Dissecting and Mitigating Diffusion Bias via Mechanistic Interpretability

Supplementary Material

A. Algorithm of DIFFLENS

In this section, we present a detailed explanation of the methodology behind our DIFFLENS. To enhance clarity, we outline the approach through two algorithms that comprehensively illustrate the key steps of DIFFLENS. In the context of the algorithms, "online" refers to performing the process during each image generation, while "offline" means executing it only once beforehand and using the obtained result directly during generation.

A.1. Dissecting Bias Mechanism

We introduce how to dissect bias mechanism, *i.e.*, disentangling bias features in the sparse semantic space (see in Sec. 4.1 for more details).

Algorithm 1: DIFFLENS: Dissecting Bias Mechanism

Input: A support set of samples $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}, \tau \in \mathbb{N}$

for $j = 1$ **to** N **do**
  Extract $\mathbf{h} = [h_1, h_2, \ldots, h_n]$ from $\epsilon_\theta(\mathbf{x}_j)$
  $\mathbf{s} = [s_1, s_2, \ldots, s_m] = \Phi(\mathbf{h})$; // Eq. (3)
  for $i = 1$ **to** m **do**
    $S(s_i; X)$ += $S(s_i; \mathbf{x}_j)$; // Eq. (6)
$\mathbf{A} = \{i_1, \ldots, i_\tau\} = \arg\operatorname{top}_\tau \{S(s_i; X) \mid s_i \in \mathbf{s}\}$;

Output: $\mathbf{A}$

We denote U-Net [50] as $\epsilon_\theta(\cdot)$ that accepts a sample $\mathbf{x}$ as input and we use it to extract hidden representations of the sample. To dissect bias mechanism, we leverage a support set of samples X consisting of N samples to identify bias feature. For each sample, the original hidden state $\mathbf{h}$ of the sample is extracted and transformed into the sparse semantic space using a k-SAE [35] through Eq. (3). Next, we localize the features that are related with bias contents by measuring the influence of these feature using a gradient-based bias attribution method as Eq. (6). For the calculation of attribution, we use Riemann approximation to estimate integral in the score of bias generation to the disentangled feature s_i as follows

$$S(s_i; \mathbf{x}) = (s_i - s_i') \cdot \frac{1}{q} \sum_{k=1}^{q} \frac{\partial F_{\mathbf{x}}(\mathbf{s}' + \frac{k}{q}(\mathbf{s} - \mathbf{s}'))}{\partial s_i}, \quad (8)$$

where $\mathbf{s}' = [s_1', \ldots, s_m']$ is a relative baseline of $\mathbf{s}$, q is the number of discrete steps or partitions used in the Riemann approximation to estimate the integral, and $\partial F_{\mathbf{x}}(\mathbf{s}' + \alpha(\mathbf{s} - \mathbf{s}'))/\partial s_i$ is the gradient of the bias measure given the input image $\mathbf{x}$, to the target feature space $\mathbf{s}$. The baseline s' can be (i) zero or (ii) a value tailored to a specific input. We use (i) for unconditional diffusion model P2 [10] and (ii) for conditional diffusion model Stable Diffusion [49]. Then we aggregate the attribution scores across all sample and every time steps. Finally, we set a threshold τ to the pick the top τ features that are highly related with specific bias content.

Note that, this process is required *only* once for one specific diffusion model, since the sparse sematic space and the features are consistent across all time steps during generation and generalizable within a model. As we utilize the same k-SAE [50] across all time steps, thereby reducing the number of parameters, the diffusion step t is not explicitly represented in Algorithm 1.

A.2. Bias Mitigation

We describe how to intervene in bias features identified within the latent space of a diffusion model in Algorithm 2. The process modifies specific elements of the semantic feature vector $\mathbf{s} = [s_1, s_2, \ldots, s_m] \in \mathbb{R}^m$ according to a set of indices $\mathbf{A} = \{i_1, \ldots, i_\tau\}$ specifying the subset of features in $\mathbf{s}$ corresponding to bias attributes to be adjusted.

Algorithm 2: DIFFLENS: Bias Mitigating

Input: A set of feature indexes $\mathbf{A} = \{i_1, \ldots, i_\tau\}$, hidden state extracted from $\epsilon_\theta(\mathbf{x})$ $\mathbf{h} = \{h_1, h_2, \ldots, h_n\} \in \mathbb{R}^n, \beta \in \mathbb{R}$

$\mathbf{s} = [s_1, s_2, \ldots, s_m] = \Phi(\mathbf{h})$; // Eq. (3)
Create $\mathbf{s}' = \{s_1', s_2', \ldots, s_m'\} = \mathbf{s}$
for $k = 1$ **to** τ **do**
  $s'_{i_k} = \begin{cases} \beta s'_{i_k} & \text{(Scaling), or} \\ s'_{i_k} + \beta & \text{(Adding)} \end{cases}$; // Eq. (7)
$\mathbf{h} = \mathbf{h} + \mathbf{W}_{\text{dec}}(\mathbf{s}' - \mathbf{s})$; // Eq. (9)

Output: $\mathbf{h}$

To map the intervened features back to the original hidden units, instead of directly reconstructing $\mathbf{h}$, we compute the difference between the intervened sparse semantic space $\mathbf{s}'$ and the original space $\mathbf{s}$, facilitating the mapping process. This approach reduces reliance on the reconstruction effect of k-SAE [35], as it only modifies the intervened parts while preserving the rest. We reference for [24] and derive the op-

eration as follows

$$\begin{aligned}\mathbf{h} &= \mathbf{h} + \hat{\mathbf{h}}' - \hat{\mathbf{h}} \\ &= \mathbf{h} + \mathbf{W}_{\text{dec}}\mathbf{s}' + \mathbf{b}_{\text{pre}} - (\mathbf{W}_{\text{dec}}\mathbf{s} + \mathbf{b}_{\text{pre}}) \\ &= \mathbf{h} + \mathbf{W}_{\text{dec}}(\mathbf{s}' - \mathbf{s}) \\ &= \mathbf{h} + \sum_{k=1}^{\tau}(s'_{i_k} - s_{i_k})f_{i_k},\end{aligned} \tag{9}$$

where $\hat{\mathbf{h}}$ is the reconstructed hidden space without intervention in Eq. (4), $\hat{\mathbf{h}}'$ is the mapped back of intervened hidden state and f_i is the i_{th} row of the decoder matrix $\mathbf{W}_{\text{dec}}$. The resulting update is applied to the hidden state $\mathbf{h}$, effectively reflecting the feature adjustments within the model's latent space.

