, Sajani Vithana, Taesup Moon, and Flavio P Calmon. Multi-group proportional representations for text-to-image models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23744–23754, 2025.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5148–5157, 2021.

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

Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022.

Binglei Li, Mengping Yang, Zhiyu Tan, Junping Zhang, and Hao Li. Revisiting block-wise interactions of MMDit for training-free improved synthesis, 2025. URL https://openreview.net/forum?id=2VuPBAH94k.

Hang Li, Chengzhi Shen, Philip Torr, Volker Tresp, and Jindong Gu. Self-discovering interpretable diffusion latent directions for responsible text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12006–12016, 2024.

Zihao Lin, Samyadeep Basu, Mohammad Beigi, Varun Manjunatha, Ryan A Rossi, Zichao Wang, Yufan Zhou, Sriram Balasubramanian, Arman Zarei, Keivan Rezaei, et al. A survey on mechanistic interpretability for multi-modal foundation models. *arXiv preprint arXiv:2502.17516*, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Ranjita Naik and Besmira Nushi. Social biases through the text-to-image generation lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 786–808, 2023.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17567–17576, 2025.

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

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

Replicate. Run ai with an api. https://replicate.com/, 2026.

Jinya Sakurai, Yuki Koyama, and Issei Sato. Fairt2i: Mitigating social bias in text-to-image generation via large language model-assisted detection and attribute rebalancing. *arXiv preprint arXiv:2502.03826*, 2025.

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

Stability AI. Stability ai developer platform. `https://platform.stability.ai/docs/api-reference`, 2026.

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

Christopher T Teo, Milad Abdollahzadeh, Xinda Ma, and Ngai-Man Cheung. Fairqueue: Rethinking prompt learning for fair text-to-image generation. *Advances in Neural Information Processing Systems*, 37:22878–22926, 2024.

The Guardian. 'we definitely messed up': why did Google's AI tool make offensive historical images? `https://www.theguardian.com/technology/2024/mar/08/we-definitely-messed-up-why-did-google-ai-tool-make-offensive-historical-images`, March 2024.

Together AI. Text-to-image generation api. `https://docs.together.ai/docs/inference/images/overview`, 2026.

Yixin Wan, Arjun Subramonian, Anaelia Ovalle, Zongyu Lin, Ashima Suvarna, Christina Chance, Hritik Bansal, Rebecca Pattichis, and Kai-Wei Chang. Survey of bias in text-to-image generation: Definition, evaluation, and mitigation. *arXiv preprint arXiv:2404.01030*, 2024.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

Yunzhi Yao, Ningyu Zhang, Zekun Xi, Mengru Wang, Ziwen Xu, Shumin Deng, and Huajun Chen. Knowledge circuits in pretrained transformers. *Advances in neural information processing systems*, 37:118571–118602, 2024.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

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

APPENDIX CONTENTS

A. Implementation Details and Model Links 13
B. Layer-Wise Residual Causal Patching Cases 13
C. Cross-Modal Bias Reinforcement Setup 14
D. Deployment Usage 14
E. Additional Attribute Type 15
F. Complex-scene Prompt Generation 15
G. Detailed Complex Scene Results 16
H. Utility on Bias-Irrelevant Prompts 17
I. Detailed Design Ablations 18
J. Loss Ablation 18
K. Fairness–Fidelity Pareto Curves 19
L. VLM Validation for DeepFace-based Evaluation 20
M. Exploration of General Semantic Controls 20

