DEBIAS ANYTHING: FAIRNESS WITH DIVERSITY WITHOUT SUPERVISION IN DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Although diffusion models produce high-quality images, they also reproduce and amplify demographic imbalances in their training data. Debiasing their generation process post-training w.r.t. some sensitive attribute usually relies on classifier guidance or explicit text extra-conditioning, but this reduces methods' applicability and output diversity. Conversely, methods promoting diversity alone do not ensure fair attribute representation. In this paper, we propose a method tackling fairness and diversity jointly that is generally applicable to any diffusion model and any sensitive attribute. To this end, an adapter connects the frozen diffusion model to a pretrained vision-language embedding space, enabling fairness and diversity guidance without sensitive-attribute annotations. For fairness, pairs of text prompts define attribute directions which guide batch composition towards specific proportions. For diversity, we introduce a score measuring disagreement between the semantic estimates derived from this representation. The formulation supports unconditional and text-conditional diffusion models, while requiring no prior knowledge or data of sensitive attribute. Experiments confirm that our method improves quality and diversity scores at comparable fairness levels.



Figure 1: Generations examples on CelebA-HQ. *Top*: the original model, 6 random draws of women; two of them appear with artifacts, the third looking away from the camera and the fourth turning her head. *Middle*: guidance with the fairness term. Three columns turn into men and the faces sharpen, but the pose collapses, every face now frontal and centered. *Bottom*: guidance with the fairness and diversity terms. The woman in the second column looks older, the man in the fourth turns his head away and gains a mustache.

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.

Table 1: Positioning among debiasing methods for diffusion models. A method may need attribute supervision (e.g. labels, reference images), may not accept an arbitrary target proportion, may not have been applied to both unconditional and text-conditioned models, and may retrain the generator. (∗) Score Guidance also presents an alternate supervision-free version, but with significantly diminished performance compared to the supervised version, which is the one we compare with.

	no attribute supervision	arbitrary proportion	arbitrary conditioning	no retraining	attribute
Fair Diffusion (Friedrich et al., 2025)	✓	✓	✗	✓	text
ITI-GEN (Zhang et al., 2023)	✗	✓	✗	✓	ref. images
Self-debiasing (Vardhana et al., 2026)	✓	✗	✓	✓	discovered
Latent editing (Kwon et al., 2023)	✓	✗	✓	✓	text, training per attribute
Latent direction (Li et al., 2024)	✓	✓	✗	✓	text, training per attribute
FairGen (Jiang et al., 2025)	✓	✓	✗	✗	text, training per attribute
Fine-tuning (Shen et al., 2024)	✗	✓	✗	✗	attribute-trained classifier
Balancing Act (Parihar et al., 2024)	✗	✓	✓	✓	labels
Inference-time alignment (Luan et al., 2026)	✗	✓	✗	✓	attribute-trained classifier
Jeffrey guidance (Razafindralambo et al., 2026)	✗	✓	✗	✓	attribute-trained classifier
DiffLens (Shi et al., 2025)	✗	✓	✓	✓	labeled probes
ICM (Zaleska et al., 2026)	✗	✓	✗	✓	labeled probes
Score Guidance (Tiwary et al., 2026)	✗/✓*	✓	✓	✓	ref. images / free text
Debias Anything (Ours)	✓	✓	✓	✓	free text

1 Introduction

Diffusion (Ho et al., 2020; Song et al., 2021b), and equivalently flow matching (Lipman et al., 2023), models are the standard unconditional generation models or text-to-image systems deployed at scale (Rombach et al., 2022; Podell et al., 2024; Wu et al., 2025). They inherit imbalances of their training data and often exaggerate them (Perera & Patel, 2023; Bianchi et al., 2023; Luccioni et al., 2023), and the guidance they sample with amplifies the imbalance further as its scale grows (Kim et al., 2026).

Debiasing them includes choosing how often a given property, the *sensitive attribute*, should appear in the generated images, and making the model comply (Choi et al., 2020; Teo et al., 2023). Two problems make debiasing diffusion models a hard problem. Firstly, existing methods have limited applicability with various restrictions: either they must rely on sensitive attribute supervision (Parihar et al., 2024; Shi et al., 2025; Zaleska et al., 2026), require model finetuning (Shen et al., 2024), are not adaptable to arbitrary attributes (Vardhana et al., 2026), or are restricted to text-to-image models (Friedrich et al., 2025; Zhang et al., 2023). Secondly, guidance, used to steer a diffusion model towards fair representation as predicted by an attribute classifier, harms output diversity: the stronger the push, the more the samples cluster around a few typical images and the rarer ones vanish (Ho & Salimans, 2022; Sadat et al., 2024; Sehwag et al., 2022). Methods that restore diversity exist (Um & Ye, 2024; Corso et al., 2024), but they do not aim for specific attribute fairness.

We propose Debias Anything, a guidance that sets the proportion of any multi-class attribute one can describe in words, with no labels and no generator retraining, and keeps the diversity of the model. The attribute is described independently of the diffusion model or prompt as short sentences, for instance *a person with eyeglasses* and *a person without eyeglasses*. An image-text encoder places sentences and images in one shared space, so sentences pairs define a fairness direction which can be used to steer generated images towards fairness. To follow that direction while images are still noisy, we train a small adapter, once per model and for no attribute in particular, that predicts the encoder's expected embedding of the final image from the inner activations of the frozen generator. Each sampling step then moves every image along a direction towards an assigned attribute class. A second term, computed from the same adapter, pushes the images back towards the rare regions the first term empties.

We compare Debias Anything with published baselines on unconditional face models and a text-to-image model, demonstrating its versatility. Despite the strong assumptions made by the baselines, especially regarding attribute supervision, Debias Anything is better than or comparable to them. On CelebA-HQ, it reaches the best or near-best fairness score on every considered attribute, with an improved image quality. On Stable Diffusion 1.5, it substantially improves fairness across all

considered attributes, with an image quality among the best. Finally, we show it can also target attributes for which no label exists, confirming its attribute independence

2 Related work

Debiasing diffusion models. Table 1 compares existing methods. Almost all of them share at least one restriction among four. First, many need attribute supervision like reference images for ITI-GEN (Zhang et al., 2023) and labeled images for DiffLens (Shi et al., 2025) and ICM (Zaleska et al., 2026). Second, some cannot debias towards arbitrary attribute proportions (Vardhana et al., 2026). Third, most rely on text conditioning in the base model, which restricts them to text-to-image models (Friedrich et al., 2025; Li et al., 2024; Jiang et al., 2025). Fourth, a few of them necessitate retraining or fine-tuning the base model (Shen et al., 2024). Our method is free from all these assumptions.

The closest methods to ours steer a whole batch towards a target distribution at sampling time. Balancing Act (Parihar et al., 2024) does so through a predictor trained on labeled U-Net activations, and every image receives the gradient of a batch-level distance, but remains a supervised method. Score Guidance (Tiwary et al., 2026) needs exemplar images of each value, except in a text version that needs no supervision but is less effective in most of the reported experiments. Both versions score the attribute in pixel space, whereas our score is read in the encoder's embedding space (Section 4.1), and both adopt a greedy assignment of images to values early in sampling, whereas ours is optimally recomputed at every step (Proposition 1).

Restoring diversity. Guidance narrows the distribution it samples (Ho & Salimans, 2022) and amplifies attribute imbalance as its scale grows (Kim et al., 2026). Sampling-time methods recover the tail by pushing towards low-density regions of a feature space (Sehwag et al., 2022), repelling the samples of a batch in pixel or feature space (Corso et al., 2024; Bhat et al., 2026), maximising their kernel entropy (Jalali et al., 2025), boosting the initial noise (Um et al., 2025), annealing the conditioning (Sadat et al., 2024), measuring rarity against a prior independent of the generator (Park & Um, 2026), or scoring rarity by how much re-denoising deforms the Tweedie estimate, with a trained classifier (Um et al., 2024) or without one (SGMS; Um & Ye, 2024). None of them constrains a sensitive attribute. Our rarity term transposes SGMS into the encoder's space through the same adapter for efficiency purposes, and it is added to the fairness term so that the tail is recovered at the target proportion.

Inferring the attribute during sampling. Classifier guidance (Dhariwal & Nichol, 2021) trains a classifier on noisy images for each attribute; training-free guidance (Chung et al., 2023; Yu et al., 2023; Bansal et al., 2024; Ye et al., 2024), like the text version of Score Guidance, applies an off-the-shelf model to the denoised image, a blurred average, early in sampling; readout guidance (Luo et al., 2024) trains light heads on inner features with task labels. Instead, our method is unsupervised and only applies an encoder on clean, non-blurry images.

The closest method to ours is that of unsupervised self-debiasing (Vardhana et al., 2026), the only debiasing method that is given no attribute at all. For each prompt, it trains a projector from the h-space of Kwon et al. (2023), to an embedding of the image that the sampler produces, then clusters the projections and steers each batch towards balanced clusters. The attribute is therefore discovered rather than named, and a target can only be set over the clusters found. Our adapter is instead trained once per generator rather than per prompt, and is simulation-free: it regresses the embedding of a clean image from noised versions of that image rather than from the sampling trajectory that produced it (Section 4.1). Sentences then name the attribute, and our batch assignment can set any proportion.

3 Preliminaries

Generative model. We focus in this paper on diffusion models, even though our method can easily be adapted to standard flow matching methods. We follow the EDM parameterisation (Karras et al., 2022) without loss of generality; other diffusion models such as DDPM (Ho et al., 2020) and VP / VE SDEs (Song et al., 2021b) may be written in the same framework.

A noisy image is $x_\sigma = x_0 + \sigma\varepsilon \sim p_\sigma$ with $x_0 \sim p_{\text{data}}$ and $\varepsilon \sim \mathcal{N}(0, I)$, and a network $D_\theta(x;\sigma)$ is pre-trained to approximate the posterior mean $\mathbb{E}[x_0 \mid x_\sigma = x]$. We write $\hat{x}_0(x) = D_\theta(x;\sigma)$ and call

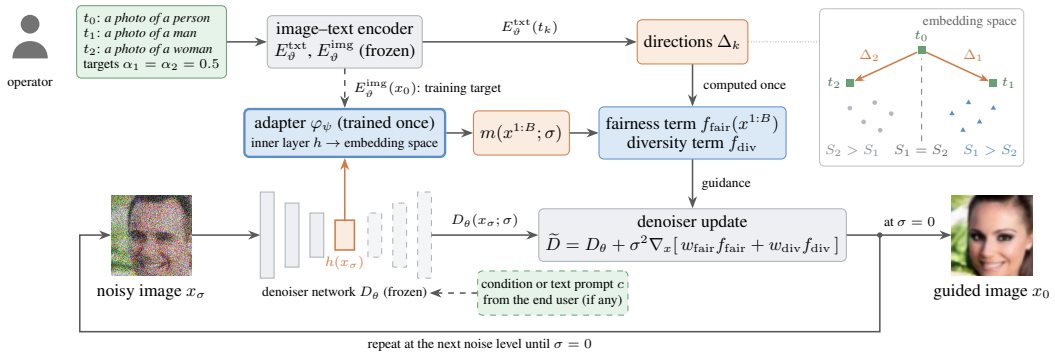

Figure 2: One guided sampling step. The prompts t_k are set once and encoded into directions Δ_k. At each step, the adapter φ maps an inner layer $h(x_\sigma)$ of the frozen denoiser to $m(x_\sigma;\sigma)$ for every image of the batch, where the fairness and diversity terms are computed.

it the *Tweedie estimate*, the model's approximation (as an average) of the denoised image given the current noisy state. Tweedie's identity (Efron, 2011), $\nabla_x \log p_\sigma(x) = (D_\theta(x;\sigma) - x)/\sigma^2$, turns the denoiser into the score of p_σ, and sampling integrates from $\sigma_{\max}$ down to $\sigma = 0$, for any $\beta > 0$:

$$\mathrm{d}x_\sigma = -(1+\beta)\sigma\,\nabla \log p_\sigma(x_\sigma)\,\mathrm{d}\sigma + \sqrt{2\beta\sigma}\,\mathrm{d}W_\sigma = (1+\beta)\frac{x - D_\theta(x_\sigma;\sigma)}{\sigma}\,\mathrm{d}\sigma + \sqrt{2\beta\sigma}\,\mathrm{d}W_\sigma. \quad (1)$$

Guidance may be used to steer the sampler towards images with a desired property without retraining the model. One chooses a scalar function f of the image and of the noise level, the *guidance term*, and a weight w, and samples from $p_\sigma(x)\,e^{w f(x;\sigma)}$ instead of p_σ. By Tweedie's identity this only changes the denoiser (Dhariwal & Nichol, 2021; Chung et al., 2023):

$$\widetilde{D}(x;\sigma) \;=\; D_\theta(x;\sigma) \;+\; w\,\sigma^2\,\nabla_x f(x;\sigma). \quad (2)$$

Classifier guidance (Dhariwal & Nichol, 2021) is the case $f = \log p_\sigma(y \mid x)$ for a class y, estimated with a classifier trained on noisy images, and classifier-free guidance (Ho & Salimans, 2022) obtains that classifier from a conditional model by Bayes' rule. The solver is untouched, and several guidance terms add inside one gradient.

Image-text embeddings. We will use a frozen image-text encoder E_ϑ^{img} / E_ϑ^{txt} mapping respectively images and texts to a shared embedding space on the unit sphere $\mathbb{S}^{d-1}$, so that matching pairs are close. Any choice of image-text encoder, such as CLIP (Radford et al., 2021) and its successors, is admissible. In this paper, we choose SigLIP 2 (Zhai et al., 2023; Tschannen et al., 2025).

4 Method

We describe here our guidance method for debiasing a pre-trained diffusion model D_θ of any kind w.r.t. any sensitive attribute, expressed in free text, with arbitrary proportions and preserved output quality & diversity, while requiring no extra annotated data and no diffusion model fine-tuning.

For this, we combine two guidance terms, one for fairness and one for diversity, and compute both from a single semantic representation of the image being generated. The proposed method therefore contains three pieces (Figure 2): an adapter that makes a clean and easy-to-use embedding space of a pre-trained image-text encoder available at every noise level of a frozen generator (Section 4.1), attribute guidance directions from pairs of prompts with the batch-level fairness term (Section 4.2), and a diversity term defined in that space (Section 4.3).