By intervening on specific bias-related features, the algorithm enables controlled adjustments to the generated content. We also provide two approaches for intervening, which are scaling and adding. Since we adopt the same operation (scaling or adding) to hidden states $\mathbf{h}$ for all time steps, the time step t is not explicitly represented in Algorithm 2.

B. Implementation Details

In this section, we provide the details of how we implement DIFFLENS. To ensure consistency with the baselines used for comparison, we adopt the DDIM [55] in both diffusion models (Sec. 5.1) used in our experiment. It modifies the original reverse process by allowing for non-Markovian updates, which reduces the number of timesteps needed for sampling without sacrificing image quality. The update rule for generating $\mathbf{x}_{t-1}$ from $\mathbf{x}_t$ can be written as

$$\mathbf{x}_{t-1} = \sqrt{\alpha_{t-1}}\, P_t(\epsilon_t^\theta(\mathbf{x}_t)) + D_t(\epsilon_t^\theta(\mathbf{x}_t)) + \sigma_t z_t, \tag{10}$$

where $\alpha_t \in [0,1]$, $\sigma_t \geq 0, \sigma_t \in \mathbb{R}$ and $z_t \sim \mathcal{N}(0, I)$. Here, ϵ_t^θ is the predicted noise from the U-Net [50]. The intermediate terms are defined as:

$$\begin{aligned}P_t(\epsilon_t^\theta(\mathbf{x}_t)) &= \frac{\mathbf{x}_t - \sqrt{1-\alpha_t}\,\epsilon_t^\theta(\mathbf{x}_t)}{\sqrt{\alpha_t}}, \\ D_t(\epsilon_t^\theta(\mathbf{x}_t)) &= \sqrt{1-\alpha_{t-1}-\sigma_t^2}\,\epsilon_t^\theta(\mathbf{x}_t).\end{aligned} \tag{11}$$

In this work, based on the finding of interference between P and D from [32], we focus mainly on P_t, which corresponds to the prediction of the clean image at timestep t.

Additionally, we provide a strategy to intervene in the identified bias features for different classes within an attribute. This approach supports both general bias mitigation and prompt-specific bias mitigation. For attributes with more than two classes (*e.g.*, age or race), we intervene in the features of one class with a uniform probability $p = \frac{1}{k}$, where k is the number of classes, while keeping the features of the remaining classes unchanged (β = 1.0 for scaling or β = 0.0 for adding in Sec. 4.3). For instance, in mitigating age bias, after identifying feature indices for "young", "adult" and "old", we intervene in the features of one of these three classes with a probability $p = \frac{1}{3}$, while leaving the features of the other classes unchanged within a batch of samples. For bias mitigation in Stable Diffusion [49], the same strategy is applied to address varying levels of bias inherent to different prompts (*e.g.*, "doctor" being male-dominated versus "receptionist" being female-dominated).

We adopt the scaling operation to intervene in bias features for different attributes and specify the intervention parameter β. For gender attribute, β is set between 1.4 and 1.5. For the attribute, $\beta = 3.0$ is used for the old class, while $\beta = 1.0$ is applied to other classes. Similarly, for race attribute, $\beta = 5.0$ is applied to the black class, $\beta = 1.5$ for Asian and Indian, and $\beta = 1.0$ for white. When mitigating bias, such as in the case of age, we use a uniform probability of $\frac{1}{3}$ to select one class per batch (*e.g.*, old). The selected class is intervened with its respective β value (*e.g.*, $\beta = 3.0$ for old), while keeping the other classes with $\beta = 1.0$. This approach ensures precise control and flexibility across various attributes and classes.

B.1. Dataset Construction

We present details about constructing datasets using in (i) training k-SAE [35], (ii) training the light-weight classifier for our DIFFLENS and H-Distribution [43], and (iii) computing attribution score of bias generation. For all three parts, we leverage the diffusion models (mentioned in Sec. 5.1) to generate image examples as our dataset.

Dataset for Training K-SAEs. To construct the dataset for training k-SAEs, we leverage the unconditional diffusion model P2 [10] to generate 35,000 image examples. For text-to-image diffusion model Stable Diffusion [49], we use prompt "A face photo of a(an) class {occupation}" to generate 100,000 images for training. The "class" represents the categories in the gender, age and race attributes, *e.g.*, male. The occupations are sampled from a prompt pool which is presented in [40].

Dataset for Training Light-weight Classifier. For the training of the light-weight classifier in the unconditional diffusion model P2 [10], we use it to generate 1,000 images for each class in an attribute. We use FairFace [25] to determine the class for these generated images. In text-to-image diffusion model Stable Diffusion [49], we use prompts "a $\{class_1\}\{class_2\}\{class_3\}$ person", where the $class_i$ represent the categories in {gender, age, race} (*e.g.*, "a male old Indian person"). We generate 1,000 images for each class in each attribute.

Dataset for Identifying Bias Features. For the support set of samples X used in identifying target bias features in Sec. 4.2, we generate 1,000 image samples for each at-

k	FID ↓	CLIP-I ↑
32	6.86	0.9516
64	5.22	0.9695
128	**3.71**	**0.9795**

Table 4. Reconstruction effects w.r.t. different choices of k, where k represents the number of activated features in k-SAE [35]. We use non-reconstructed (original) outputs as the reference dataset for calculating both FID and CLIP-I metrics.

tribute (gender, age and race) in both the unconditional and text-to-image diffusion model. We use the trained light-weight classifier to discriminate the hidden space representations of these samples for specific classes in an attribute.

B.2. Training K-SAE on Hidden Space

In Eqs. (3) and (4), the k-SAE contains an encoder $\mathbf{W}_{\text{enc}}$ and a decoder $\mathbf{W}_{\text{dec}}$ with the same initialization of parameters. For backpropagation through TopK operation, we use straight through estimator. We use the DDIM [55] to obtain the middle block representations in the U-Net [50] for each image sample introduced in Sec. 4.2. The loss we use is the reconstruction error introduced in Eq. (5). The dimension m of the sparse representation space is set to 4096 in P2 [10] and 5120 in Stable Diffusion[49]. We train k-SAE in P2 [10] using 2 $\times$ RTX2080Ti GPU and in Stable Diffusion [49] using 2 $\times$ A100 GPU.