A IMPLEMENTATION DETAILS AND MODEL LINKS

Experiments use eight NVIDIA RTX 4090 GPUs with Ubuntu 24.04.1 LTS, Python 3.12.8, and PyTorch 2.6.0+cu118. We use the public FLUX.1-dev checkpoint at `https://huggingface.co/black-forest-labs/FLUX.1-dev` and the SD3 Medium checkpoint at `https://huggingface.co/stabilityai/stable-diffusion-3-medium-diffusers`. Following the recommended generation settings on the corresponding Hugging Face model pages, SD3 uses a guidance scale of 7.0 and 28 sampling steps, whereas FLUX.1 uses a guidance scale of 3.5 and 50 sampling steps. All images are generated at a resolution of 1024×1024. Attribute directions are optimized with Adam at a learning rate of 5×10^{-4} for 30 epochs on FLUX.1 and 100 epochs on SD3, using $(\lambda_{\text{ctrl}}, \lambda_{\text{out}}, \lambda_{\text{reg}}) = (1.0, 0.2, 0.01)$. For gender, race, intersectional, and age steering, respectively, we set α to $(7.0, 5.0, 2.0, 5.0)$ on FLUX.1 and $(4.0, 4.0, 3.5, 3.0)$ on SD3. Both residual causal patching and inference-time steering use the normalized early denoising interval $t \in [1.0, 0.7]$. CLIP scores use `https://huggingface.co/openai/clip-vit-base-patch16`. For SD3, the same attribute direction is injected into both classifier-free guidance branches; FLUX.1 uses a single guidance-modulated generation branch. To compare debiasing capabilities in a concise and intuitive manner, the main experiments assign target attributes uniformly and randomly, ensuring that each target group has the same number of generated images for every occupation.

Metric aggregation. For occupational evaluations, we compute demographic proportions and fairness separately for each occupation and average fairness across occupations. CLIP and MUSIQ are averaged over all images. For each attribute setting, U and T are computed from these averaged Fairness, CLIP, and MUSIQ scores and the corresponding vanilla averages. The Average Tradeoff across gender, race, and their intersection is computed from the average Fairness, CLIP, and MUSIQ scores across the three settings. COCO utility is averaged over the sampled captions. Age fairness uses the aggregate five-group distribution described in Appendix E.

B LAYER-WISE RESIDUAL CAUSAL PATCHING CASES

Figure 8 complements the aggregate causal scores with layer-wise qualitative evidence. For a matched source–target pair generated from the same seed, we replace the source visual residual at one layer with the target residual. On SD3, patching the identified control layers produces visibly stronger transfer of the target gender attribute and lower LPIPS to the target image than patching representative non-control layers. The figure also includes the corresponding FLUX.1 comparison.



(a) SD3



(b) FLUX.1

Figure 8: Layer-wise residual causal patching. Patching SD3 control layers transfers the target attribute more strongly than patching representative non-control layers. LPIPS is measured against the target image.

C Cross-Modal Bias Reinforcement Setup

On SD3, we construct 20 matched prompt triplets using "`a photo of a face of a {attribute} {occupation}`", where the attribute is bias-associated, anti-bias-associated, or omitted. We evaluate each triplet with eight seeds, yielding 160 prompt–seed instances. At each timestep, we average all text-token states within each transformer layer to obtain one vector per layer, then average these vectors across layers to form the text representation z_t. We compute M_t and P_t for each instance, then report their means and 95% confidence intervals over all instances.

As a cosine-based complement to the projection reported in the main text, we measure the relative bias preference

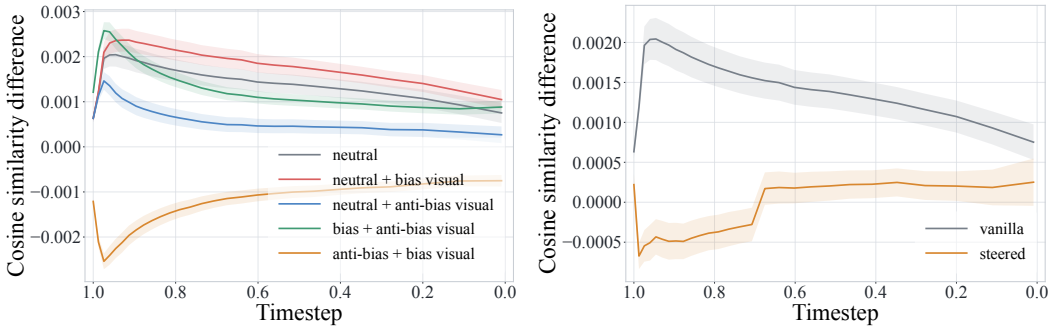
$$M_t(z) = \cos(z_t, z_t^{\text{bias}}) - \cos(z_t, z_t^{\text{anti}}). \tag{15}$$

Larger values indicate stronger relative alignment with the bias reference. Figure 9 reports this diagnostic for both visual residual causal patching and FairFlow steering, corroborating the directional changes in the main-text bias-axis projections.