4.1 A clean simulation-free semantic embedding of the noisy image

To build guidance terms without labeled data, we use the pre-trained image-text encoder E_ϑ^{img} / E_ϑ^{txt}, which turns any property described in words into a direction for guidance in image space. However, a guidance term f needs to be evaluated on noisy images x_σ (cf. Section 3), on which the image-text

encoder was never trained on. This raises the question of how to obtain the encoder's embedding $m(x_\sigma; \sigma)$ of the final image from the noisy one.

Choice of encoder input. Two options have been considered in the literature.

The simplest option, used by training-free guidance (Chung et al., 2023; Yu et al., 2023; Bansal et al., 2024), is to denoise x_σ in one step to the Tweedie estimate $\hat{x}_0 \approx \mathbb{E}[x_0 \mid x_\sigma]$ and to embed that image to $m(x_\sigma; \sigma) = E_\vartheta^{\text{img}}(\hat{x}_0)$. In this case, E_ϑ^{img} is applied to an average of images, necessarily blurry for high σs and therefore out of distribution (OOD) for E_ϑ^{img}, making it unreliable. This distribution shift is particularly strong in the first phases of the trajectory when the coarse content of a sample, and with it an attribute such as gender, is decided (Raya & Ambrogioni, 2023; Biroli et al., 2024; Gagneux et al., 2025), i.e. when guidance needs to act. We refer to Section D.1 for more details.

Another option explored by Vardhana et al. (2026) is to embed the generated sample obtained after integrating Equation (1). To avoid the simulation cost at each guidance step, this necessitates the training of an adapter network trained to map x_σ or some proxy like internal activations of $D_\theta(x; \sigma)$ to the true embedding $m(x_\sigma; \sigma) = E_\vartheta^{\text{img}}(x_\sigma + \int_\sigma^0 [\text{Equation (1)}])$. This solves the issue of blurry OOD inputs for the encoder, but ties guidance to the choice of generating process, including discretization errors and other guidance terms.

Instead, we propose to embed $x_0 = x_\sigma - \sigma\varepsilon$. When learning an adapter to regress this target, then:

$$m(x_\sigma; \sigma) \;=\; \mathbb{E}[E_\vartheta^{\text{img}}(x_0) \mid x_\sigma]. \tag{3}$$

It combines the advantages of the two previous methods, being simulation-free and computed in-distribution, at the cost of averaging in the embedding space. As shown in Section 4.2, this is a mild approximation as, due to its semantic nature, averaging in the embedding space better preserves attribute information than in the pixel space.

Our adapter. To approximate m, we also use an adapter taking as input the denoiser bottleneck activation $h(x_\sigma)$, which Kwon et al. (2023) identified as a semantic latent space (the h-space), and which the denoiser computes in any case during the generation process. In a U-Net (Ronneberger et al., 2015), h can be the output of the mid-block; in a diffusion transformer (Peebles & Xie, 2023) h can then be the tokens of an intermediate block (Yu et al., 2025). The adapter network φ_ψ, with parameters ψ, maps this activation to a unit vector $\varphi_\psi\big(h(x_\sigma);\, \sigma\big) \;\in\; \mathbb{S}^{d-1}$ and is trained with a cosine loss on pairs $(x_\sigma, E_\vartheta^{\text{img}}(x_0))$ at random noise levels, whose minimizer is the direction of m:

$$\mathcal{L}(\psi) \;=\; 1 - \mathbb{E}_{x_0,\sigma,\varepsilon} \left\langle \varphi_\psi(h(x_\sigma); \sigma),\; E_\vartheta^{\text{img}}(x_0) \right\rangle, \tag{4}$$

Only ψ is trained; the denoiser and the encoder stay frozen and no attribute label enters. The adapter is trained once per generator and serves every attribute afterwards (Section C).

4.2 Fairness guidance from text prompts

For a K-valued sensitive attribute, we are provided with a description of each value $k \in [\![1, K]\!]$ by a sentence t_k, and a neutral sentence t_0: for age, for instance, t_0 = *a photo of a person*, t_1 = *a photo of a young person*, t_2 = *a photo of a middle-aged person*, t_3 = *a photo of an old person*. We will build a fairness guidance term using only these as information about the attribute.

Zero-shot classifier. Thanks to the image-text encoder, a zero-shot classifier (Radford et al., 2021) can be built by comparing each attribute to the neutral prompt. Each value k defines a direction Δ_k in the joint embedding space, and projecting an image embedding from Section 4.1 on Δ_k gives a logit S_k that increases the closer the image is to t_k than to t_0 (Figure 2, top right):

$$\Delta_k = \frac{E_\vartheta^{\text{txt}}(t_k) - E_\vartheta^{\text{txt}}(t_0)}{\|E_\vartheta^{\text{txt}}(t_k) - E_\vartheta^{\text{txt}}(t_0)\|}, \qquad S_k(x_\sigma; \sigma) = \big\langle m(x_\sigma; \sigma),\, \Delta_k \big\rangle \in [-1, 1]. \tag{5}$$

Directions are computed once before sampling. The difference of embeddings keeps what distinguishes the two sentences and cancels what they share (Bolukbasi et al., 2016; Lyu et al., 2023). In practice, we compute the logits with the adapter output $\varphi_\psi(h(x_\sigma); \sigma)$ in place of $m(x_\sigma; \sigma)$. Note that the same adapter serves every attribute.

At $\sigma = 0$, $m(x_0; 0) = E^{\text{img}}_\vartheta(x_0)$ so it is the true embedding of the clean image. This defines a clean-image one-versus-rest classifier as $p(Y_k = 1 \mid x_0) \propto \mathrm{e}^{wS_k(x_0;0)}$, with $w > 0$ the guidance strength and $Y_k = \delta_{y=k}$. By linearity, $S_k(x_\sigma; \sigma) = \mathbb{E}[S_k(x_0; 0) \mid x_\sigma]$: the logit of the noisy image is the expected logit of the clean one. Our classifier thus averages the log of the clean classifier:

$$\bar{p}_\sigma(Y_k = 1 \mid x_\sigma) \;=\; \exp \mathbb{E}[\log p(Y_k = 1 \mid x_0) \mid x_\sigma] \;\propto\; \mathrm{e}^{wS_k(x_\sigma;\sigma)}. \tag{6}$$

It is not the ideal such classifier, which is the average of the clean classifier (Section 3):

$$p_\sigma(Y_k = 1 \mid x_\sigma) \;=\; \mathbb{E}[p(Y_k = 1 \mid x_0) \mid x_\sigma] \;\propto\; \mathbb{E}[\mathrm{e}^{wS_k(x_0;0)} \mid x_\sigma], \tag{7}$$

but which is intractable in a zero-shot, attribute-independent setting. The gap is as follows.

Lemma 1 (expected logit)**.** *With* $V_k(x) = \mathrm{Var}[S_k(x_0; 0) \mid x_\sigma]$,

$$e^{-2w}\,\tfrac{w^2}{2}\,V_k(x) \;\le\; \log p_\sigma(Y_k = 1 \mid x_\sigma) - \log \bar{p}_\sigma(Y_k = 1 \mid x_\sigma) \;\le\; e^{2w}\,\tfrac{w^2}{2}\,V_k(x). \tag{8}$$

The proof is in Section A.1. The gap between the ideal classifier and ours therefore vanishes when V_k approaches 0, that is, once the model has decided the attribute.

Guidance. Given target proportions $\alpha_1, \ldots, \alpha_K \ge 0$ summing to one and a batch of B images, the objective is that a fraction α_k of the batch carries attribute value k. The objective is enforced by batch-wise assignment: each image i is assigned a value $\pi(i)$ and guided towards it with our classifier of Equation (6). With integer quotas $B_k \approx \alpha_k B$ summing to B, let Π be the set of maps $\pi\colon \{1, \ldots, B\} \to \{1, \ldots, K\}$ that assign exactly B_k images to value k. The assignment, solved exactly by the Hungarian algorithm (Kuhn, 1955), maximises the total logit of the batch, so that each image is pushed, as far as the quotas allow, towards the value it is already closest to.

$$\pi^\star \;=\; \arg\max_{\pi \in \Pi} \sum_{i=1}^{B} a_{i\pi(i)}, \qquad f_{\text{fair}}(x^{1:B}) \;=\; \sum_{i=1}^{B} a_{i\pi^\star(i)}, \qquad a_{ik} = S_k(x_i; \sigma). \tag{9}$$

Up to a constant, $w\, f_{\text{fair}} = \sum_i \log \bar{p}_\sigma(y = \pi^\star(i) \mid x_i)$ is the log of our classifier summed over the batch, so guiding with f_{fair} is guiding each image towards its assigned value.

Proposition 1. *Let* $A = (a_{ik}) \in \mathbb{R}^{B \times K}$*. Then (i)* $f_{\text{fair}} = \max_{\pi \in \Pi} \sum_i a_{i\pi(i)}$ *is convex and piecewise linear in* A*, and wherever the maximiser* $\pi^\star$ *is unique it is differentiable with* $\partial f_{\text{fair}} / \partial a_{ik} = \mathbf{1}[\pi^\star(i) = k]$*; (ii) the guided denoiser of Equation (2) with* $f = f_{\text{fair}}$ *is then:*

$$\widetilde{D}(x_i; \sigma) \;=\; D_\theta(x_i; \sigma) \;+\; w_{\text{fair}}\, \sigma^2\, \nabla_{x_i} f_{\text{fair}}(x^{1:B}) \;=\; D_\theta(x_i; \sigma) \;+\; w_{\text{fair}}\, \sigma^2\, \nabla_{x_i} a_{i\pi^\star(i)}, \tag{10}$$

so each image is pushed towards the value it is assigned to.

The proof is in Section A.1. By Equation (10), pushing each image towards its assigned value is guidance with f_{fair}, and recomputing the assignment at every step recomputes its gradient. Because the assignment couples the B trajectories, Equation (9) is a joint guidance term over the batch in the sense of Corso et al. (2024). Still, unlike Parihar et al. (2024), where every image receives the gradient of a single batch-level distance between the predicted and the target attribute distributions, each image is guided by its own logit, towards the value it is assigned to and therefore compatible with.

Binary vs multi-class. A multi-class zero-shot classifier on the same scores would lead to the same method with fewer guarantees: the assignment is unchanged, and the guidance term gains the normalisation $-\log \sum_j e^{wS_j(x_0;0)}$; this term is not linear in the embedding, so Lemma 1 would no longer hold exactly. For a binary attribute, the neutral sentence is not needed and the assignment may be reduced to ranking the batch along a single direction (Section A.2).

4.3 The diversity term and combined guidance

Diversity guidance. We start from SGMS (Um & Ye, 2024) and move its construction into the encoder's space, available from the adapter at no extra cost. SGMS scores diversity by re-noising the Tweedie estimate to σ_s, denoising it again, and measuring how far the image moved:

$$\hat{x}_s \;=\; \mathrm{sg}[\hat{x}_0(x)] + \sigma_s \varepsilon, \qquad f_{\text{SGMS}}(x) \;=\; d_{\text{pix}}\big(\hat{x}_0(x),\, D_\theta(\hat{x}_s; \sigma_s)\big), \tag{11}$$

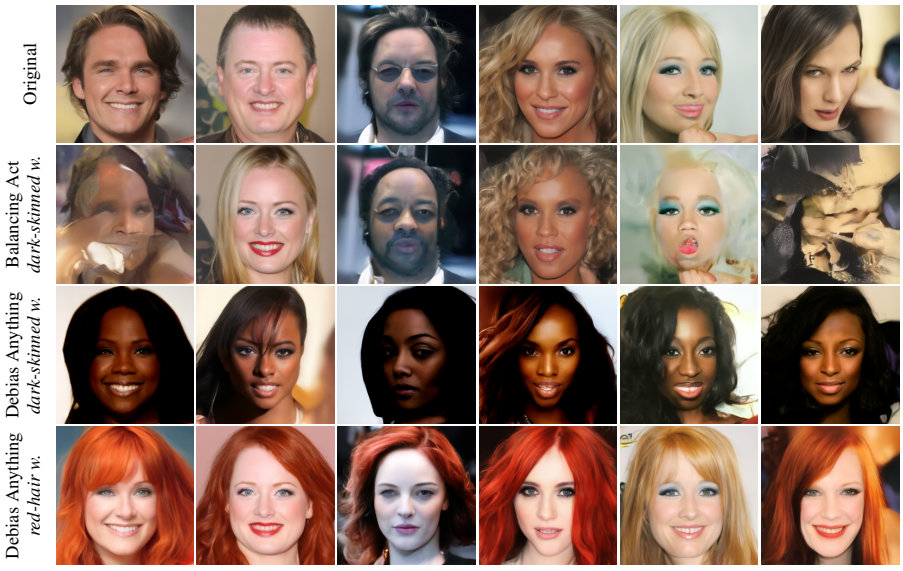

Figure 3: CelebA-HQ, P2 backbone, six shared seeds per column. Top: unguided samples. Second row: Balancing Act steered towards a dark-skinned woman with its labelled predictor. Third row: Debias Anything on the same target, given as two sentences. Bottom: Debias Anything on a second attribute with no CelebA-HQ label, red hair, from the same seeds.

with d_{pix} an LPIPS (Zhang et al., 2018) or ℓ^2 distance and $\text{sg}[\cdot]$ a stop-gradient.

Re-denoising pulls an image towards the modes of the model, so an image already on a mode barely moves while an image in the tail does. Measured in pixel space, however, the displacement counts pose, background and lighting rather than semantics. We propose instead a score that measures the disagreement between two estimates of the embedding of the image being generated:

$$f_{\text{div}}(x) \;=\; 1 - \big\langle m(x_\sigma;\sigma),\; m(\hat{x}_s;\sigma_s)\big\rangle \;\approx\; 1 - \big\langle \varphi_\psi(h(x_\sigma);\sigma),\; \varphi_\psi(h(\hat{x}_s);\sigma_s)\big\rangle. \tag{12}$$

Section A.3 describes a variant that takes the second estimate from the encoder instead.