Reconstruction Effect with K-SAE. We present the effect of the reconstruction by using k-SAE in Tab. 4. The P2 [10] model is used to compare outputs with and without reconstruction and the non-reconstructed outputs is used as a reference dataset for both FID and CLIP-I metrics. As we can see in Tab. 4, as k goes larger, we obtain better reconstruction effect with the decreasing of FID and increasing CLIP-I. However, the reconstruction is not the key factor, as we can see in Tab. 5, there is a trade-off between k and our efficacy of bias mitigation. We provide visual comparisons of reconstruction quality for P2 [10] and Stable Diffusion [49] in Figs. 7 and 8. Images reconstructed using k-SAE [35] are almost indistinguishable from those generated by the original diffusion model P2 [10]. For Stable Diffusion [49], using prompts detailed in Appendix C.1, most semantic features are well reconstructed, although certain elements, such as the background, may not be fully preserved. This discrepancy is likely due to the richer semantic content of Stable Diffusion [49] compared to P2 [10]. Importantly, since our DIFFLENS does not heavily depend on the reconstruction quality of k-SAE [35], its performance in bias mitigation remains unaffected.

Training Results. We present the details about the training of k-SAEs [50], including hyper-parameters and learning curves. We train these on the hidden space without conditioning on the time step t, using the same model across all time steps. This approach significantly reduces the parameter size. We use learning rates of 0.001 and 0.005, with batch sizes of 8 and 100, for P2 [10] and Stable Diffusion [49], respectively. For the training curve, we leverage Fraction of Variance Unexplained (FVU) [35] which is a related metric of interest, measuring the total amount of the original activation that is not "explained" or reconstructed well by k-SAE [35]. We present it as below

$$FVU = \frac{\mathcal{L}(\mathbf{h})}{\text{var}[\mathbf{h}]}, \tag{12}$$

where $\mathbf{h}$ is the hidden state, $\mathcal{L}(\mathbf{h})$ is defined in Eq. (5) and var represents the variance of $\mathbf{h}$. A lower FVU indicates better reconstruction performance, as more of the original activation is captured by the k-SAE model. The training curves in Fig. 9 show that the k-SAE models for both P2 [10] and Stable Diffusion [49] are trained to perform well on the FVU metric.

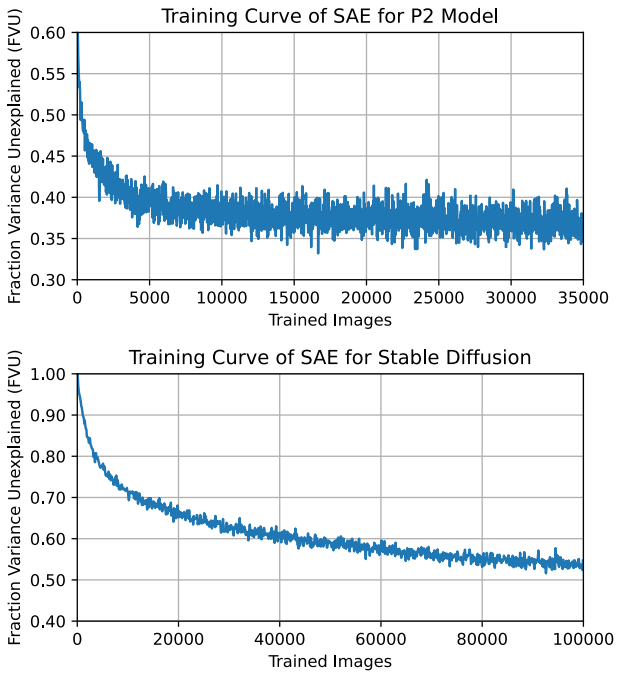

Figure 9. Training curves of SAE for P2 Model and Stable Diffusion, showing the Fraction of Variance Unexplained (FVU) metric over the number of training images.

B.3. Training Light-weight Classifier

To accurately locate target features within the hidden space of diffusion models, we train a classifier to identify which feature are more closely related to categories within a given attribute (*e.g.*, male and female for gender attribute) using the attribution method described in Sec. 4.2. Here we give the details about how to train such a classifier. After obtaining the dataset in Appendix B.1, we follow the setting

Figure 7. Comparison between images generated by original P2 model (top) and using k-SAE (bottom). The reconstructed and original images are almost identical, indicating effective reconstruction quality.

Figure 8. Comparison between images generated by original Stable Diffusion model (top) and using k-SAE (bottom). We observe minimal differences between the reconstructed and original images, indicating effective reconstruction quality.

of [43] for training. Then we obtain the middle block hidden representations in the U-Net [50] through DDIM [55] with respective to all time steps. The classifiers $C_a^h(\mathbf{h}_t, t)$ are trained as linear heads over the obtained $\mathbf{h}_t$, conditioned on time t and with respective to attribute a. We train the classifier in both diffusion models using one RTX2080Ti GPU. As shown in Fig. 10, the classifiers for P2 [10] and Stable Diffusion [49] achieve high accuracy across all three attributes: gender, age, and race. Specifically, the gender attribute comprises 2 classes, age consists of 3 classes, and race includes 4 classes. Notably, as the time steps approach the clean image state (closer to 1), the classifier accuracy consistently improves.

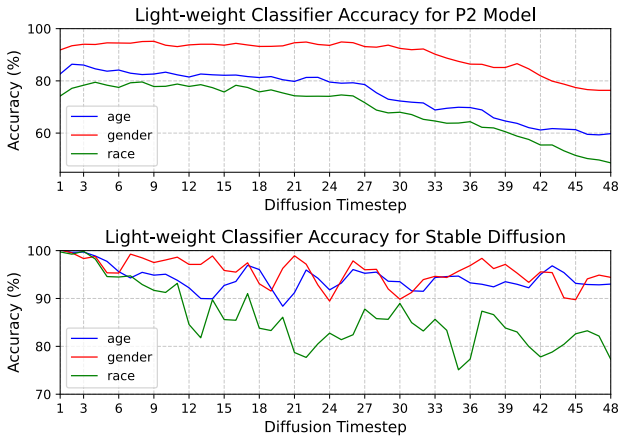

Figure 10. Light-weight classifier accuracy for P2 Model and Stable Diffusion for all three attributes (gender, age and race). We follow [43] and delete time time step 49. We use 1,000 images for each class in each attribute when training.