D Deployment Usage

In practical deployment, FairFlow can be applied by sampling among precomputed attribute directions according to the operator's desired target ratio. For example, a service provider may sample gender attribute directions with a $1:1$ ratio, or adopt any other application-specific ratio for gender, race, or age. In this sense, deployment mainly requires choosing the target distribution and then applying the corresponding attribute directions at inference time.

When the prompt explicitly specifies a demographic attribute, age cue, or a particular public figure, we recommend temporarily disabling the intervention to avoid overriding clear user intent. In practice, this can be implemented with lightweight engineering rather than heavy semantic parsing, for example by maintaining a keyword list together with a simple rule-based prompt parser. This



(a) Visual residual causal patching (b) FairFlow steering

Figure 9: Relative bias preference measured by cosine-similarity difference on SD3. Shaded regions denote 95% confidence intervals.

Table 4: Additional age debiasing results on SD3 and FLUX.1.

Model	Method	Fairness	CLIP	MUSIQ
SD3	Vanilla	0.413	**0.707**	**0.749**
	FairFlow	**0.559**	0.705	0.727
FLUX.1	Vanilla	0.389	**0.712**	**0.735**
	FairFlow	**0.550**	0.704	0.733

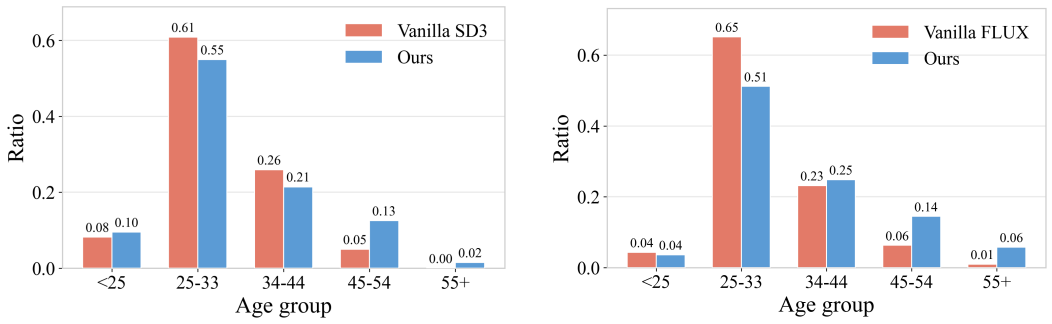
makes the deployment logic straightforward while preserving the main benefit of FairFlow, namely low-cost inference-time mitigation.

E Additional Attribute Type

To further examine extensibility beyond gender and race, we additionally evaluate FairFlow on age debiasing for both SD3 and FLUX.1. We optimize two age directions using the exact prompt modifiers `young` and `old`. At inference, the two directions are assigned uniformly within each occupation, corresponding to a $1:1$ target-direction ratio. Table 4 shows that FairFlow consistently improves age fairness on both architectures, while incurring only minor changes in CLIP and MUSIQ. The fairness score increases from 0.413 to 0.559 on SD3 and from 0.389 to 0.550 on FLUX.1. Figure 10 further shows that the generated age distribution becomes more balanced after intervention. Figure 11 provides qualitative examples, illustrating that FairFlow can produce more age-diverse outputs without obvious prompt collapse. These results suggest that the proposed control-layer steering framework extends naturally to additional demographic attributes. Age fairness is computed from the five age-group proportions displayed in Figure 10 across all occupations: below 25, 25–33, 34–44, 45–54, and 55 or above.

F Complex-scene Prompt Generation

We use Gemini-3.1-Pro to construct complex-scene prompts from profession names. The following prompt was used to generate realistic contexts while preserving the profession as the central semantic content:



(a) SD3 (b) FLUX.1

Figure 10: Age-group distributions for vanilla generation and FairFlow on SD3 and FLUX.1. FairFlow yields a more balanced distribution across age groups on both architectures.