Combined guidance. The diversity term is added to the fairness guidance of Equation (10):

$$\widetilde{D}(x;\sigma) \;=\; D_\theta(x;\sigma) \;+\; \sigma^2\,\nabla_x\big[\,w_{\text{fair}}\,f_{\text{fair}}(x) \;+\; w_{\text{div}}\,f_{\text{div}}(x)\,\big]\,. \tag{13}$$

Compatibility with any conditioning and other guidance The presented method is independent of how the model is conditioned, applies to unconditional and text-conditional models alike, and composes with any other guidance already in use, since Equation (13) would simply gain a term. When classifier-free guidance (Ho & Salimans, 2022) is used, the adapter only reads from prompt-conditioned branch.

5 Experiments

We compare in this section the proposed method, Debias Anything, with published, more restrictive (Table 1), baselines at matched fairness, on unconditional and a text-to-image models.

Setup. We test Debias Anything on three generators: an EDM (Karras et al., 2022) trained on CelebA (Liu et al., 2015) at 64×64 (deferred to Section D), the P2 model of Choi et al. (2022) trained on CelebA-HQ (Karras et al., 2018) at 256×256, and Stable Diffusion 1.5 (Rombach et al., 2022). On CelebA the attributes are: *gender* and *eyeglasses*, binary with a target of one half; on CelebA-HQ: *gender*, *race* and *eyeglasses*; on Stable Diffusion: *gender*, *age* and *race*, on occupation prompts that do not mention these attributes. The target proportion is uniform for each of these attribute possible values. Fairness is measured by the discrepancy FD between the achieved and target proportions,

Table 2: CelebA-HQ, P2 backbone, protocol of Balancing Act (10 000 images, one seed); all rows are our runs. FD is CLEAM-corrected except on race, which has no label to correct against. Guided rows are shaded from best (dark, bold) to worst (light); the bracket marks the rows matched on discrepancy.

	Gender		Race		Eyeglasses	
Method	FD↓	FID↓	FD↓	FID↓	FD↓	FID↓
Original	0.257	54.9	0.606	62.8	0.666	73.9
Score Guidance (Tiwary et al., 2026)	**0.001**	64.7	0.012	65.4	0.075	66.8
Balancing Act (Parihar et al., 2024)	0.043	50.1	0.182	**52.0**	0.187	50.5
Debias Anything, quality setting	0.057	50.7	0.159	54.2	0.183	45.4
+ diversity term	0.050	**48.7**	0.147	52.7	0.184	**45.2**
Debias Anything, fairness setting	0.008	50.7	**0.008**	58.2	**0.065**	59.8
+ diversity term	0.008	50.0	**0.008**	58.0	**0.065**	59.8

CLEAM-corrected where possible (Teo et al., 2023). Quality is assessed w.r.t. a debiased reference set with FID (Heusel et al., 2017) (and also MIND from Berthet et al. (2026) in the appendix), by precision and recall (Kynkäänniemi et al., 2019), and diversity with the Vendi score (Friedman & Dieng, 2023), a diversity measure counting the effective number of distinct images without a reference set. CelebA-HQ and Stable Diffusion comparisons follow the protocols of Balancing Act (Parihar et al., 2024) and DiffLens (Shi et al., 2025), so that their published numbers can be reused. Prompts, baselines, metric definitions, weights, and other details are included in Section B.

5.1 Comparison with baselines

CelebA-HQ. Every number in Table 2 comes from our own runs of the three methods under the protocol of Balancing Act (Parihar et al., 2024). For Score Guidance (Tiwary et al., 2026), the numbers we obtain differ from those reported in the original paper. We reproduced its attribute classifier, and we believe the gap comes from the reference set behind FID, which might not be the same as Balancing Act; the Score Guidance version we run is the stronger of the two suggested in the paper. Table 2 reports our method at two weights, *fairness setting* and *quality setting*, since fairness guidance might reduce generation quality.

At the *fairness setting* weight the discrepancy is the lowest in race and eyeglasses and, in gender, second best behind **Score Guidance, which compensates for its low discrepancy with worse** FID: at similar discrepancy, our FID is 7 to 14 points below the best version of the concurrent method. At the *quality setting* weight, we set the discrepancy to approach Balancing Act's by tweaking w_{fair}, so that only FID differs between the two bracketed rows: it is five points lower for us on eyeglasses, and 2.2 and 0.6 points higher on race and gender. Adding the diversity term at the same weight lowers discrepancy and FID together on gender and race, and takes FID below Balancing Act on gender and eyeglasses. So the diversity term closes the FID gap that the fairness term leaves, still without any label. Figure 3 shows additional advantage of the proposed method: the same model steered towards a target *red-hair woman* that CelebA-HQ does not have a label for. Thus, **Debias Anything generates the image proportion closer to the target than a classifier trained on labeled images**, and with the diversity term **it gives a lower or comparable** FID **at comparable discrepancy.**

Stable Diffusion. The baseline rows of Table 3 are the numbers the DiffLens and ICM papers report under the setting of Shi et al. (2025) and Zaleska et al. (2026). We re-ran the unguided model under the same setup and evaluation and recovered the corresponding numbers reported by DiffLens, so we directly report all their published numbers. **Every baseline knows the attribute in advance**, through labels, a classifier or a training run for that attribute. **Ours does not**: the adapter is trained once, on prompts about jobs that say nothing about gender, age or race, and the attribute enters only at sampling, as two sentences, with one weight for all three attributes. Figure 11 shows doctors and firefighters before and after guidance on shared seeds.

With much less information about the attributes, our method brings the discrepancy from 0.56–0.75 down to 0.20 on all three attributes, and image quality stays close to the unguided model's: FID is second on gender and age, behind DiffLens, and first on race. Section E reports the detailed explanation on the residual discrepancy. Some methods go lower on one attribute, always with

Table 3: Stable Diffusion 1.5, protocol of Shi et al. (2025), whose reported numbers fill the baseline rows bar ICM. All baselines incorporate attribute knowledge (labeled data, attribute-specific training); ours only gets two sentences. Guided rows are shaded from best (dark, bold) to worst (light).

	Gender		Age		Race	
Method	FD↓	FID↓	FD↓	FID↓	FD↓	FID↓
Original	0.564	120.1	0.752	120.1	0.558	120.1
Attribute-specific						
Latent editing (Kwon et al., 2023)	0.408	166.1	0.682	200.9	0.524	153.1
h-distribution (Parihar et al., 2024)	0.222	151.7	0.506	147.7	0.544	126.9
Latent direction (Li et al., 2024)	0.305	129.4	0.052	113.8	**0.175**	128.3
Fine-tuning (Shen et al., 2024)	0.050	161.5	0.746	161.5	0.198	161.5
DiffLens (Shi et al., 2025)	**0.046**	**112.8**	**0.049**	**99.2**	0.401	119.9
ICM (Zaleska et al., 2026)	0.087	122.1	0.133	114.9	0.266	116.9
Attribute-unsupervised						
Debias Anything	0.200	113.8	0.206	110.6	0.195	**114.7**

attribute-specific training or supervision, and none holds across the three; only ours is below 0.21 on all three. Where DiffLens is ahead, on gender and age, it is advantaged by attribute knowledge: labeled images locate the features that carry the attribute inside the network, and it edits those features directly. Thus, in the setting of Shi et al. (2025), **our unsupervised results are better than or comparable to those of methods that leverage labeled data**.

5.2 Ablations on CelebA

Results are included in Section D; selected findings are the following. **Reading the score through the adapter is what makes the rare attribute reachable.** At equal weights on eyeglasses it drives the discrepancy from 0.573 to 0.32 while lowering MIND from 20.51 to 8.9, whereas the same score read from the encoder on the denoised image never passes 0.489 and raises MIND (Table 5 and Figure 6). **Every baseline diversity term improves on the fairness term alone, but ours lowers the discrepancy most** (Table 8). On gender all four lower MIND from 2.52 to about 2.0, and our term lowers the discrepancy from 0.031 to 0.006; on eyeglasses they lower the discrepancy and raise Vendi, while MIND moves within one standard deviation and FID rises by up to 1.2. **Our method is resilient to changing the wording of attribute reference sentences**, as shown in Figure 10.

6 Conclusion

We introduce Debias Anything, a generative model guidance that sets the proportions of a sensitive attribute described in words, in unconditional and text-to-image diffusion models, with no labels and no generator retraining. A generic adapter gives access to the embedding space of an image-text encoder; from it, a batch-level fairness term enforces the target proportions and a diversity term counteracts the loss of diversity that guidance causes. Although it receives very little attribute information, Debias Anything is better than or comparable to methods that rely on attribute supervision.

Limitations. The method has the limits of its parts. The reconstituted attribute relies on the image-text embedding, with the encoder's own biases. Guidance can only push towards images the model can produce: on Stable Diffusion, male nurses are rare and remain hard to reach. Steering one attribute moves others, as eyeglasses moved gender on CelebA, so attribute combinations should be considered. The proportion is enforced on a batch and not at sample level. The two guidance weights have a narrow useful range that we select by sweeps as detailed in the appendix. Our evidence is on human faces only, at uniform target proportions, with single seeds on CelebA-HQ and Stable Diffusion.

Future work. We believe that the work has several future directions. The adapter itself could serve to audit a model's biases at a fraction of the cost of generating and classifying its samples, or to stop a biased trajectory early. The same recipe, one adapter from an inner layer to an embedding shared with text, carries over to video and audio diffusion models, whose encoders exist. Finally, the proportions we enforce are measured by classifiers; putting the correction of their errors inside the loop, rather than after it, would let the sampler aim at the true proportion instead of the counted one.

AI USE STATEMENT

In this work, we used generative AI tools to help derive our theoretical analysis and as writing and coding assistants.

We have not used generative AI assistants to generate synthetic data, to develop theoretical models or conceptual frameworks, to propose or refine hypotheses, to design or give feedback on the methodology or the experiments, to implement the method, or to interpret the results; translation, dataset cleaning and reformatting, and qualitative or thematic data analysis are not applicable to this work.

We used generative AI tools for the following tasks: to draft parts of the paper and edit it for readability, to assist in theorem proving, to create and edit software code for evaluation and plotting, to produce the LaTeX of tables and figures, to identify and summarize relevant literature, and to format references. We have reviewed all AI-assisted work. Every AI-drafted sentence was checked by the authors against the tables and the cited papers; every number in the paper comes from our own runs or from the cited papers; AI-edited code was read and tested by the authors before use. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work makes a generator represent groups of people in proportions an operator chooses, and the mechanism that sets a proportion to one half can set it to zero. Any method that controls an attribute offers the same lever, including the label-based methods we compare with. Furthermore, the method makes a chosen distribution enforceable and says nothing about which distribution is fair.

The attributes we steer (gender, race, age, skin tone and eyeglasses) are named in text and assessed by classifiers we trained and external tools like FairFace or FID, that carry biases of their own and may impose fixed categories, binary gender and four races among some of them. Our discrepancy numbers are only as good as the used classifiers.

REPRODUCIBILITY STATEMENT

Section B lists the generators, attributes and prompts, the baselines and what each is given, the metric definitions, and the protocol of every comparison. Section C gives the adapter's layers, loss, and training data, Section D every ablation table with seeds and standard deviations, and Section A.1 the proofs. Code and adapter checkpoints will be open-sourced with the paper release.

REFERENCES

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *International Conference on Learning Representations*, 2024.

Quentin Berthet, Yu-Han Wu, Clément Crepy, Romuald Elie, Klaus Greff, and Michaël E. Sander. MIND: Monge Inception distance for generative models evaluation. *arXiv preprint arXiv:2605.06797*, 2026.

Pradhaan S. Bhat, Rishubh Parihar, Abhijnya Bhat, and R. Venkatesh Babu. Don't settle at the mode! Mitigating diversity collapse in pretrained flow models via feature self-guidance. In Paolo Favaro, Zuzana Kukelova, Atsuto Maki, Anna Rohrbach, Konrad Schindler, and Federico Tombari (eds.), *The European Conference on Computer Vision (ECCV)*, pp. 334–352, Cham, Switzerland, September 2026. Springer Nature Switzerland.

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 *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1493–1504, New York, New York, United States of America, 2023. Association for Computing Machinery.

Giulio Biroli, Tony Bonnaire, Valentin de Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 15(1):9957, 2024.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29, pp. 4349–4357. Curran Associates, Inc., 2016.

Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11472–11481, June 2022.

Kristy Choi, Aditya Grover, Trisha Singh, Rui Shu, and Stefano Ermon. Fair generative modeling via weak supervision. In Hal Daumé, III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1887–1898. PMLR, July 2020.

Min Jin Chong and David Forsyth. Effectively unbiased FID and Inception score and where to find them. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6069–6078, June 2020.

Hyungjin Chung, Jeongsol Kim, Michael T. Mccann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations*, 2023.

Gabriele Corso, Yilun Xu, Valentin de Bortoli, Regina Barzilay, and Tommi Jaakkola. Particle guidance: non-I.I.D. diverse sampling with diffusion models. In *International Conference on Learning Representations*, 2024.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann Dauphin, Percy S. Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794. Curran Associates, Inc., 2021.

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

Dan Friedman and Adji Bousso Dieng. The Vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023.

Felix Friedrich, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Patrick Schramowski, Sasha Luccioni, and Kristian Kersting. Auditing and instructing text-to-image generation models on fairness. *AI and Ethics*, 5(3):2103–2123, 2025.

Anne Gagneux, Ségolène Martin, Rémi Gribonval, and Mathurin Massias. The generation phases of flow matching: a denoising perspective. *arXiv preprint arXiv:2510.24830*, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, June 2016.

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 Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna Wallach, Rob Fergus, Swaminathan Venkata Narayana Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30, pp. 6626–6637. Curran Associates, Inc., 2017.

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020.

Mohammad Jalali, Haoyu Lei, Amin Gohari, and Farzan Farnia. SPARKE: Scalable prompt-aware diversity and novelty guidance in diffusion models via RKE score. In Danielle Belgrave, Cheng Zhang, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, and Nancy Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 119943–119980. Curran Associates, Inc., 2025.

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

Kimmo Kärkkäinen and Jungseock Joo. FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1548–1558, January 2021.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

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

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In Sanmi Koyejo, Shakir Mohamed, Alekh Agarwal, Danielle Belgrave, Kyunghyun Cho, and Alice Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 26565–26577. Curran Associates, Inc., 2022.

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

Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10(60): 1755–1758, 2009.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Harold W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1–2):83–97, 1955.

Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. In *International Conference on Learning Representations*, 2023.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In Hanna Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32, pp. 3927–3936. Curran Associates, Inc., 2019.

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

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 3730–3738, December 2015.

Hao Luan, See-Kiong Ng, and Chun Kai Ling. Inference-time attribute distribution alignment for unconditional diffusion. *arXiv preprint arXiv:2605.07456*, 2026.

Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Evaluating societal representations in diffusion models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 56338–56351. Curran Associates, Inc., 2023.

Grace Luo, Trevor Darrell, Oliver Wang, Dan B. Goldman, and Aleksander Holynski. Readout guidance: Learning control from diffusion features. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8217–8227, June 2024.

Yueming Lyu, Tianwei Lin, Fu Li, Dongliang He, Jing Dong, and Tieniu Tan. DeltaEdit: Exploring text-free training for text-driven image manipulation. *arXiv preprint arXiv:2303.06285*, 2023.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In Hal Daumé, III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7176–7185. PMLR, July 2020.

Rishubh Parihar, Abhijnya Bhat, Abhipsa Basu, Saswat Mallick, Jogendra Nath Kundu, and R. Venkatesh Babu. Balancing Act: Distribution-guided debiasing in diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6668–6678, June 2024.

Sol Park and Soobin Um. Beyond generative priors: Minority sampling with JEPA-guided diffusion. In *Proceedings of the 43rd International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, July 2026.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in GAN evaluation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11410–11420, June 2022.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In Hanna Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32, pp. 8026–8037. Curran Associates, Inc., 2019.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, October 2023.

Malsha V. Perera and Vishal M. Patel. Analyzing bias in diffusion-based face generation models. In *IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1–10, September 2023.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April 2018.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, July 2021.

Gabriel Raya and Luca Ambrogioni. Spontaneous symmetry breaking in generative diffusion models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 66377–66389. Curran Associates, Inc., 2023.

Raphaël Razafindralambo, Rémy Sun, Frédéric Precioso, Jes Frellsen, and Pierre-Alexandre Mattei. Towards more general control of diffusion models using Jeffrey guidance. *arXiv preprint arXiv:2606.13240*, 2026.

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

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi (eds.), *Medical Image Computing and Computer-Assisted Intervention — MICCAI 2015*, pp. 234–241, Cham, Switzerland, 2015. Springer International Publishing.

Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M. Weber. CADS: Unleashing the diversity of diffusion models through condition-annealed sampling. In *International Conference on Learning Representations*, 2024.

Vikash Sehwag, Caner Hazirbas, Albert Gordo, Firat Ozgenel, and Cristian Canton Ferrer. Generating high fidelity data from low-density regions using diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11492–11501, June 2022.

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

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

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8192–8202, June 2025.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021b.

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 Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 3732–3784. Curran Associates, Inc., 2023.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, June 2016.

Christopher T. H. Teo, Milad Abdollahzadeh, and Ngai-Man Cheung. On measuring fairness in generative models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 10644–10656. Curran Associates, Inc., 2023.

Piyush Tiwary, Prabhav Verma, and Prathosh AP. Debiasing diffusion models via score guidance. *Transactions on Machine Learning Research*, 2026.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Soobin Um and Jong Chul Ye. Self-guided generation of minority samples using diffusion models. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *The European Conference on Computer Vision (ECCV)*, pp. 414–430, Cham, Switzerland, October 2024. Springer Nature Switzerland.

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

Soobin Um, Beomsu Kim, and Jong Chul Ye. Boost-and-Skip: A simple guidance-free diffusion for minority generation. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 60561–60589. PMLR, July 2025.

Korada Sri Vardhana, Shrikrishna Lolla, and Soma Biswas. Fully unsupervised self-debiasing of text-to-image diffusion models. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1221–1230, March 2026.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

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 Amir Globerson, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 22370–22417. Curran Associates, Inc., 2024.

Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. FreeDoM: Training-free energy-guided conditional diffusion model. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23174–23184, October 2023.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025.

Katarzyna Zaleska, Łukasz Popek, Monika Wysoczańska, and Kamil Deja. Attention, may I have your decision? Localizing generative choices in diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30854–30863, June 2026.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, October 2023.

Cheng Zhang, Xuanbai Chen, Siqi Chai, Chen Henry Wu, Dmitry Lagun, Thabo Beeler, and Fernando De la Torre. ITI-GEN: Inclusive text-to-image generation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3969–3980, October 2023.

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

APPENDIX CONTENTS

A Additional method details **16**

A.1 Proofs of theoretical results . . . 16

A.2 Binary attributes . . . 17

A.3 Alternative diversity term . . . 17

A.4 Density baseline . . . 17

B Experimental details **17**

B.1 Generators . . . 18

B.2 Attributes, prompts and sample sizes . . . 18

B.3 Sampling . . . 18

B.4 Baselines . . . 18

B.5 Metrics . . . 19

B.6 Protocols . . . 20

C Adapter architecture and training **20**

D Ablations on CelebA **21**

D.1 Where the score is computed: h-space against pixel space . . . 21

D.2 Fairness weight . . . 22

D.3 Diversity weight . . . 24

D.4 Diversity on top of the fairness term at the selected setting . . . 26

D.5 Diversity terms alone . . . 27

D.6 Qualitative results . . . 30

E Full comparison under the DiffLens protocol **32**

A ADDITIONAL METHOD DETAILS

A.1 PROOFS OF THEORETICAL RESULTS

Proof of Lemma 1. Fix k, σ and x_σ. Write $p(Y_k = 1 \mid x_0) = c\, e^{w S_k(x_0;0)}$, with c independent of x_0, and $S = S_k(x_0;0) \in [-1,1]$ for the logit of the clean image, a function of x_0 under the posterior given x_σ. Let $\Lambda(t) = \log \mathbb{E}[e^{tS} \mid x_\sigma]$. Then $\log p_\sigma(Y_k = 1 \mid x_\sigma) = \log c + \Lambda(w)$ by Equation (7). By linearity, $\Lambda'(0) = \mathbb{E}[S \mid x_\sigma] = S_k(x_\sigma;\sigma)$, so $\log \bar{p}_\sigma(Y_k = 1 \mid x_\sigma) = \log c + w\, S_k(x_\sigma;\sigma) = \log c + w\,\Lambda'(0)$ by Equation (5) and Equation (6), and the gap is $\Lambda(w) - \Lambda(0) - w\,\Lambda'(0)$. Since S is bounded, Λ is smooth and $\Lambda''(t) = \mathrm{Var}_t[S]$, the variance under the posterior tilted by e^{tS}. Taylor's formula with integral remainder gives:

$$\Lambda(w) - \Lambda(0) - w\,\Lambda'(0) \;=\; \int_0^w (w - t)\,\mathrm{Var}_t[S]\,\mathrm{d}t. \tag{14}$$

Since $S \in [-1,1]$, the density of the tilted posterior with respect to the posterior, $e^{tS}/\mathbb{E}[e^{tS} \mid x_\sigma]$, lies in $[e^{-2t}, e^{2t}]$. Hence, for every $b \in \mathbb{R}$, $\mathbb{E}_t[(S-b)^2]$ lies between e^{-2t} and e^{2t} times $\mathbb{E}[(S-b)^2 \mid x_\sigma]$; taking the minimum over b gives $e^{-2t}V_k(x) \le \mathrm{Var}_t[S] \le e^{2t}V_k(x)$. Bounding $e^{\pm 2t}$ by $e^{\pm 2w}$ on $[0,w]$ and using $\int_0^w (w-t)\,\mathrm{d}t = w^2/2$ gives Equation (8). □

Proof of Proposition 1. (i) f_{fair} is the maximum of the finitely many linear functions $A \mapsto \sum_i a_{i\pi(i)}$, $\pi \in \Pi$, hence convex and piecewise linear. Where the maximiser $\pi^\star$ is unique, the inequalities that make it the maximiser are strict and remain so on a neighbourhood of A, where f_{fair} therefore coincides with $\sum_i a_{i\pi^\star(i)}$; its partial derivatives are $\mathbf{1}[\pi^\star(i) = k]$. (ii) Since a_{ik} depends on x_i

only, $\nabla_{x_i} f_{\text{fair}} = \sum_k \mathbf{1}[\pi^\star(i) = k]\, \nabla_{x_i} a_{ik} = \nabla_{x_i} a_{i\pi^\star(i)}$ wherever (i) applies, and substituting into Equation (2) gives Equation (10). □

A.2 Binary attributes

For a binary attribute, each value is described against the other instead of against a neutral sentence. With t_+ and t_- the two sentences, for gender *a photo of a woman* and *a photo of a man*,

$$\Delta = \frac{E_\vartheta^{\text{txt}}(t_+) - E_\vartheta^{\text{txt}}(t_-)}{\|E_\vartheta^{\text{txt}}(t_+) - E_\vartheta^{\text{txt}}(t_-)\|}, \qquad \Delta_1 = \Delta, \qquad \Delta_2 = -\Delta. \tag{15}$$

The clean logit $S(x_0; 0) = \langle E_\vartheta^{\text{img}}(x_0), \Delta \rangle$ is positive when the image is closer to t_+ than to t_-, and its sign is the zero-shot classifier of Section 3. With $a_i = S(x_i; \sigma)$, the noisy logit of Equation (5) along Δ, the scores of Equation (9) are $a_{i1} = a_i$ and $a_{i2} = -a_i$. Writing $s_i = +1$ if $\pi(i) = 1$ and $s_i = -1$ otherwise, the total score of an assignment is $\sum_i s_i a_i$, with exactly $B_+ = B_1$ positive signs. Moving a $+1$ from a smaller score to a larger one cannot decrease this sum, so it is maximised by giving the positive signs to the B_+ largest scores (the rearrangement inequality). The assignment of Equation (9) thus reduces to ranking the batch along Δ,

$$s_i = \begin{cases} +1 & \text{if } a_i \text{ is among the } B_+ \text{ largest scores of the batch,} \\ -1 & \text{otherwise,} \end{cases} \qquad f_{\text{fair}}(x^{1:B}) = \sum_{i=1}^{B} s_i\, a_i, \tag{16}$$

and Equation (10) becomes $\widetilde{D}(x_i; \sigma) = D_\theta(x_i; \sigma) + w_{\text{fair}}\, \sigma^2\, s_i\, \nabla_{x_i} a_i$: each image is pushed so that its score moves towards the pole it is assigned to. The maximiser is unique wherever the B_+-th and $(B_+ + 1)$-th largest scores differ. The CelebA experiments of Section 5 use this case.

A.3 Alternative diversity term

A second diversity term, which we name `SigLIPMS`, takes the second estimate of Equation (12) from the frozen image encoder applied to the Tweedie estimate:

$$f_{\text{div}}(x) = 1 - \big\langle m(x_\sigma; \sigma),\ \text{sg}\Big[E_\vartheta^{\text{img}}(\hat{x}_0(x))\Big] \big\rangle \approx 1 - \big\langle \varphi_\psi(h(x_\sigma); \sigma),\ \text{sg}\Big[E_\vartheta^{\text{img}}(\hat{x}_0(x))\Big] \big\rangle. \tag{17}$$

The term uses the very discrepancy that motivated the adapter: $m(x_\sigma; \sigma) = \mathbb{E}[E_\vartheta^{\text{img}}(x_0) \mid x_\sigma]$ takes the expectation in embedding space and $E_\vartheta^{\text{img}}(\hat{x}_0(x))$ in pixel space; the two coincide when the posterior is concentrated on one image, and they differ when it spreads over several. The term is thus small where the model already knows what it is generating, on the modes, and large in the tail. The two terms are the two natural ways to obtain a second estimate, one external and one internal, and they fail differently: `SigLIPMS` inherits the encoder's behaviour on blurred images, while our original diversity term inherits the adapter's own errors. Section D compares them per attribute.

A.4 Density baseline

The diversity terms of Sections A.3 and 4.3 compare two estimates of the same image. A second family, after Sehwag et al. (2022), compares $m(x_\sigma; \sigma)$ with a fixed reference set $\mathcal{R}$ of 10,000 embeddings $E_\vartheta^{\text{img}}(x_0)$ of real images and measures rarity relative to the data. We use the mean distance to the κ nearest neighbours (`avgKNN`),

$$f_{\text{div}}(x) = \frac{1}{\kappa} \sum_{r \in \mathcal{N}_\kappa(x)} \big\| m(x_\sigma; \sigma) - r \big\|, \tag{18}$$

with $\mathcal{N}_\kappa(x)$ the κ nearest neighbours of $m(x_\sigma; \sigma)$ in $\mathcal{R}$, computed, as in Section 4.2, with the adapter output in place of $m(x_\sigma; \sigma)$; both lie on the unit sphere, so this is a function of the inner products $\langle \varphi_\psi(h(x_\sigma); \sigma), r \rangle$, $r \in \mathcal{R}$. The two families measure different quantities, rarity relative to the data against the posterior spread of the model. They are compared in Section D. In every case, $w_{\text{div}} > 0$ increases the term and therefore pushes towards low-density regions.

B Experimental details

This section contains all details needed to reproduce Section 5, except for adapters which are described in Section C.

Implementation details. We implemented our method using PyTorch (Paszke et al., 2019) with the help of SciPy (Virtanen et al., 2020) for the matching algorithm. We will release in open source our code and adapter checkpoints together with the paper release. Every run uses one NVIDIA A100 (40 GB) GPU.

B.1 Generators

On CelebA (Liu et al., 2015) we train an EDM model (Karras et al., 2022) at 64×64: a U-Net (Ronneberger et al., 2015) of 19.4 M parameters (base width 96, multipliers $(1, 2, 2, 2)$, two residual blocks per level in the encoder and three in the decoder, self-attention at 16^2, 8^2 and in the middle block, dropout 0.1), trained with the EDM loss and preconditioning on the 162,770 training images (centre crop 140, no flip) with Adam (Kingma & Ba, 2015), batch 256, EMA 0.9999 and a learning rate of $2 \cdot 10^{-4}$ divided by 2, 10 and 10 at plateaus of the validation loss; the EMA weights of lowest validation loss are reached after 1,487 epochs ($9.5 \cdot 10^5$ steps, 8.5 days). On CelebA-HQ (Karras et al., 2018) we use the released P2 model (Choi et al., 2022), an ADM U-Net (Dhariwal & Nichol, 2021) of 93.6 M parameters at 256×256 (width 128, multipliers $(1, 1, 2, 2, 4, 4)$, attention at 16^2) with a linear β schedule over 1,000 steps. The text-to-image model is the released Stable Diffusion 1.5 from Rombach et al. (2022) (860 M-parameter U-Net, 512×512).