C. Experimental Settings

C.1. Evaluation Pipeline

In this section, we describe the overall steps for evaluation based on metrics in Sec. 5.1. In unconditional diffusion model P2 [10], we generate 10,000 images for each attribute, including gender, age, race, for each method. We use FairFace [25] as the classifier for different classes in all three attributes. In text-conditional diffusion model Stable Diffusion [49], we use four prompts for evaluation. The prompt used is "a face of an {occupation}" where the "occupation" includes doctor, firefighter, nurse, and receptionist and two of which are male-biased and two female-biased. The first two is based on [43] and the latter two prompts are from [40] which suggest bias for female and also exhibit age and race biases. For every prompts, we generate 500 images for each method. We will introduce how we measure FD, FID and CLIP-I/T in the subsequent sections.

C.2. Evaluation Metric

Fairness Discrepancy (FD). For evaluating bias mitigation, we use the metric FD in [43], and we provide details as fol-

lows. To measure the fairness of generated images with respect to a particular attribute a. Given a well-trained classifier for attribute a, denoted as C_a, the FD score is calculated as

$$\|\bar{\mathbf{p}} - \mathbb{E}_{\mathbf{x}\sim p_\theta(\mathbf{x})}[\mathbf{y}]\|_2 \,, \tag{13}$$

where $\bar{\mathbf{p}}$ is the target distribution over the attribute classes, which can be uniform, representing the ideal fair distribution, $\mathbf{y}$ is the softmax output of the classifier $C_a(\mathbf{x})$ for the generated sample $\mathbf{x}$, and $p_\theta(\mathbf{x})$ is the distribution of the generated images. For P2 [10], we use the result of all images (10,000) in each method. For Stable Diffusion [49], we average the results of four prompts, each with 500 images.

Fréchet Inception Distance (FID). The Fréchet Inception Distance (FID) [22] is a metric commonly used to assess the quality and diversity of generated images in comparison to real images. Given two sets of images—one generated and one real—the FID measures the distance between the feature distributions of these two sets, as extracted by a pre-trained Inception network. Let $\mathcal{N}(\mu_r, \Sigma_r)$ and $\mathcal{N}(\mu_g, \Sigma_g)$ represent the multivariate Gaussian distributions of real and generated images, respectively, where μ_r and Σ_r are the mean and covariance of the real image features, and μ_g and Σ_g are the mean and covariance of the generated image features. The FID is defined as:

$$\text{FID} = \|\mu_r - \mu_g\|^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\right), \tag{14}$$

where $\|\mu_r - \mu_g\|^2$ measures the squared difference between the means of the distributions, and $\text{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\right)$ measures the difference in covariances.

A lower FID score indicates a closer match between the distributions of real and generated images. The reference dataset (real images) we use for unconditional diffusion model is CelebA-HQ dataset [26], while for text-to-image diffusion models is the FFHQ [27]. When calculating FID for Stable Diffusion [49], we evaluate images in all four prompts (2,000) simultaneously for each method.

CLIP-I and CLIP-T. We follow [54] to measure the similarity between originally generated images and images after bias mitigation. We use two metrics based on CLIP embeddings which are CLIP-I and CLIP-T scores. The CLIP-I score compares the similarity between the original image embedding $\mathbf{e}_{\text{img}}^{\text{orig}}$ and the debiased image embedding $\mathbf{e}_{\text{img}}^{\text{gen}}$, both extracted using the CLIP model. Formally, we compute the cosine similarity between $\mathbf{e}_{\text{img}}^{\text{orig}}$ and $\mathbf{e}_{\text{img}}^{\text{gen}}$ as

$$\text{CLIP-I} = \frac{\mathbf{e}_{\text{img}}^{\text{orig}} \cdot \mathbf{e}_{\text{img}}^{\text{gen}}}{\|\mathbf{e}_{\text{img}}^{\text{orig}}\|\|\mathbf{e}_{\text{img}}^{\text{gen}}\|}. \tag{15}$$

This score reflects the preservation of visual features after bias mitigation.

The CLIP-T score, on the other hand, measures the alignment between the debiased image and a text prompt. Given a text prompt embedding $\mathbf{e}_{\text{text}}$ and the debiased image embedding $\mathbf{e}_{\text{img}}^{\text{gen}}$, the CLIP-T score is calculated as

$$\text{CLIP-T} = \frac{\mathbf{e}_{\text{text}} \cdot \mathbf{e}_{\text{img}}^{\text{gen}}}{\|\mathbf{e}_{\text{text}}\|\|\mathbf{e}_{\text{img}}^{\text{gen}}\|}. \tag{16}$$

The CLIP model we use is CLIP ViT-L/14 [48]. Note that, when we compare CLIP-I/T scores, we should also take FD score into account. The reason is that if one method is defective in bias mitigation, most of the time, it will not alter the generated images much, resulting in a high CLIP-I/T score. Therefore, we mainly consider methods that are effective in bias mitigation when comparing CLIP-I/T score. For CLIP-I in P2 [10], we use the average result of every generated images (10,000) in each method. For Stable Diffusion [49], we use the average result of the prompts in each method (*i.e.*, calculating average CLIP-I/T for each image with a prompt and then average within four prompts).

C.3. Baselines

In this section, we provide details of the implementation of baseline methods along with specific information on each approach. For all baselines, we strictly follow their setting and record their results. All baselines except for Fintuning [54], are based on the bottleneck layer of the U-Net [50] in diffusion models.

Activation is based on previous work on interpreting GPT-2 using GPT-4 employing internal activation to analyze individual features [6]. It suggests that activation in internal neurons contains meaningful information. In addition, [12] shows the capability to edit internal neurons to affect the performance of a neural network. Based on these, we adopt a similar setting where feature editing is performed directly on the original activation in the latent space. For a fair comparison, we constrain it on the bottleneck layer of the U-Net [50].

Latent Editing, as described in [32], they learn a latent vector to steer unbiased image generation through the bottleneck layer of the U-Net [50] model. We adopt their approach to learn latent vectors for each class (*e.g.*, "old") across different attributes (*e.g.*, "age"). To mitigate bias, we apply linear scaling to the learned vectors and incorporate them into the original bottleneck layer as described in their methodology.

H-Distribution, introduced by [43], employs distributional loss on bottleneck layer as guidance in diffusion models. We directly use the h-classifier for gender attribute they provided, while we train the multi-class h-classifier for age and gender, following the methods introduced in the paper.