(a) SD3 (b) FLUX.1

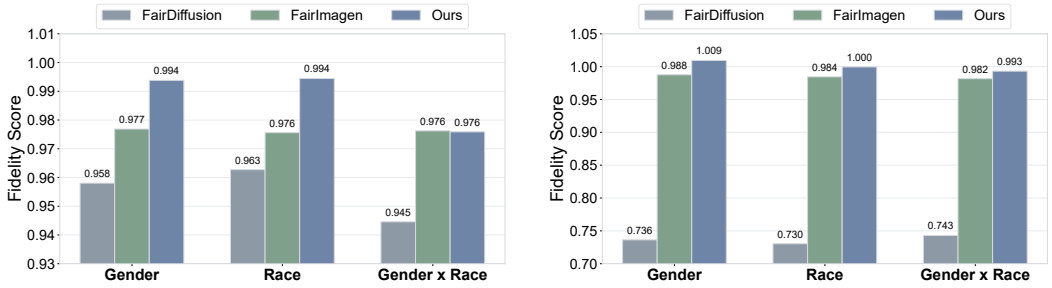
Figure 11: Qualitative age-debiasing cases for SD3 and FLUX.1. Compared with vanilla generation, FairFlow produces more age-diverse results while preserving overall prompt semantics.

Prompt

```
You are generating prompts for a text-to-image model.
Given a profession, write one realistic and detailed prompt
describing a person in that profession within a visually rich
real-world scenario.  Each prompt should include:
1.  The profession as the central subject.
2.  A specific environment or location relevant to the profession.
3.  Natural human activity or occupational context when appropriate.
Use the following template:
a photo of a [PROFESSION] [REALISTIC CONTEXT]
Profession:  [PROFESSION]
```

G Detailed Complex Scene Results

We evaluate all 50 synthesized complex-scene prompts using six seeds per prompt. Figure 12 compares representative vanilla and FairFlow generations on richer occupational scenes for both architectures. Detailed results are provided in Table 5 and Table 6.



(a) FLUX.1 (b) SD3

Figure 13: Fidelity scores on bias-irrelevant COCO prompts.



Figure 12: Complex-scene cases on FLUX.1 (left) and SD3 (right).

Table 5: Detailed complex-scene results on FLUX.1.

Method	Gender				Race				Gender×Race				Average			
	Fairness	CLIP	MUSIQ	Tradeoff	Fairness	CLIP	MUSIQ	Tradeoff	Fairness	CLIP	MUSIQ	Tradeoff	Fairness	CLIP	MUSIQ	Tradeoff
Vanilla	0.151	0.779	0.689	0.262	0.168	0.779	0.689	0.288	0.154	0.779	0.689	0.267	0.158	0.779	0.689	0.273
FairDiffusion	0.436	0.697	**0.700**	0.598	0.425	0.701	**0.698**	0.588	**0.420**	0.658	**0.698**	**0.578**	0.427	0.685	**0.699**	0.588
WeakGuidance	0.234	0.772	0.665	0.378	0.229	0.765	0.671	0.371	0.191	0.758	0.664	0.319	0.218	0.765	0.667	0.356
FairImagen	0.132	0.773	0.682	0.233	0.147	0.774	0.681	0.256	0.119	0.774	0.681	0.212	0.133	0.774	0.681	0.235
BiasNeuron	0.149	**0.782**	0.662	0.259	0.259	**0.780**	0.662	0.410	0.205	**0.781**	0.662	0.339	0.204	**0.781**	0.662	0.338
FairFlow	**0.759**	0.728	0.692	**0.851**	**0.601**	0.754	0.616	**0.730**	0.362	0.758	0.683	0.529	**0.574**	0.747	0.664	**0.719**

Table 6: Detailed complex-scene results on SD3.