B.2 Attributes, prompts and sample sizes

On CelebA (Liu et al., 2015) the attributes are gender (*a photo of a man* / *a photo of a woman*) and eyeglasses (*a person with eyeglasses or sunglasses* / *a person without eyeglasses or sunglasses*). On CelebA-HQ (Karras et al., 2018), following Balancing Act (Parihar et al., 2024), gender keeps the CelebA pair, race adds *a photo of a black person* / *a photo of a white person*, and eyeglasses is *a photo of a person with eyeglasses* / *a photo of a person without eyeglasses*. All these attributes are binary and use the form of Section A.2. On Stable Diffusion (Rombach et al., 2022), following DiffLens (Shi et al., 2025), images are generated from *A face of a* {*doctor, firefighter, nurse, receptionist*}. Gender is binary (*a photo of a woman* / *a photo of a man*); age and race use the K-valued form of Section 4.2 with $t_0 =$ *a photo of a person*, and $t_1, t_2, t_3 =$ *a photo of a young / middle-aged / old person* for age, $t_1, \ldots, t_4 =$ *a photo of a white / black / asian / indian person* for race, the four FairFace groups of the DiffLens protocol. The assignment of Equation (9) is solved with the Hungarian algorithm (Kuhn, 1955) on a cost matrix in which the column of value k is repeated B_k times, so that the quotas are met exactly. The targets are uniform, $\alpha_k = 1/K$. Each configuration is sampled with five seeds of 5,000 images on CelebA, one seed of 10,000 images per attribute on CelebA-HQ, and one seed of 500 images per prompt on Stable Diffusion.

B.3 Sampling

CelebA: deterministic Heun solver (Karras et al., 2022), 100 steps on $\sigma \in [0.002, 80]$ ($\rho = 7$), EMA weights, guidance for $\sigma \le \sigma_{\max} = 10$ (67 steps), batch size $B = 100$; SGMS (Um & Ye, 2024) and our diversity term of Equation (12) act at every fifth network evaluation and re-noise at the fixed level $\sigma_s = 0.96$ of Equation (11), `SigLIPMS` (Section A.3) at every evaluation. **CelebA-HQ**: DDIM (Song et al., 2021a), 50 steps, $\eta = 1$, guidance at every step, batch size $B = 50$; the unguided model and Balancing Act (Parihar et al., 2024) are sampled with the released code on the same schedule with batch size $B = 100$. **Stable Diffusion**: Euler with trailing timesteps, 30 steps, classifier-free guidance (Ho & Salimans, 2022) at scale 7.5, guidance for $\sigma \in [0.2, 4]$ (21 steps), batch size $B = 100$ images of a single prompt. A run takes 18–25 min per 5,000 images on CelebA, 2.1–2.9 h per 10,000 on CelebA-HQ and 37 min per 2,000 on Stable Diffusion.

B.4 Baselines

On CelebA-HQ we rerun Balancing Act (Parihar et al., 2024), a predictor trained on labelled attributes from the same U-Net activations, at its published setting (scale 1,500, batch size 100), and the exemplar variant of Score Guidance (Tiwary et al., 2026), which steers a share of each batch towards the DDIM-inverted scores of eight exemplar images per class, with the authors' code and gender preset on all three attributes. Every method is then evaluated in the same way, under the protocol of Balancing Act: the same reference sets and our own attribute classifiers (Section B.5). On

Stable Diffusion we evaluate with the released code of DiffLens (Shi et al., 2025) and reproduce its unguided row (FD 0.565, 0.747 and 0.560 on gender, age and race against 0.564, 0.752 and 0.558 reported; FID 117.1 against 120.1), so that our numbers are comparable with those it reports for latent editing (Kwon et al., 2023), h-space distribution guidance (Parihar et al., 2024), discovered latent directions (Li et al., 2024), fine-tuning (Shen et al., 2024) and DiffLens itself; ICM (Zaleska et al., 2026), which locates the attribute with labeled probes, is added under the same protocol. On CelebA our two diversity terms are compared with SGMS (Um & Ye, 2024), the pixel-space score they transpose.

For Figure 3, Balancing Act is run in its per-sample mode, in which every image is guided towards a fixed attribute value, here the combination of its gender and race predictors set to female and Black, at the same scale as the single-attribute runs. Its batch-level distribution mode, which matches a target proportion over the batch, did not reach the compound target in our runs.

B.5 METRICS

Reference sets. On CelebA (Liu et al., 2015) the reference has the size of the generated set, drawn for the diversity experiments from the minority set of SGMS (Um & Ye, 2024), the 10,000 training images of highest AvgkNN (Sehwag et al., 2022) (5 nearest neighbours, ResNet-50 features (He et al., 2016)), and for the fairness experiments with replacement from the training set by a sampler balanced on the attribute. On CelebA-HQ (Karras et al., 2018), as in Balancing Act (Parihar et al., 2024), it is 5,000 images balanced with the attribute labels for gender and eyeglasses (2,500 per class; the 1,468 images with eyeglasses are drawn with replacement), and for race, which has no label, the 2,500 images most and the 2,500 least similar to *a black person* under CLIP ViT-B/32 (Radford et al., 2021). On Stable Diffusion we follow DiffLens (Shi et al., 2025), which takes FFHQ (Karras et al., 2019) as the real reference for text-to-image models and computes FID on the 2,000 images of the four prompts at once.

Quality and diversity. FID (Heusel et al., 2017) is computed on Inception-V3 features (Szegedy et al., 2016) on CelebA and CelebA-HQ, and with Clean-FID (Parmar et al., 2022) on Stable Diffusion. On CelebA we add MIND (Berthet et al., 2026), an optimal-transport distance on the same features (1,000 random projections, equal set sizes) that drops the Gaussian assumption of FID, known to misrank diffusion models (Stein et al., 2023) and to bias the estimate with sample size (Chong & Forsyth, 2020); precision and recall(Kynkäänniemi et al., 2019), all with five nearest neighbours, and density and coverage (Naeem et al., 2020), which separate fidelity from coverage; and the Vendi score (Friedman & Dieng, 2023) with a cosine kernel, which needs no reference. On Stable Diffusion, CLIP-T and CLIP-I (Radford et al., 2021; Shi et al., 2025) are the similarity of a sample to its prompt and to the unguided sample of the same prompt; like FD, they are computed per occupation and averaged over the four.

Fairness. On CelebA and CelebA-HQ, FD is computed from hard predictions: $\mathrm{FD} = \|(p, 1-p) - (\frac{1}{2}, \frac{1}{2})\|_2 = \sqrt{2}\,|p - \frac{1}{2}|$, with p the proportion of class 0 among hard predictions. For gender and eyeglasses, p is the CLEAM-corrected proportion p^* (Teo et al., 2023) of class 0 (female, no eyeglasses). On CelebA the attribute is counted by two CNNs trained on the training split with class-balanced sampling, at 98.8% (gender) and 99.5% (eyeglasses) validation accuracy, with per-class accuracies $(\alpha_0, \alpha_1) = (0.987, 0.989)$ and $(0.997, 0.981)$. On CelebA-HQ, as in Balancing Act (Parihar et al., 2024), gender and eyeglasses are counted by ResNet-18s (He et al., 2016) pretrained on ImageNet and fine-tuned on 80% of CelebA-HQ (97.9% and 98.2% balanced test accuracy), with per-class recalls $(0.989, 0.968)$ and $(0.998, 0.966)$. Race, which has no label, is counted by FairFace (Kärkkäinen & Joo, 2021) restricted to its White and Black outputs; since the lack of labels prevents a proper evaluation of this classifier, race is not CLEAM-corrected and p is the raw proportion of White predictions. On Stable Diffusion, as in DiffLens (Shi et al., 2025), faces are cropped with dlib (King, 2009) and classified by FairFace (Kärkkäinen & Joo, 2021) into two genders, three ages (0–19, 20–59, 60+) and four races (White, Black, East and Southeast Asian, Indian, not renormalised), and FD is the distance of the mean softmax output to the uniform target, without CLEAM correction, averaged over the four occupations.

Table 4: Adapter training, on one A100 GPU.

Backbone	Data	Noise levels	Optimisation	Epochs	Time
EDM, CelebA	162,770 training images	σ log-uniform on $[0.002, 80]$	AdamW, batch 512, lr 10^{-3}	50	2.9 h
P2, CelebA-HQ	28,000 (+2,000 val.)	t uniform on the 49 guided steps	AdamW, batch 32, lr 10^{-3}, wd 10^{-4}	47 of 60	4.2 h
SD 1.5	49,000 generated (+1,000 val.)	σ uniform on $[0.03, 14.6]$	AdamW, batch 64, lr $5 \cdot 10^{-4}$, wd 0.05, warm-up and cosine	47 of 50	11 h

B.6 Protocols

On CelebA we keep the w_{fair} of lowest MIND in the sweep of Section D.2 (5 on gender, 200 on eyeglasses) and, for each diversity term, the w_{div} of best diversity in the ablation of Section D.3; entries are means over five seeds, with standard deviations in Section D.4. On CelebA-HQ the fairness setting of Table 2 uses $w_{\text{fair}} = 20$ / 100 / 500 on gender / race / eyeglasses, and the quality setting 5 / 30 / 180. With the diversity term in the fairness setting, we choose $w_{\text{div}} = 40$, re-noising at $\sigma_s = 0.5\,\sigma$ every second step for $t \le 800$, with the race weight lowered to 75. On Stable Diffusion $w_{\text{fair}} = 3{,}000$ on every attribute.

C Adapter architecture and training

The adapter φ_ψ of Section 4.1 is learned from clean images only. Each training image x_0 is noised at two independent levels, $x_\sigma = x_0 + \sigma\varepsilon$ (on P2 (Choi et al., 2022), $x_t = \sqrt{\bar\alpha_t}\,x_0 + \sqrt{1-\bar\alpha_t}\,\varepsilon$), passed through the frozen backbone, and $\varphi_\psi(h(x_\sigma);\sigma)$ is regressed onto the image embedding $E^{\text{img}}_\vartheta(x_0)$ of the clean image, given by SigLIP 2 ViT-B/16 (Tschannen et al., 2025) (768 dimensions), under the cosine loss of Equation (4). No sampler trajectory is used. The targets are fixed, one per image and precomputed once, so a contrastive loss would have nothing to discriminate that this regression does not already do.

Unconditional backbones. On EDM (Karras et al., 2022) and P2 (Choi et al., 2022), h is the output of the U-Net (Ronneberger et al., 2015) middle block (192 and 512 channels at 8×8), read as 64 tokens with a learned position embedding, pooled by attention with a learned query and a residual MLP, normalised, modulated by FiLM (Perez et al., 2018) on a sinusoidal embedding of the noise level ($\frac{1}{4}\ln\sigma$ on EDM, $t/999$ on P2) initialised at zero, and mapped to 768 dimensions by a two-layer MLP followed by ℓ^2 normalisation (0.79 M and 2.42 M parameters).

Text-conditional backbone. The adapter reads the prompt-conditioned branch $D_\theta(x;\sigma,c)$ of classifier-free guidance (Ho & Salimans, 2022):

$$D^\gamma_\theta(x;\sigma) \;=\; D_\theta(x;\sigma,\varnothing) \;+\; \gamma\,\big(D_\theta(x;\sigma,c) - D_\theta(x;\sigma,\varnothing)\big), \tag{19}$$

with c the prompt, $\varnothing$ the empty prompt and γ the guidance scale. Its input combines the mid-block and the four decoder blocks of that branch:

$$h(x_\sigma) \;=\; \textstyle\sum_{\ell=1}^{5} \beta_\ell\,\pi_\ell\big(h_\ell(x_\sigma)\big), \qquad \beta = \mathrm{softmax}(b), \tag{20}$$

where h_ℓ are the block activations (1280×8^2 to 320×64^2), π_ℓ timestep-conditioned bottlenecks to $512 \times 8 \times 8$ and b learned logits; a six-layer transformer (width 512) with attention pooling maps the 64 tokens to the embedding space (21.2 M parameters in all). Since the logits are read before the combination Equation (19), they do not depend on γ. A readout of the middle block alone reaches the gender target as well, but not eyeglasses, which it tends to paint as a bright mask on the lower face.

Data for text-to-image. The adapter must read the h-space of this model, so its training images are generated by Stable Diffusion 1.5 (Rombach et al., 2022) itself: 50,000 images at 512×512 (Euler with trailing timesteps, 30 steps, classifier-free guidance (Ho & Salimans, 2022) at scale 7.5), in 6.2 h.

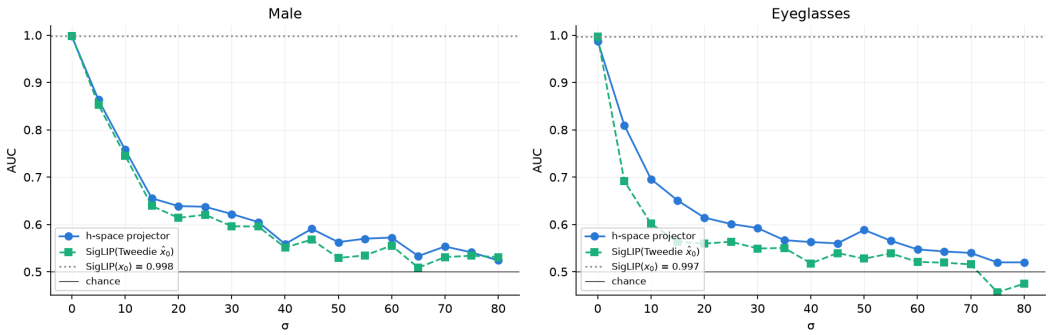

Figure 4: What the CelebA adapter reads of the attribute as the noise grows. AUC of the logit $S(x_\sigma; \sigma)$ along the direction Δ of Equation (15), against the CelebA label, on 500 images per class of the test split, for gender (*a photo of a man* / *a photo of a woman*, left) and eyeglasses (*a photo of a person wearing eyeglasses* / *a photo of a person*, right). Circles use the adapter output $\varphi_\psi(h(x_\sigma); \sigma)$, squares the encoder on the Tweedie estimate, $E^{\text{img}}_\vartheta(\hat{x}_0)$; the dotted line is the encoder on the clean image, $E^{\text{img}}_\vartheta(x_0)$. Over the guidance window $\sigma \leq 10$ the adapter matches the Tweedie estimate on gender and exceeds it on eyeglasses, at no decoding cost.