Latent Direction, introduced by [33], identifies interpretable semantic directions within the latent space of text-to-image diffusion models, specifically in the U-Net [50]

bottleneck layer. This method optimizes concept-specific latent vectors by reconstructing images that exclude certain features in the text prompt while leveraging the pre-trained model's semantic knowledge. We learn concept vectors for gender (male and female), age (young, adult, and old), and race (white, black, Asian, and Indian) following their setting. These vectors are then combined with equal probabilities to mitigate bias.

Finetuning, introduced by [54], uses distributional alignment loss to adjust generated images to align with a user-defined target distribution. This approach integrates pre-trained classifiers to estimate class probabilities and fine-tune Stable Diffusion [49]. The released fine-tuned model is utilized for generation.

D. More Results

D.1. Sensitivity Analysis of DIFFLENS

In sensitivity analysis, we follow the experiment settings in Sec. 5.5, which aims at bias mitigation, *i.e.*, generating balanced outputs. We show that our DIFFLENS is relatively stable within different hyper-parameters.

K-SAE Hyper-parameters. We adjust different k values to evaluate the effects of the k-SAEs [50] on bias mitigation as shown in Tab. 5. As k increases, it is harder to achieve a balanced outputs, though the reconstruction effect may be better. We conclude that out method does not heavily rely on how well we reconstruct the images (see in Algorithm 2 at Appendix A), instead, how well the semantic features are disentangled and intervened plays a more important role. We choose k with 32 showing the best performance in bias mitigation.

Bias Attribution Parameters. We experiment on different number of target semantic features τ to be located. The results are summarized in Tab. 6 where we selected $\tau = 30$ features as the incorporated setting in our experiments, balancing both low FID of 31.93 and high CLIP-I score of 0.9479 in debiasing gender attribute. We observe that selecting too few features for bias mitigation risks omitting critical attributes related to biased content. Conversely, selecting a greater number of features facilitates better generation quality and semantic coherence, as evidenced by improvements in FID and CLIP-I metrics.

D.2. Visual Results on Conditional Diffusion Model

We provide the visual results mentioned in Sec. 5.2 for Stable Diffusion [49]. We include results for the prompt "A face of a firefighter" for illustration. We use male-to-female ratio to measure how well the effect of bias mitigation is achieved for each method. As shown in Fig. 11, our DIFFLENS effectively mitigates bias while preserving generation quality. In contrast, Latent Editing [32] and H-Distribution [43] struggle to produce balanced outputs and generate distorted images. While Finetuning [54] achieves high-quality images, it faces challenges in achieving well-balanced generation.

k	Gender (2)		
	FD ↓	FID ↓	CLIP-I ↑
Original	0.226	33.38	-
32	**0.002**	**31.93**	0.9479
64	0.005	31.94	**0.9524**
128	0.008	32.24	0.9113

Table 5. Impact of different TopK parameters in k-SAE on debiasing gender attribute. TopK means preserve the k features while deactivating the rest in the sparse semantic space (see in Sec. 4.1).

τ	Gender (2)		
	FD ↓	FID ↓	CLIP-I ↑
Original	0.226	33.38	-
10	0.003	33.01	0.9446
20	**0.001**	32.94	0.9466
30	0.002	**31.93**	**0.9479**

Table 6. Evaluation of bias mitigation for gender attribute w.r.t. various choices of feature number τ, where τ means the number of identified target features in Sec. 4.2. We base on our choice of $k = 32$ for this evaluation.

D.3. Visual Results for Ablation Study

We provide qualitative results for our ablation study in Sec. 5.5. As shown in Fig. 16, directly selecting neurons with the highest activation values as bias features (Activation method "Act@L") performs poorly, supporting the claim in Sec. 4.1. Selecting features (disentangled by k-SAE [35]) according to activation value performs even worse (see in "Act@S"). Comparing "Attr@L" with DIFFLENS, although "Attr@L" has effect in debiasing, it suffers distorted outputs while DIFFLENS are able to achieve excellent performance in debiasing and preserve image quality, illustrating the importance of disentangling neurons for better control.

D.4. Case Study

D.4.1 Accurate Attribution of Bias Features

We provide more examples with our DIFFLENS corresponding to Sec. 5.3. Figure 12 provides visual examples for comparison with baseline methods.

Following the same settings for generating a male-to-female ratio of 7:3 as described in Sec. 5.3, our DIFFLENS effectively preserves semantic features unrelated to the target attribute (gender), such as facial expressions, eyeglasses, background, and race.

In contrast, H-Distribution [43] may unintentionally alter non-target attributes, such as race. For instance, in the

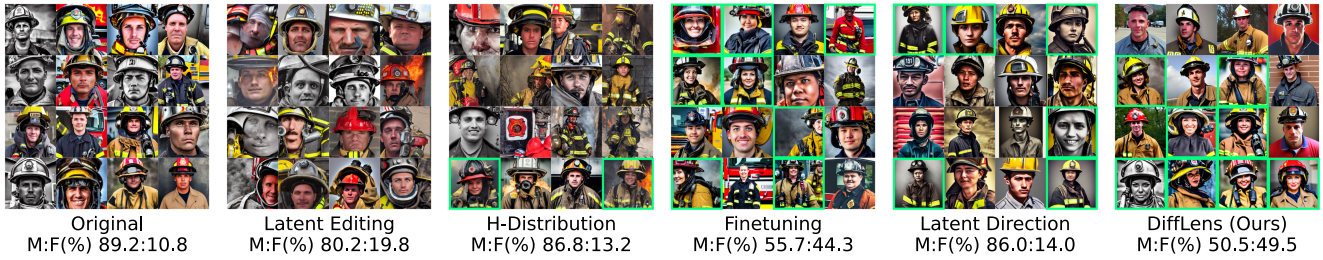

Figure 11. Comparison of randomly sampled original and debiased images generated by Stable Diffusion of the "firefighter" occupation. Various baseline debiasing methods are compared. The minority group (female) is highlighted with green bounding boxes for easier viewing. "M:F" refers to the male-to-female ratio.