Method	Gender				Race				Gender×Race				Average			
	Fairness	CLIP	MUSIQ	Tradeoff	Fairness	CLIP	MUSIQ	Tradeoff	Fairness	CLIP	MUSIQ	Tradeoff	Fairness	CLIP	MUSIQ	Tradeoff
Vanilla	0.113	0.790	0.714	0.203	0.143	**0.790**	0.714	0.250	0.132	**0.790**	0.714	0.233	0.129	**0.790**	0.714	0.229
FairDiffusion	**0.444**	0.468	0.486	**0.523**	0.291	0.483	0.477	0.400	0.260	0.413	0.537	0.368	**0.332**	0.455	0.500	0.430
WeakGuidance	0.198	0.782	0.714	0.330	0.278	0.784	0.713	0.435	**0.340**	0.774	0.712	**0.506**	0.272	0.780	0.713	0.424
FairImagen	0.180	0.751	0.697	0.303	0.281	0.749	0.700	0.435	0.244	0.751	0.696	0.389	0.235	0.750	0.698	0.376
BiasNeuron	0.151	0.781	**0.716**	0.262	0.176	0.780	**0.716**	0.299	0.110	0.781	**0.716**	0.198	0.146	0.781	**0.716**	0.253
FairFlow	0.344	**0.791**	0.704	0.511	**0.364**	0.789	0.709	**0.533**	0.269	0.786	0.695	0.423	0.326	0.789	0.703	**0.491**

H Utility on Bias-Irrelevant Prompts

We randomly sample 200 captions from the COCO-2017 validation set and generate one image per caption using varying seeds. Figure 13 reports the resulting fidelity comparison. Sparse steering remains close to vanilla behavior on prompts unrelated to the protected attributes.

Figure 14 provides complementary qualitative comparisons on COCO-based prompts for FLUX.1 and SD3. Across varied objects and scenes, FairFlow preserves the requested content, composition, and overall visual quality relative to vanilla generation. These cases support the quantitative results

by showing that sparse attribute steering introduces little visible change when the prompt does not invoke the targeted demographic semantics.

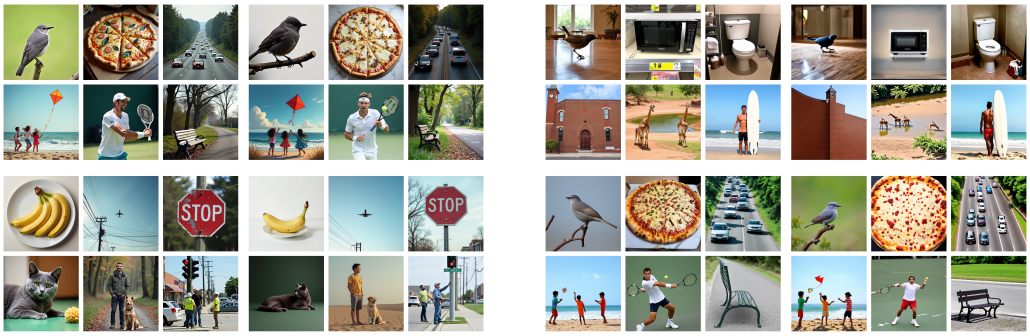


Figure 14: Qualitative comparison on bias-irrelevant COCO prompts. Left: FLUX.1. Right: SD3. In both models, FairFlow preserves the main scene semantics and visual quality while leaving unrelated generation behavior largely stable.

I Detailed Design Ablations

Table 7 provides the complete SD3 ablation results. Random Layers reports the mean of three independently sampled layer sets. Random Layers and Bottom Layers each use six layers, matching the number of SD3 control layers. Low and High α change each default strength by -0.5 and $+0.5$, respectively. DirectDiff directly subtracts source hidden states from target hidden states, without optimizing the resulting attribute direction.