Each sample stores the clean latent x_0, its prompt and the target E^{img}_ϑ of the decoded image; at every training step x_0 is noised and passed through the frozen U-Net. Four templates (*a photo of*, *a portrait photo of*, *a realistic centered portrait of*, *a face of*) crossed with *a person* and the four evaluation occupations are used (20 prompts). None name any attribute, so the attribute in each image comes from the model's own bias, not from the text conditioning.

Training. Table 4 lists the data, noise levels and cost of each adapter. The noise levels cover the range where guidance acts: log-uniform over the full EDM range on CelebA, the 49 guided DDIM steps on P2, and uniform over the Stable Diffusion schedule (Rombach et al., 2022). On CelebA-HQ and Stable Diffusion the kept epoch is the one of best validation retrieval (R@1 among the validation targets). Figure 4 shows what the CelebA adapter retains of the attribute as the noise grows.

D Ablations on CelebA

We test the following design choices: where the score is read, the fairness weight, the diversity weight, and the choice of diversity term; a last section runs each diversity term on its own, without the fairness term. The chosen diversity term is then used in CelebA-HQ and Stable Diffusion experiments.

The following properties are shared between all runs: the EDM backbone of Section B, the deterministic Heun sampler, guidance active for $\sigma \leq \sigma_{\max} = 10$, and the fairness term of Section 4.2 read in the h-space through the adapter. Unless a section says otherwise, FID, MIND, precision, recall, density and coverage are computed against real CelebA images resampled to be balanced on the attribute being steered, as many as the generated images, and every FD is CLEAM-corrected. In every table each column is shaded from its best value (dark, bold) to its worst (light), within each block where a table has several; in the sweep tables the shading orders the settings of one method.

D.1 Where the score is computed: h-space against pixel space

"Pixel" here denotes the fairness term computed from SigLIP applied to the Tweedie estimate $\hat{x}_0$, the readout that Section 4.1 argued moves the expectation through a nonlinearity evaluated off the support of clean images, whereas the h-space estimate moves it through a linear functional. Figure 6 is the visible counterpart on the compound target of Figure 5, and Table 5 and Figure 7 the quantitative one on the two binary attributes, at equal guidance weights.

Every sample of the top row of Figure 6 carries both attributes. At an equal budget the pixel-space score reaches the target for part of the batch only, and more than doubling the number of steps does not close the gap: several samples still miss the sunglasses, the skin tone, or both. The failure is

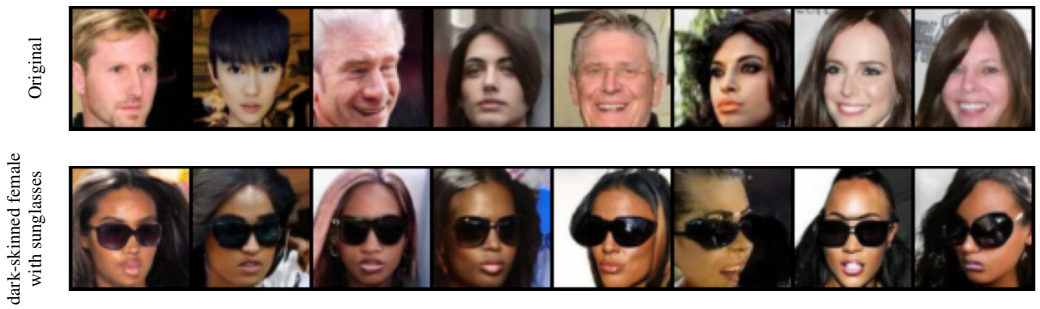

Figure 5: Steering an attribute that has no label. Top: unguided samples from a diffusion model trained on CelebA. Bottom: the same eight initial noises, guided towards *dark-skinned female with sunglasses*. The attribute is handed to the sampler as this sentence and a contrasting one describing its absence. CelebA has no label for it, and neither the generator nor any classifier was trained.



Figure 6: Same compound target, *dark-skinned female with sunglasses*, three ways of reading the guidance term. Top: from $\varphi(h(x_\sigma);\sigma)$. Middle: from $E_\vartheta^{\text{img}}(\hat{x}_0)$ at the same sampling budget. Bottom: from $E_\vartheta^{\text{img}}(\hat{x}_0)$ with more than twice the budget. Eight samples per row.

not uniform across the batch, which is what an uncontrolled, off-support evaluation produces: at the noise levels where the guidance acts, $\hat{x}_0$ is an average across modes and its embedding points at no single one of them.

On the strongly skewed eyeglasses attribute the two estimates are not comparable in effect. The h-space score drives the discrepancy from 0.614 to 0.156 and improves MIND and FID until $w = 200$ (MIND 8.9, FID 11.6); the pixel-space score stalls at a discrepancy of 0.489, never entering the range the h-space covers, and pays for that small gain with a rising MIND ($19.1 \rightarrow 21.0$) and a rising FID ($16.2 \rightarrow 22.1$). On the nearly balanced gender attribute the ranking reverses but the margin collapses: the pixel-space score takes MIND on every guided row, by 2% at the selected weight $w = 5$ and 13% at $w = 50$, against a seed-to-seed standard deviation of 5 to 10% of MIND measured on the same configuration, so the two are indistinguishable there. The h-space score is also cheaper, its gradient stopping at the mid-block instead of traversing the full denoiser and the encoder.

D.2 Fairness weight

Table 5 doubles as the fairness-weight sweep. On both attributes the discrepancy falls monotonically with w_{fair}, but quality does not: MIND and FID first improve, as the guidance corrects the imbalance, reach a minimum, and then degrade. On eyeglasses the minimum is at $w_{\text{fair}} = 200$; between it and $w_{\text{fair}} = 600$, MIND rises from 8.9 to 15.1 and FID from 11.6 to 21.2 while the discrepancy gains

Table 5: h-space against pixel-space fairness term at equal weight w; FD on the steered attribute; both share the $w = 0$ run, 5000 images, one seed. Bold: the better MIND of the two on each row.

	h-space guidance			pixel guidance		
w	MIND	FD	FID	MIND	FD	FID
Eyeglasses attribute (FD = 0.614 *unguided*)						
0	19.13	0.614	16.23	19.13	0.614	16.23
50	**15.14**	0.534	13.81	19.69	0.611	16.92
100	**11.81**	0.441	12.14	19.88	0.598	17.66
150	**9.97**	0.377	11.60	20.38	0.587	18.56
200	**8.92**	0.320	11.63	20.51	0.573	19.15
400	**10.64**	0.198	15.51	20.86	0.534	20.81
600	**15.12**	0.156	21.17	20.95	0.489	22.07
Gender attribute (FD = 0.0963 *unguided*)						
0	2.49	0.0963	6.31	2.49	0.0963	6.31
2	2.31	0.0644	6.24	**2.21**	0.0395	6.19
5	2.26	0.0302	6.25	**2.21**	0.0201	6.22
10	2.41	0.0094	6.44	**2.34**	0.0097	6.39
20	2.85	0.0021	6.87	**2.64**	0.0076	6.70
40	3.83	0.0010	7.69	**3.49**	0.0027	7.47
50	4.30	0.0010	8.08	**3.81**	0.0018	7.79

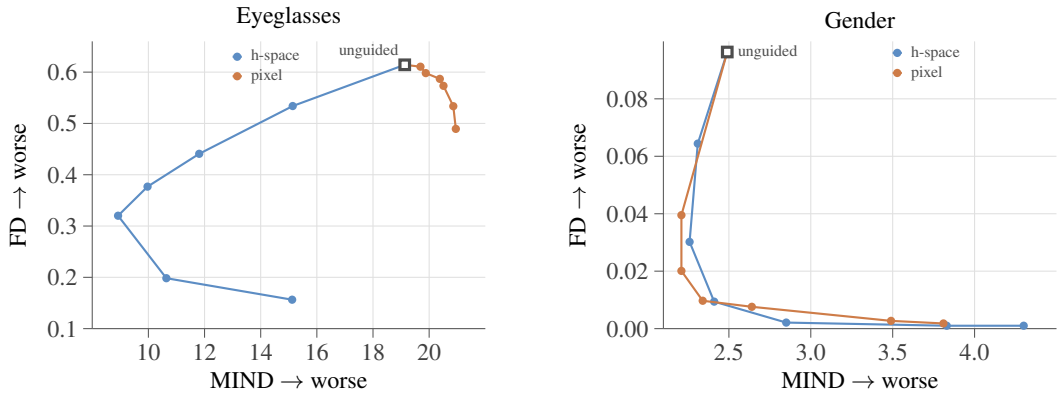

Figure 7: FD–MIND trade-off, h-space versus pixel-space fairness term; each polyline follows its own sweep in order of increasing guidance weight w, and the white square is the shared unguided model. Left: on eyeglasses the two curves do not overlap, pixel guidance stays pinned in the top-right corner, drifting *rightwards* as w grows, while h-space sweeps down and to the left, then turns back up past $w = 200$ as fidelity starts to cost, so no pixel setting matches any h-space setting on both axes. Right: on gender the two sweeps trace the same short hook, pixel running marginally inside h-space; both bottom out in MIND around $w = 5$ and then climb steeply as FD is driven to zero.

only its last 0.16. On gender the minimum is at $w_{\text{fair}} \approx 5$ and the climb past it is steep, MIND reaching 4.3 at $w_{\text{fair}} = 50$ for a discrepancy already at 0.001. The setting that maximises quality is therefore not the setting that maximises parity, and the useful operating range ends well before parity is reached. Steering eyeglasses also moves the gender balance: the gender discrepancy rises from 0.096 to 0.343 at $w_{\text{fair}} = 600$, a collateral shift that a multi-attribute target would have to control.

D.3 Diversity weight

Four diversity terms are compared from here on: `PerturbProj`, the term of Equation (12); `SigLIPMS`, its variant of Section A.3; and the two baselines of Section B.4, `SGMS` and `avgKNN`. With the fairness weight fixed at the value selected above, each term is swept over its own weight; $w_{\text{div}} = 0$ reproduces the fairness-only run in every block, which serves as a consistency check, and $\star$ marks the setting carried forward to Section D.4.

Tables 6 and 7 use 2000 images per setting, so their FID and MIND are inflated relative to Tables 5 and 8 and comparable only within a sweep; the selected settings are re-evaluated at 5,000 images and five seeds in Table 8.

Table 6: Diversity-weight sweep on eyeglasses, fairness weight fixed at 200. $w_{\text{div}} = 0$ is the fairness-only baseline, 2000 images, one seed; $\star$ marks the setting carried to Table 8.

Method	w_{div}	MIND↓	FD↓	FID↓	Vendi↑	Prec.↑	Rec.↑	Cov.↑	Dens.↑
`avgKNN`	0	10.76	0.336	**19.21**	6.47	**0.814**	0.679	**0.893**	**1.218**
	2	10.56	0.332	**19.21**	6.52	0.812	0.689	**0.893**	1.208
	5*	**10.44**	0.330	19.39	6.58	0.805	0.690	**0.893**	1.190
	10	10.46	0.330	19.78	6.69	0.795	**0.697**	0.880	1.159
	20	11.24	0.317	20.77	6.88	0.778	0.692	0.870	1.131
	40	14.72	**0.309**	23.80	**7.31**	0.765	0.696	0.835	1.035
`SGMS`	0	10.75	0.336	**19.18**	6.48	**0.815**	0.683	**0.892**	**1.217**
	10	**10.64**	0.332	19.52	6.57	0.807	0.695	0.887	1.186
	20*	10.68	0.327	19.93	6.63	0.797	0.694	0.881	1.171
	40	10.84	0.327	20.72	6.78	0.778	0.696	0.867	1.120
	60	11.29	0.323	21.75	6.97	0.766	0.704	0.860	1.085
	80	12.08	**0.316**	22.91	**7.12**	0.756	**0.714**	0.852	1.047
`SigLIPMS`	0	10.77	0.336	19.21	6.47	**0.813**	0.682	0.894	**1.215**
	20	9.89	0.327	19.02	6.54	0.810	**0.685**	**0.897**	1.210
	40	9.43	0.317	**18.92**	6.59	0.806	0.672	0.889	1.200
	80	**9.07**	0.298	19.21	6.72	0.806	0.675	0.886	1.180
	120*	9.49	0.287	19.99	6.85	0.807	0.678	0.879	1.165
	160	10.26	0.283	20.72	6.94	0.800	0.657	0.875	1.153
	200	11.21	0.280	21.56	7.07	0.795	0.669	0.863	1.168
	240	12.58	0.277	22.57	7.15	0.790	0.668	0.855	1.135
	300	15.05	**0.267**	24.31	**7.26**	0.794	0.655	0.830	1.094
`PerturbProj`	0	10.77	0.336	19.21	6.47	**0.813**	0.682	0.894	**1.215**
	20	10.60	0.334	**19.19**	6.50	0.812	0.680	0.889	1.212
	40	10.48	0.332	19.21	6.52	0.811	0.682	0.886	1.209
	80	10.31	0.326	19.29	6.55	0.804	**0.688**	0.886	1.201
	120	10.27	0.323	19.42	6.57	0.802	0.684	0.889	1.198
	160	10.22	0.320	19.49	6.59	0.806	0.678	0.890	1.196
	200	10.18	0.317	19.58	6.61	0.805	0.677	0.892	1.200
	240	10.09	0.313	19.62	6.65	0.802	0.669	**0.895**	1.187
	300*	**10.05**	0.309	19.87	6.72	0.787	0.673	0.881	1.185
	350	10.08	**0.307**	20.09	**6.77**	0.790	0.672	0.883	1.177

The columns disagree, so no setting dominates. On eyeglasses (Table 6) FID and density are best at or near $w_{\text{div}} = 0$ while discrepancy and Vendi keep improving to the end of every sweep; the retained value is a compromise, and for `SGMS` and `SigLIPMS` it is deliberately past the MIND minimum, buying discrepancy and Vendi at a small MIND cost. Only `SigLIPMS` improves MIND, discrepancy and FID together over its own baseline, at w_{div} between 40 and 80; the other three buy discrepancy

Table 7: Diversity-weight sweep, on gender, fairness weight fixed at 10. $w_{\text{div}} = 0$ is the fairness-only baseline, 2000 images, one seed; $\star$ marks the setting carried to Table 8.