Figure 12. Comparison in accurate identification of bias features. Our DIFFLENS preserves overall image semantics such as smile and eyeglasses while other methods frequently introduce distortions or lose important details.

left example of the first row and the middle example of the last row, the generated images exhibit changes in racial attributes. Additionally, the background in their generated images may vary significantly, as seen in the middle and right examples of the fourth row. Latent Editing [32], on the other hand, may generates distorted or unrealistic images, such as the left example from the last row. It also tends to entangle attributes like age and gender, as demonstrated in the left example of the second row, where a male image appears to incorporate old-age features. For original images that are male, in the target 7:3 male-to-female ratio, our DIFFLENS can maintain these images almost un-

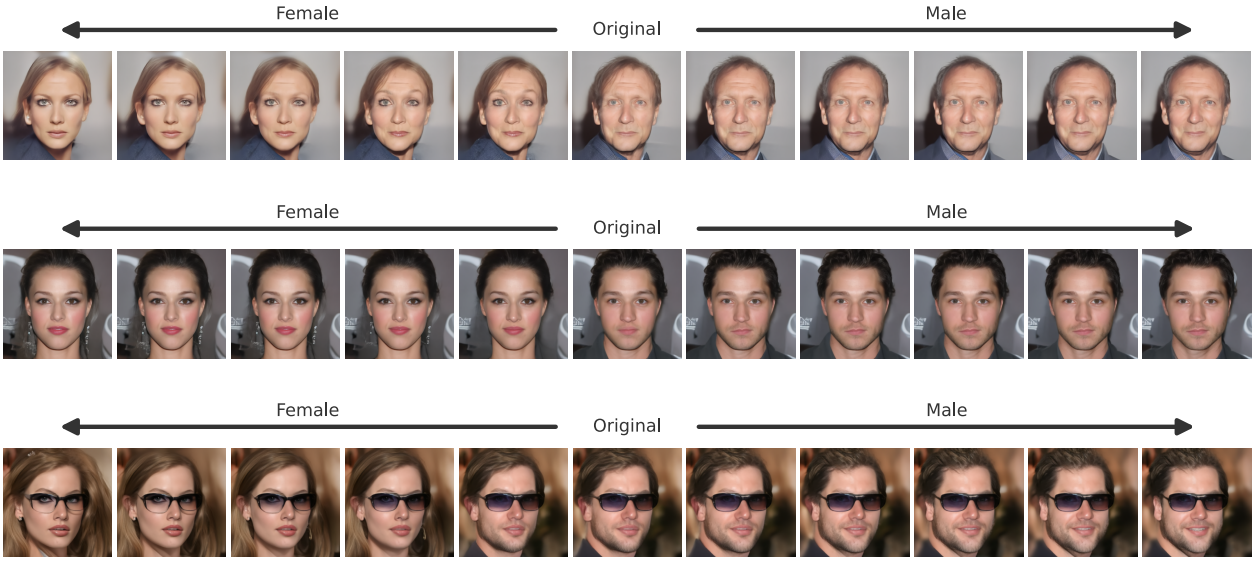

Figure 13. Scaling up "female" or "male" features. We adopt the same settings as outlined in Sec. 5.4. Our DIFFLENS demonstrates smooth editing and highlights its ability to exert fine-grained control over bias levels.

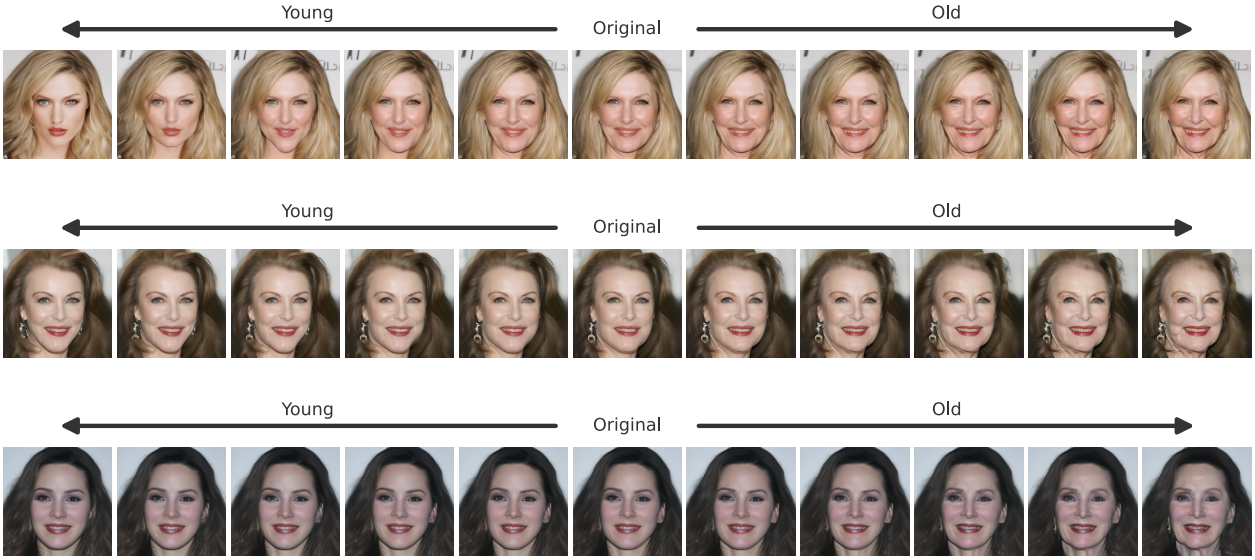

Figure 14. Scaling up "young" or "old" features. We adopt the same settings as outlined in Sec. 5.4. Our DIFFLENS demonstrates smooth editing and highlights its ability to exert fine-grained control over bias levels.

changed (e.g., the middle example in the fourth row and the right example in the second row in Fig. 12), aligning with the desired ratio. In contrast, other methods may inadvertently alter the original images or transform them into female representations, failing to preserve the intended male attributes.

D.4.2 Fine-grained Control and Editing

More examples showing our control over bias level with finer granularity (see in Sec. 5.4) are provided within this section. The settings outlined in Sec. 5.4 are applied, where generated images are randomly sampled across a broad spectrum of ratios (*e.g.*, male-to-female and young-to-old).