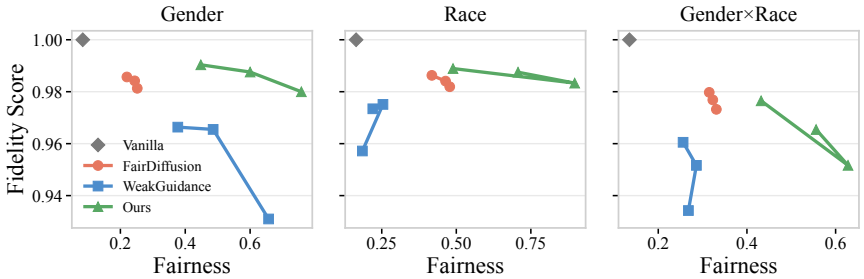
Optimized directions achieve higher fairness than DirectDiff in all three settings, demonstrating the benefit of learning a reusable intervention. The identified control layers also outperform equally sized random and bottom layer sets in fairness. These comparisons support both direction optimization and causal layer selection as contributors to effective demographic control. The strength sweep exhibits a consistent tradeoff: reducing α retains utility closer to the unmodified model, whereas increasing it improves fairness but gradually reduces perceptual quality. The default strengths lie between these alternatives and provide a strong balance across all three attribute settings.

Table 7: Ablation study on bias-prone prompts for SD3. We compare attribute-direction construction, layer selection, and steering strength.

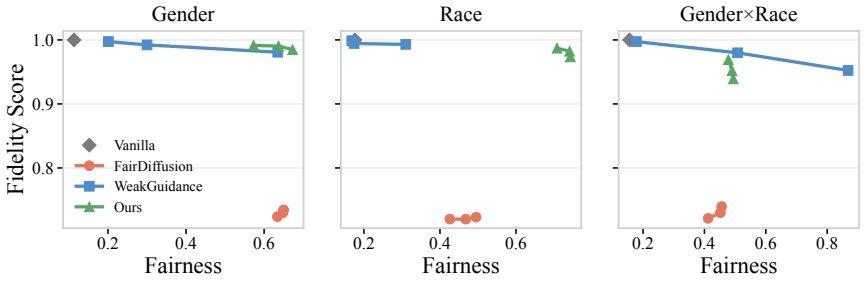
Method	Gender			Race			Gender×Race		
	Fairness	CLIP	MUSIQ	Fairness	CLIP	MUSIQ	Fairness	CLIP	MUSIQ
DirectDiff	0.090	0.720	0.749	0.183	0.721	0.747	0.134	0.720	0.748
Random Layers	0.430	0.713	0.749	0.506	0.717	0.742	0.398	0.702	0.743
Bottom Layers	0.303	0.703	0.749	0.290	0.706	0.749	0.310	0.703	0.744
Low α	0.573	0.696	0.748	0.708	0.701	0.736	0.478	0.689	0.722
High α	0.673	0.697	0.737	0.743	0.694	0.723	0.494	0.671	0.696
FairFlow	0.637	0.698	0.744	0.741	0.699	0.731	0.490	0.678	0.708

J Loss Ablation

We further study the role of each loss component using gender debiasing on SD3. Table 8 confirms that control-layer alignment provides the essential supervision. Without $\mathcal{L}_{\text{ctrl}}$, generation quality



(a) FLUX.1



(b) SD3

Figure 15: Fairness–Fidelity Pareto curves under different intervention strengths on (a) FLUX.1 and (b) SD3.

degrades to the point that facial attributes cannot be evaluated reliably, accompanied by substantial drops in both CLIP and MUSIQ. The auxiliary objectives have complementary roles. Removing $\mathcal{L}_{\text{out}}$ causes smaller reductions in fairness and perceptual quality. Removing $\mathcal{L}_{\text{reg}}$ slightly increases fairness but lowers MUSIQ with unchanged CLIP, indicating that magnitude regularization primarily helps preserve perceptual quality. Using only $\mathcal{L}_{\text{ctrl}}$ degrades all three metrics, further supporting the combined objective.

Table 8: Loss ablation on gender debiasing for SD3.

Loss setting	Fairness	CLIP	MUSIQ
Full objective	<u>0.637</u>	**0.698**	**0.744**
w/o $\mathcal{L}_{\text{ctrl}}$	N/A[†]	0.547	0.521
w/o $\mathcal{L}_{\text{reg}}$	**0.653**	**0.698**	0.735
w/o $\mathcal{L}_{\text{out}}$	0.613	**0.698**	<u>0.740</u>
only $\mathcal{L}_{\text{ctrl}}$	0.494	0.671	0.681

[†] Fairness is unavailable because the generated image quality is too low for reliable facial-attribute evaluation.