Method	w_{div}	MIND↓	FD↓	FID↓	Vendi↑	Prec.↑	Rec.↑	Cov.↑	Dens.↑
avgKNN	0	2.76	0.0104	12.41	6.18	0.829	0.723	0.970	**1.133**
	2	2.65	0.0075	12.40	6.24	0.828	0.723	0.969	1.124
	5*	**2.54**	0.0039	**12.39**	6.32	**0.832**	0.732	**0.973**	1.113
	10	2.56	**0.0017**	12.54	6.45	0.820	0.734	0.963	1.094
	20	3.42	0.0019	13.28	6.72	0.824	0.732	0.967	1.060
	40	8.34	0.0121	16.81	**7.26**	0.792	**0.749**	0.902	0.936
SGMS	0	2.77	0.0104	**12.41**	6.18	**0.830**	0.724	0.969	**1.135**
	5	2.66	0.0075	12.42	6.22	0.828	0.718	0.971	1.117
	10	2.60	0.0068	12.46	6.27	0.826	0.722	**0.973**	1.106
	20*	**2.51**	**0.0010**	12.54	6.38	0.824	0.738	0.970	1.076
	40	2.72	0.0039	12.88	6.52	0.808	0.753	0.964	1.035
	60	3.29	0.0017	13.47	6.68	0.786	0.750	0.960	0.985
	80	4.47	0.0012	14.58	**6.90**	0.771	**0.756**	0.960	0.949
SigLIPMS	0	2.76	0.0104	**12.41**	6.18	0.829	0.723	0.970	1.133
	20*	**2.54**	**0.0017**	12.42	6.28	**0.836**	0.724	**0.972**	**1.136**
	40	2.58	0.0031	12.59	6.39	0.827	0.723	0.971	1.111
	80	3.71	0.0128	13.60	6.65	0.821	**0.726**	0.966	1.085
	120	5.81	0.0186	14.99	6.83	0.819	0.707	0.945	1.071
	160	9.39	0.0360	17.11	**7.01**	0.818	0.714	0.939	1.038
PerturbProj	0	2.77	0.0104	12.41	6.18	0.830	0.726	0.970	**1.134**
	10	2.74	0.0082	12.40	6.19	0.830	0.725	0.971	**1.134**
	20	2.70	0.0068	12.39	6.21	**0.834**	0.725	0.972	1.133
	40	2.66	0.0053	**12.38**	6.23	0.832	0.723	0.973	1.130
	80	2.58	0.0039	12.39	6.26	**0.834**	0.728	**0.975**	1.124
	120*	2.56	0.0031	12.47	6.31	0.833	0.719	0.974	1.107
	160	**2.55**	0.0039	12.54	6.36	0.833	**0.732**	0.973	1.107
	200	2.60	**0.0017**	12.63	**6.42**	0.828	0.724	0.971	1.095

and Vendi while giving up FID, precision, coverage and density, the four moving as one. On gender (Table 7) all four terms improve MIND and discrepancy over their baseline, and coverage holds or rises while density falls only slightly, which is why the final table shows no fidelity penalty there. Every sweep is unimodal in MIND and turns sharply past its optimum (`avgKNN` reaches 8.34 at $w_{\text{div}} = 40$, `SigLIPMS` 9.39 at 160), so the useful range is narrow.

D.4 Diversity on top of the fairness term at the selected setting

Table 8 compares the four diversity terms at the fairness weight of Section D.2 and the diversity weight of Section D.3, as means and standard deviations over five sampling seeds of $5,000$ images; it is the table behind the diversity findings of Section 5.2. Table 9 repeats the comparison with $50,000$ images per configuration and a single seed.

Table 8: Diversity terms on top of the fairness term at the selected weights, 5000 images, 5 seeds; time is wall-clock generation of one run and is not ranked.

	time (s)	FD↓	FID↓	MIND↓	Vendi↑
Gender, $w_{fair} = 5$					
Debias Anything	1613 ± 665	0.031 ± 0.004	6.59 ± 0.04	2.52 ± 0.13	6.26 ± 0.03
+ `SGMS`	1380 ± 5	0.023 ± 0.005	6.46 ± 0.05	2.05 ± 0.14	$\mathbf{6.46 \pm 0.02}$
+ `avgKNN`	1635 ± 713	0.021 ± 0.003	$\mathbf{6.40 \pm 0.11}$	2.06 ± 0.15	6.42 ± 0.03
+ `SigLIPMS`	2179 ± 962	0.011 ± 0.004	6.48 ± 0.07	2.13 ± 0.17	6.41 ± 0.04
+ `PerturbProj`	1476 ± 592	$\mathbf{0.006 \pm 0.003}$	6.44 ± 0.10	$\mathbf{2.03 \pm 0.15}$	6.43 ± 0.04
Eyeglasses, $w_{fair} = 200$					
Debias Anything	1619 ± 662	0.304 ± 0.004	11.47 ± 0.12	9.06 ± 0.48	6.62 ± 0.02
+ `SGMS`	1418 ± 51	0.300 ± 0.002	11.87 ± 0.10	8.81 ± 0.44	6.80 ± 0.01
+ `avgKNN`	1599 ± 678	0.300 ± 0.003	$\mathbf{11.46 \pm 0.09}$	$\mathbf{8.64 \pm 0.41}$	6.73 ± 0.03
+ `SigLIPMS`	2162 ± 936	$\mathbf{0.270 \pm 0.005}$	12.65 ± 0.29	9.14 ± 0.83	$\mathbf{7.03 \pm 0.04}$
+ `PerturbProj`	1703 ± 755	0.288 ± 0.004	12.38 ± 0.16	8.82 ± 0.56	6.91 ± 0.03

Table 9: As Table 8, with 50 000 images per configuration and one seed.

	time (s)	FD↓	FID↓	MIND↓	Vendi↑
Gender, $w_{fair} = 5$					
Debias Anything	11065	0.024	2.48	2.14	6.33
+ `SGMS`	15747	0.018	2.24	1.61	**6.55**
+ `avgKNN`	10942	0.015	**2.22**	1.63	6.50
+ `SigLIPMS`	15560	0.008	2.29	1.72	6.48
+ `PerturbProj`	11417	**0.002**	2.23	**1.59**	6.52
Eyeglasses, $w_{fair} = 200$					
Debias Anything	11120	0.311	7.50	8.78	6.69
+ `SGMS`	13829	0.307	7.84	8.42	6.88
+ `avgKNN`	10923	0.307	**7.47**	8.32	6.81
+ `SigLIPMS`	14712	**0.272**	8.45	8.79	**7.08**
+ `PerturbProj`	12431	0.293	8.13	**8.28**	7.00

On gender the gains are unambiguous. All four terms lower MIND from 2.52 to 2.03–2.13, two to three standard deviations, and improve the discrepancy, FID and Vendi at the same time. `PerturbProj` reaches the best discrepancy, 0.006 against 0.031, and the best MIND; the four terms are within one standard deviation of each other on MIND, so the choice between them is driven by the discrepancy. On eyeglasses the terms buy discrepancy ($0.304 \to 0.270$ for `SigLIPMS`, 0.288 for `PerturbProj`, both many standard deviations) and Vendi ($6.62 \to 7.03$), while the five MIND means, 8.64 to 9.14, all lie within one standard deviation (0.4 to 0.8) of one another: on this attribute the diversity terms cannot be ranked on MIND, and the two semantic terms do not improve it. FID rises from 11.47 to 12.4–12.7 for the two semantic terms and stays put for `avgKNN`. Since the reference is balanced on the attribute, a lower discrepancy alone would lower FID, so the rise is a genuine departure from the reference manifold: the semantic terms leave it, which is the price of the coverage they recover, whereas the density term, measured against real embeddings, stays on it.

Vendi, which needs no reference, moves the other way, from 6.62 to 7.03, and that recovered spread is the quantity the fairness term had taken away. The 50 000-sample run reproduces every one of these orderings; FID falls by a factor of two to three with the larger sample, as expected from its bias at small sample size (Chong & Forsyth, 2020), MIND falls by 0.3 to 0.6 and Vendi barely moves. Generation time is dominated by the sampler; in the 50,000-image runs, the diversity terms add up to approximately 43% overhead, with the largest increases for SGMS and `SigLIPMS`.

D.5 Diversity terms alone

The earliest runs measure each diversity term without the fairness term, against the 10 000 CelebA training images with the highest average distance to their five nearest neighbours in ResNet-50 feature space, following Sehwag et al. (2022): against this set a term scores well only if it reaches the rare part of the data. These runs predate the protocol of the rest of this appendix: $\sigma_{\max} = 20$, 1000 images per setting, prompt *"a photo of a man"*, and every metric below, MIND included, is against this rare subset and not against the balanced reference of the other sections.

Selecting each term at its lowest MIND would be a mistake. `SGMS` reaches MIND $= 27.6$ at $\sigma_s = 0.4$, $w_{\text{div}} = 200$ (Table 14), but its precision has fallen from 0.97 to 0.77 by then and the images at that setting carry visible artefacts: the term has left the model's manifold, and the metric, which only asks how close the samples are to the rare reference images, does not tell that apart from reaching the tail of the data. Table 10 therefore keeps, for each term, the setting with the lowest MIND among those whose precision stays at or above 0.95, a floor 0.02 below the unguided model. Under that floor `avgKNN` reaches the lowest MIND, 62.7, with the best coverage and Vendi; `SigLIPMS` comes second at 71.2 with the best density, `PerturbProj` third, and `SGMS`, the term with the largest recall gain, last at 101.0, because its recall is bought with precision, which the floor forbids.

Table 10: Each diversity term alone, at its lowest-MIND setting with precision ≥ 0.95 in Tables 11 to 14.

Method	Setting	MIND↓	Prec.↑	Rec.↑	Cov.↑	Dens.↑	Vendi↑
Original	—	124.10	0.972	0.389	0.809	3.010	5.94
`avgKNN`	$\kappa = 50,\ w_{\text{div}} = 30$	**62.71**	0.954	0.425	**0.874**	2.875	**7.33**
`SGMS`	$\sigma_s = 0.8,\ w_{\text{div}} = 40$	100.96	0.955	**0.468**	0.841	2.783	6.63
`PerturbProj`	$\sigma_s = 0.4,\ w_{\text{div}} = 250$	92.34	**0.958**	0.386	0.837	2.909	6.56
`SigLIPMS`	$w_{\text{div}} = 250$	71.17	0.953	0.332	0.865	**3.076**	6.96

Two things qualify this ranking. The measurement favours `avgKNN`: its score is a distance to a bank of SigLIP embeddings of real CelebA images (Section A.4), and the reference set is the rarest part of the same data, so the term is pushed towards the region it is then evaluated on, while the two semantic terms measure the model's own posterior spread and are given no such hint. And the ranking does not survive the fairness term. Once $w_{\text{fair}} > 0$ (Table 8) `PerturbProj` and `SigLIPMS` reach the lowest discrepancy on both attributes, `PerturbProj` the lowest MIND on gender and `SigLIPMS` the highest Vendi on eyeglasses, and `avgKNN` keeps only FID on eyeglasses. The single-term table selects a diversity score; the combined tables select a debiasing method, and the two disagree.

Three regularities hold across the four terms. `avgKNN` is nearly insensitive to its neighbourhood size: at its strongest weight, MIND varies by 4.6 across seven values of κ from 1 to 100, against a 46-point swing driven by the weight alone. For `SGMS` and `PerturbProj` the re-noising level is non-monotone, $\sigma_s = 0.2$ being markedly worse than 0.4, so the useful range is $\sigma_s \in [0.4, 0.6]$ for `SGMS`, while `PerturbProj` barely leaves the model's manifold at any setting (precision never below 0.956) and only $\sigma_s = 0.8$ gains recall. `SGMS` is the only term that converts precision into recall at scale in this setting, which is also why it is the one term whose low-MIND settings fail the precision floor of Table 10; `SigLIPMS` has a turning point: MIND bottoms out at $w_{\text{div}} = 250$ and rises beyond it while recall falls from its peak at 150, so weights above 250 are harmful.