We release the results of how we transform the original image along two gender directions in Fig. 13. As we can see in Fig. 13, when editing towards male or female, we preserve generation quality and the visual feature coherence such as eyeglasses and expressions. We also shows the control over two age directions which are young and old in Fig. 14. Along the two directions of editing, the hair style

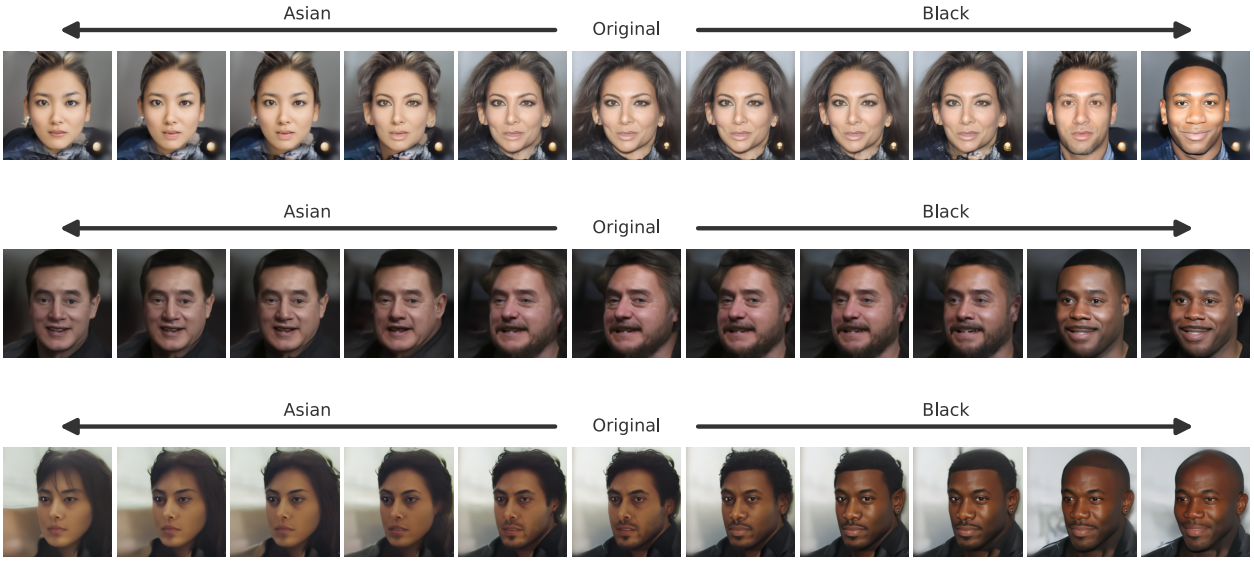

Figure 15. Scaling up "Asian" or "Black" features. We adopt the same settings as outlined in Sec. 5.4. Our DIFFLENS demonstrates smooth editing and highlights its ability to exert fine-grained control over bias levels.

Figure 16. Qualitative results in ablation study of DIFFLENS in P2 [10] for gender attribute. We abbreviate Original as "Orig.", neuron activations as "Act.", and bias attributions as "Attr". "@L" denotes operations on the original latent space, and "@S" on sparse semantic space.

and expressions are preserved in all three examples, even for earrings (middle example). In Fig. 15, we show the results of editing race along Asian and black direction. We are able to achieve a successful editing in both two directions. Additionally, for racial attributes, distinct hairstyles are often observed, such as short hair that is commonly associated with black individuals in Fig. 15.

D.4.3 Fine-grained Control across Multi-attribute

Addressing social biases across multiple attributes such as gender and age requires a comprehensive approach that ensures fairness without compromising image quality. Our approach enables fine-grained control over *multiple* attributes *simultaneously*, allowing for unbiased and consistent outputs across diverse settings.

In this section, we aim at demonstrating that our DIFFLENS are able to mitigate bias with multi-attribute (*e.g.*, gender and age) rather than only one attribute (*e.g.*, gender). In addition, we also illustrate the ability to control over the bias level within multi-attribute. Specifically, we identify the bias feature indexes for the target multi-attributes (details provided in Appendix A). These features are then simultaneously intervened (*e.g.*, male and old features). Visual results presented in Fig. 17 illustrate that DIFFLENS effectively transforms images in the male and old directions without overlap (*i.e.*, the male and old directions do not interfere with each other). This demonstrates the disentanglement achieved within the sparse semantic space, as discussed in Sec. 4.1. We do not compare other baselines except for H-Distribution because in [43], they mention that Latent Editing [32] are unable to mitigate bias in case of multiple attributes.