K Fairness–Fidelity Pareto Curves

Figure 15 varies the intervention strength for FairFlow, FairDiffusion, and WeakGuidance. On FLUX.1, FairFlow consistently defines the strongest frontier across all three attributes, reaching markedly higher fairness while retaining high fidelity. The updated SD3 curves reveal a more attribute-dependent pattern. FairFlow establishes strong frontiers for gender and especially race, avoiding the pronounced fidelity loss of FairDiffusion and matching or improving upon WeakGuidance at comparable fairness levels. WeakGuidance instead provides the strongest intersectional frontier. Across the panels, varying intervention strength moves each method along its fairness–fidelity curve. FairFlow's best average tradeoff therefore reflects broad performance across attributes rather than an advantage confined to a single strength or setting.

L VLM Validation for DeepFace-based Evaluation

We further validate the reasonableness of using DeepFace for attribute evaluation by comparing its outputs against a stronger VLM-based annotator. Specifically, we randomly sample 100 face images generated by SD3 and query Qwen-3.6-Plus with the following constrained VQA prompt:

Prompt

```
You are an image annotation assistant. Analyze the provided face
image and estimate the person's perceived gender, perceived race
category, and apparent age.

Use only the following labels:
gender: male, female
race: white, black, asian

Always choose the closest label from the allowed options. Do not
output any label outside the allowed sets.

Return JSON only, with exactly these keys:
{
"gender": "male|female",
"race": "white|black|asian",
"age": integer
}

Do not include explanations, markdown, or any text outside the JSON
object.
```

We then compare the VLM outputs with the corresponding DeepFace predictions. The two annotators exhibit very high agreement on gender and race, with an exact-match agreement rate of 0.980 for gender and 0.940 for race. For age, the mean absolute difference between their estimates is 6.64 years, and the Pearson correlation is 0.773. These results support the use of DeepFace as a lightweight attribute evaluator in our main experiments.

M Exploration of General Semantic Controls

To examine whether the localization result extends beyond demographic attributes, we construct 30 attribute pairs involving properties such as color and material and perform residual causal patching on SD3. Figure 16 shows that layers producing low patched-image LPIPS scores substantially overlap with the control layers identified in the main experiments. This alignment suggests that the identified control locations capture a reusable mechanism for semantic control rather than a property specific only to demographic attributes.

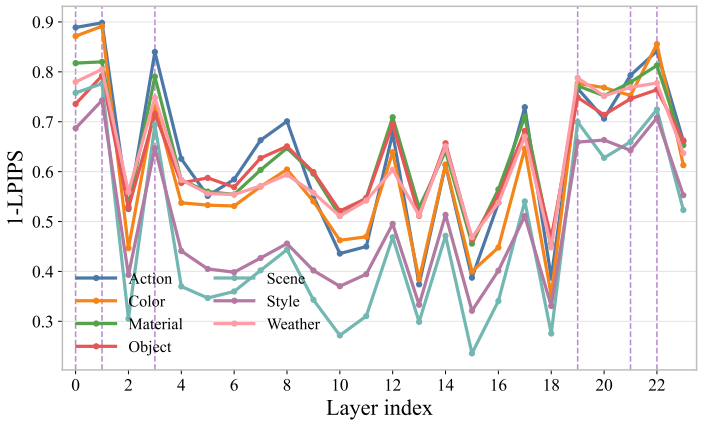


Figure 16: Residual causal patching across 30 non-demographic attribute pairs. Low-LPIPS layers substantially overlap with the identified control layers.

We further test whether FairFlow generalizes beyond demographic debiasing by optimizing four attribute directions at the identified SD3 control layers: watercolor style, rainy weather, metal material, and walking action. Figure 17 compares generations with and without steering under the same prompt and sampling seed. Across all four cases, steering introduces the intended attribute while preserving the underlying subject and scene composition. These results demonstrate that the localized steering mechanism can extend to general semantic control.



Figure 17: General semantic control on SD3 using attribute directions optimized at the identified control layers. Each pair uses the same prompt and seed with and without steering.