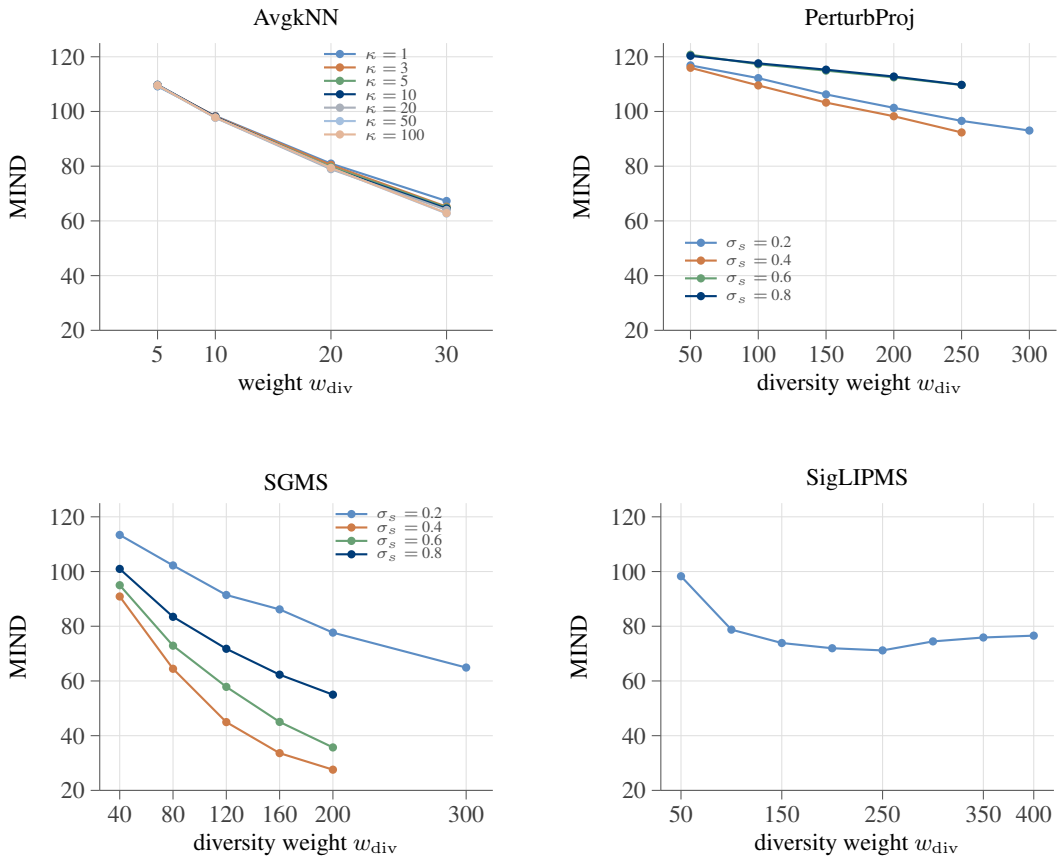

Figure 8: MIND against the diversity guidance weight w_{div}, one curve per hyperparameter configuration. All four panels share the MIND axis. `avgKNN` is nearly insensitive to κ over two orders of magnitude (seven superimposed curves, $\kappa = 1$ to $\kappa = 100$). For `SGMS` and `PerturbProj` the perturbation scale is non-monotone: $\sigma_s = 0.2$ is markedly *worse* than 0.4, so the useful range is $\sigma_s \in [0.4, 0.6]$. `SGMS` is the only method reaching MIND below 60, and it does so at a precision below 0.9 (Table 14). `SigLIPMS` turns around: MIND bottoms out at $w_{\text{div}} = 250$ and rises again beyond it.

Table 11: `avgKNN` alone, 1,000 images: precision, recall and MIND per (κ, w_{div}).

κ	w_{div}	Precision↑	Recall↑	MIND↓
1	5	**0.971**	0.383	109.19
	10	0.964	0.392	98.36
	20	0.955	0.412	80.95
	30	0.946	0.435	67.28
3	5	0.969	0.394	109.65
	10	0.963	0.403	98.21
	20	0.964	0.381	80.46
	30	0.944	0.433	65.19
5	5	0.969	0.391	109.53
	10	0.967	0.404	97.88
	20	0.960	0.397	79.73
	30	0.949	0.441	64.91
10	5	0.970	0.389	109.76
	10	0.965	0.403	98.07
	20	0.961	0.407	79.26
	30	0.949	**0.449**	64.53
20	5	0.969	0.386	109.71
	10	0.969	0.398	97.72
	20	0.960	0.412	78.91
	30	0.951	0.424	64.04
50	5	0.970	0.384	109.42
	10	0.965	0.407	97.72
	20	0.966	0.408	79.30
	30	0.954	0.425	**62.71**

Table 12: `SigLIPMS` alone, 1,000 images: precision, recall and MIND per w_{div}.

w_{div}	Precision↑	Recall↑	MIND↓
50	**0.965**	0.342	98.28
100	0.957	0.382	78.76
150	0.954	**0.405**	73.87
200	0.952	0.349	71.96
250	0.953	0.332	**71.17**
300	0.944	0.345	74.46
350	0.948	0.316	75.89
400	0.941	0.299	76.56

Table 13: `PerturbProj` alone, 1,000 images: precision, recall and MIND per $(\sigma_s, w_{\text{div}})$.

σ_s	w_{div}	Precision↑	Recall↑	MIND↓
0.2	50	0.970	0.364	116.85
	100	0.973	0.354	112.19
	150	0.967	0.336	106.24
	200	0.959	0.322	101.33
	250	0.965	0.330	96.56
	300	0.963	0.330	93.03
0.4	50	0.971	0.358	115.97
	100	0.970	0.379	109.52
	150	0.964	0.398	103.25
	200	0.956	0.374	98.26
	250	0.958	0.386	**92.34**
0.6	50	**0.974**	0.351	120.74
	100	0.972	0.350	117.26
	150	0.971	0.356	114.93
	200	0.972	0.374	112.46
	250	0.972	0.365	109.68
0.8	50	**0.974**	0.365	120.34
	100	**0.974**	0.389	117.60
	150	0.971	0.406	115.28
	200	0.968	0.416	112.77
	250	0.971	**0.421**	109.73

Table 14: `SGMS` alone, 1,000 images: precision, recall and MIND per $(\sigma_s, w_{\text{div}})$.

σ_s	w_{div}	Precision↑	Recall↑	MIND↓
0.2	40	**0.972**	0.390	113.38
	80	0.960	0.420	102.23
	120	0.939	0.410	91.43
	160	0.933	0.411	86.16
	200	0.921	0.425	77.67
	300	0.897	0.436	64.90
0.4	40	0.947	0.475	90.88
	80	0.916	0.508	64.45
	120	0.856	0.499	44.96
	160	0.820	0.481	33.61
	200	0.766	0.519	**27.57**
0.6	40	0.946	0.475	95.00
	80	0.923	0.530	72.89
	120	0.897	0.528	57.84
	160	0.859	0.540	45.02
	200	0.849	**0.567**	35.70
0.8	40	0.955	0.468	100.96
	80	0.938	0.460	83.45
	120	0.907	0.495	71.76
	160	0.874	0.529	62.30
	200	0.855	0.531	54.99

D.6 QUALITATIVE RESULTS

Every figure of this section shares its seeds across rows, so a column is one starting noise under several regimes and the panels can be read image by image. Together with Figure 1 they show three things: the guidance acts on individual images and leaves what it does not name, our method changes a face more cleanly than a labelled predictor, and the wording of the sentences matters little.

CelebA-HQ against Balancing Act (Figure 9). Six seeds, three of them women in the unguided row, under Balancing Act and under Debias Anything at matched discrepancy. Both methods turn the same two columns, the fifth and sixth, into men, and both leave the first three women and the man in the fourth column as they are, so the two batches end at the same proportion by the same choice of images. The difference is in the faces they produce: the man Balancing Act makes in the fifth column carries a blurred, uneven skin and a distorted mouth, while ours is a clean face with the seed's hair and lighting, and the sixth column shows the same gap in a milder form. That is the visible side of the FID difference in Table 2: at equal discrepancy, guidance through a labelled predictor pays in image quality where guidance through the adapter does not.

Sensitivity to the sentences (Figure 10). The same six seeds and the same weight $w_{\text{fair}} = 20$, with five different pairs of sentences naming gender: woman and man, female and male person, girl's and boy's portrait, lady and gentleman, and close-up photos of a woman and of a man. All five pairs turn the same two columns, the fifth and sixth, and leave the four others untouched; the five men in the fifth column differ only in fine detail, a lock of hair or the shape of the smile. The attribute the guidance reads is the direction between the two sentences, and that direction hardly moves with the wording, so an operator need not search for the right words.

Stable Diffusion (Figure 11). Six seeds each for *A face of a doctor* and *A face of a firefighter*, unguided and under Debias Anything at a gender target of one half. The unguided model draws five men and one woman for the doctor and six men for the firefighter, the skew of Table 3. Guidance turns two doctor columns and three firefighter columns into women, reaching three of six in each, and takes the images whose logit was already closest: the woman of the sixth doctor column is left alone. White coats and stethoscopes, helmets and turnout gear, the framing and the expression of each seed stay; only the face changes, and with it hair and, on two firefighters, a hood under the helmet. These are the samples behind the gender row of Table 3.



Figure 9: Gender on CelebA-HQ, six shared seeds: the unguided model, Balancing Act, and Debias Anything at matched discrepancy.



Figure 10: Sensitivity to the prompt pair, gender on CelebA-HQ, $w_{\text{fair}} = 20$, six shared seeds. Top: unguided. Then, one pair per row: *a photo of a woman* / *a photo of a man*; *a female person* / *a male person*; *a girl's portrait* / *a boy's portrait*; *a photo of a lady* / *a photo of a gentleman*; *a woman's close-up photo* / *a man's close-up photo*.



Figure 11: Stable Diffusion 1.5, prompts *A face of a doctor* (top pair) and *A face of a firefighter* (bottom pair), six shared seeds each: the unguided model and Debias Anything at a gender target of one half. Two doctor columns and three firefighter columns change gender; clothing, pose, background and expression stay.

E Full comparison under the DiffLens protocol

Table 16 completes Table 3 with the two CLIP scores of Section B.5, and the last paragraph of this section traces our residual discrepancy to its sources. CLIP-T lies within 0.01 of the unguided model for every method on every attribute, so none of them, ours included, trades prompt adherence for the attribute. CLIP-I separates the methods into two groups: the label-located methods (Shi et al., 2025; Zaleska et al., 2026) and fine-tuning (Shen et al., 2024) keep 0.85 to 0.92 of the original, the latent-space editors (Kwon et al., 2023; Li et al., 2024) 0.81 to 0.88, and ours sits at 0.84 to 0.85, the level of the h-space distribution guidance (Parihar et al., 2024). Race is the attribute on which DiffLens and ICM keep the most of the original image, above 0.90, and also the one on which they move the discrepancy least.

Where the residual discrepancy comes from. The FD of Table 3 is an average over the four occupations of the protocol, and Table 15 gives the four terms. On gender they are far apart: doctors reach 0.021 and firefighters 0.104, nurses stay at 0.296 and receptionists at 0.378. For a binary attribute FD $= \sqrt{2}\,|p - \frac{1}{2}|$ (Section B.5), so these are 51%, 57%, 71% and 77% women. The unguided model explains the split. Seshadri et al. (2024) measured the share of women that Stable Diffusion 1.4 draws for each occupation under four prompt templates: 100% for *nurse* under every template and 99.8 to 100% for *receptionist*, against 1 to 16% for *firefighter* and 4 to 58% for *doctor*. These shares account for the unguided row of Table 3: an all-female or all-male batch scores 0.71, and averaging the four templates gives per-occupation discrepancies of 0.71, 0.71, 0.63 and 0.32, whose mean of 0.59 is close to the 0.56 we measure on Stable Diffusion 1.5 with the prompt of the protocol. Reaching one half means creating male nurses and receptionists that the model never draws on its own, whereas female firefighters and doctors exist in its output already; the adapter, trained on the model's own images of these jobs, has seen just as few. Guidance moves nurses and receptionists from all women to three quarters, and stops there. The other guidance method of Table 3, the h-distribution guidance of Parihar et al. (2024), is trained on gender labels and stops at 0.222, above ours; the editing methods pass this point because they change what the model computes, whereas guidance reweights what it draws.

Table 15: Debias Anything on Stable Diffusion 1.5, discrepancy per occupation. Each entry is the distance of the mean FairFace output to the uniform target on the 500 images of one prompt; the last row is their average, the FD of Table 3.

Occupation	FD gender↓	FD age↓	FD race↓
doctor	0.021	0.092	0.208
firefighter	0.104	0.276	0.193
nurse	0.296	0.077	0.206
receptionist	0.378	0.381	0.174
mean (Table 3)	0.200	0.206	0.195

Table 16: Stable Diffusion 1.5 (Rombach et al., 2022) under the protocol of DiffLens (Shi et al., 2025), all four metrics. Baseline rows other than ICM (Zaleska et al., 2026) are as reported by Shi et al. (2025), which gives fine-tuning (Shen et al., 2024) a single FID, CLIP-I and CLIP-T for the three attributes. Guided rows are shaded from best (dark, bold) to worst (light); equal values share a shade. The unguided row has no CLIP-I by definition.

		Gender				Age				Race			
Method	Attribute	FD↓	FID↓	CLIP-I↑	CLIP-T↑	FD↓	FID↓	CLIP-I↑	CLIP-T↑	FD↓	FID↓	CLIP-I↑	CLIP-T↑
Original	—	0.564	120.06	–	0.6155	0.752	120.06	–	0.6155	0.558	120.06	–	0.6155
Feature or weight editing													
Latent editing (Kwon et al., 2023)	text, training per attribute	0.408	166.11	0.8253	0.6005	0.682	200.90	0.8527	0.6122	0.524	153.05	0.8804	0.6086
Latent direction (Li et al., 2024)	text, training per attribute	0.305	129.37	0.8058	0.6091	0.052	113.81	0.8151	0.6067	**0.175**	128.30	0.8211	0.6132
Fine-tuning (Shen et al., 2024)	attribute classifier	0.050	161.47	**0.8779**	**0.6095**	0.746	161.47	0.8779	0.6095	0.198	161.47	0.8779	0.6095
DiffLens (Shi et al., 2025)	labels	**0.046**	**112.83**	0.8501	0.6090	**0.049**	**99.17**	0.8778	0.6057	0.401	119.86	0.9096	0.6149
ICM (Zaleska et al., 2026)	labels	0.087	122.08	0.8501	0.6090	0.133	114.87	**0.9195**	**0.6150**	0.266	116.88	**0.9099**	**0.6172**
Guidance on the model's distribution													
h-distribution (Parihar et al., 2024)	labels	0.222	151.68	0.8475	0.6087	0.506	147.71	0.8345	0.6098	0.544	126.90	0.8255	0.6100
Debias Anything	text	0.200	113.85	0.8381	0.6090	0.206	110.55	0.8409	0.6060	0.195	**114.71**	0.8527	0.6111

On age the residual is concentrated in the same way. Doctors and nurses reach 0.092 and 0.077, firefighters and receptionists stay at 0.276 and 0.381. The unguided model puts nearly every face in one bin, its age discrepancy of 0.752 being close to the 0.82 of a single-bin batch, and its firefighters and receptionists are, as far as its samples show, middle-aged men and young women; a third of either at 60+ is far from what it draws on its own. A second effect may add a floor: FairFace counts 0–19, 20–59 and 60+, while the guidance is given *a young*, *a middle-aged* and *an old person*, and a face the encoder moves towards *young* is most likely in its twenties, which the counter files under 20–59. The two low values on doctor and nurse show that the sentences do reach the counted bins where the model can follow. On race the residual is even across occupations, 0.17 to 0.21, so it belongs to the attribute and not to any job; the four sentences name the four groups the counter uses, and ours is the lowest race discrepancy in Table 3.