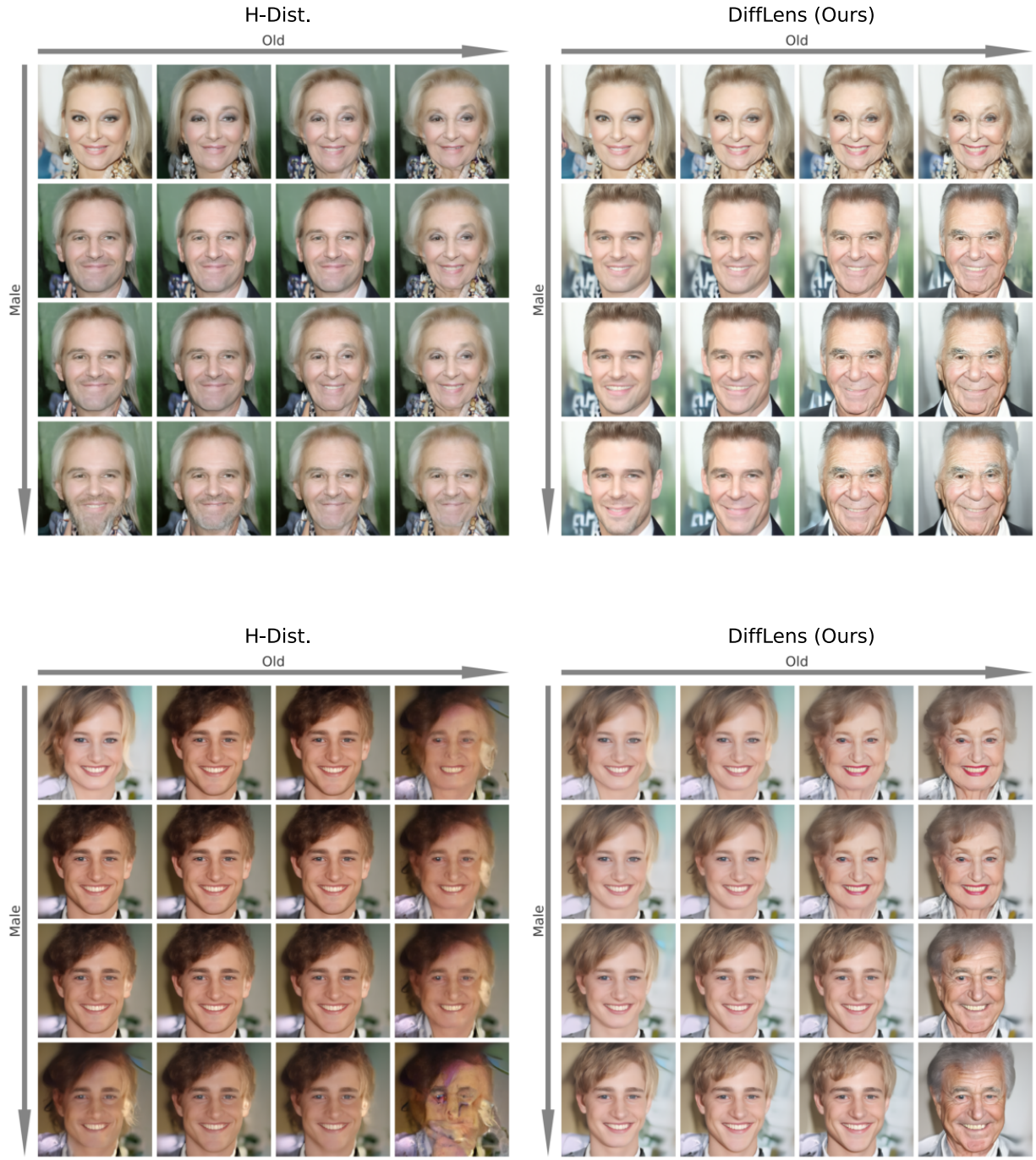

Figure 17. Scaling up "old" and "male" features at the same time. We try to control over the bias level across multi-attribute simultaneously rather than a single attribute, illustrating our ability to disentangle different bias features and accurate identification of these features. We provide two examples for control over age and gender attributes.

D.4.4 Control of Other Bias Mechanisms

Our DIFFLENS can also disentangle gender attribute with other features. We conduct additional experiments and present the results in Fig. 18. As illustrated in the figure, we can independently control gender attributes (male and female) along with features in our wide investigation in Sec. 5.6 such as "Side Pose", "Short Hair" and "Smile".

Figure 18. Decoupling and independently controlling gender and other features. We select three features presented in Sec. 5.6 that are "Side Pose", "Short Hair" and "Smile".

E. Further Discussions

E.1. Complexity of Social Bias

Defining social bias in AI systems is inherently fraught with complexities due to the fluid, multidimensional nature of social attributes such as gender and ethnicity, which resist discrete categorization and are shaped by sociocultural contexts [3]. Unlike measurable technical metrics, biases in systems like text-to-image models often reflecting historical inequities or stereotyping patterns rather than explicit labels. For instance, synthetic depictions of fictive humans lack inherent social identities, forcing evaluators to rely on subjective interpretations of visual features (*e.g.*, skin tone or hairstyles) that may not align with real-world self-identification.

The relative instability effects for race attributes in Tabs. 1 and 2 may possibly due to the complexity of social bias and the evaluation is inherently challenging [34], which requires more robust assessment methods. Additionally, these attributes in dataset may be imbalanced, making model learning unstable. However, our method balances debiasing effect (FD [43]) and generation quality (FID [22] and CLIP-score used in [54]).

E.2. Monosemanticity in Sparse Autoencoder

A monosemantic feature corresponds to one individual concept recognized by the model, in contrast to polysemantic neurons associating multiple unrelated concepts [4]. While an SAE automatically disentangles neuron spaces [35], its semantic space may not fully align with human-interpretable concepts (see Sec. 3.1 of [4]).

According to [35], the used k-SAE encourages feature orthogonality, helps disentangle neuron space towards monosemanticity. It inherently supports disentangled feature learning [35]. It is empirically supported by low pairwise cosine similarity among the learned SAE features in DIFFLENS with mean value of 0.04 and maximum value 0.17, which somehow indicates dissimilarity and disentanglement in sparse feature space.

Bias attribute, *e.g.*, gender, is a human-defined compound concept, may relate to multiple features discovered by SAEs, for example, short-hair and mustache features. Our method identifies these associations and can further separate them (as shown in Fig. 18). Features capture aspects of our target concepts, but they may not fully represent the concepts themselves. To better align these features with our interpretable concepts is a valuable future direction.

E.3. Quantitative Metric for Changes along One Direction

Our work explicitly measures specific attribute balancing (FD [43]) and overall generation quality (FID [22] and

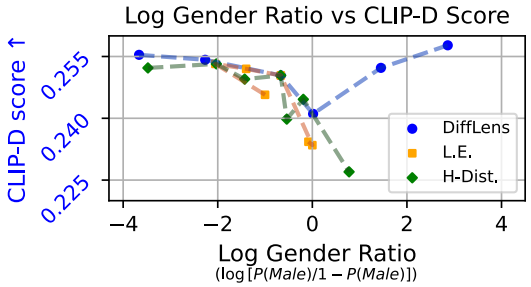

Figure 19. Alignment (CLIP-D score) to specific attribute changing across gender ratio (x-axis). The Log Gender Ratio reflects the log of male to female ratio in the generated images, with 0 indicating balance. Our DIFFLENS offers better alignment with each direction.

CLIP-score that is used in [54]). To further assess specific attribute changes, we introduce

$$\text{CLIP-D} = \frac{\mathbf{e}_{\text{attr}} \cdot \mathbf{e}_{\text{img}}^{\text{gen}}}{\|\mathbf{e}_{\text{attr}}\| \|\mathbf{e}_{\text{img}}^{\text{gen}}\|}, \tag{17}$$

where $\mathbf{e}_{\text{attr}}$ is the target attribute text (e.g., "male") embedding and $\mathbf{e}_{\text{img}}^{\text{gen}}$ is the generated image embedding. Figure 19 shows that our DIFFLENS achieves better alignment to specific attribute change direction (higher CLIP-D) across varying gender bias ratios.

E.4. Ethics Discussion

In this work, we try to address fairness in both unconditional and conditional diffusion models by proposing a framework that identifies and isolates bias mechanisms and control bias levels in generated contents. While our method does not prescribe a universal definition of fairness because ethical interpretations may vary across contexts. Our method should enable practitioners to enforce distributions considered appropriate for their applications.

All experiments were conducted using publicly available datasets and pre-trained models that are permitted for academic research. Our study prioritizes transparency in methodology and outcomes. We urge future researchers to critically evaluate the societal implications of their chosen distributions to mitigate unintended harms.

Our approach treats debiased attributes as discrete categories, thereby overlooking individuals who do not neatly fit traditional classifications (*e.g.*, those identifying as non-binary gender or of mixed race). This is a significant research question and needs to be addressed by future work.