RESEARCH-ARTICLE

Bimodal Debiasing for Text-to-Image Diffusion: Adaptive Guidance in Textual and Visual Spaces

LIU YU, University of Auckland, Auckland, AUK, New Zealand

JIAJUN SUN, University of Electronic Science and Technology of China, Chengdu, Sichuan, China

PING KUANG, University of Electronic Science and Technology of China, Chengdu, Sichuan, China

RUI ZHOU, University of Electronic Science and Technology of China, Chengdu, Sichuan, China

FAN ZHOU, University of Electronic Science and Technology of China, Chengdu, Sichuan, China

ZHIKUN FENG, University of Electronic Science and Technology of China, Chengdu, Sichuan, China

Open Access Support provided by:

University of Electronic Science and Technology of China

University of Auckland

Published: 27 October 2025

Citation in BibTeX format

MM '25: The 33rd ACM International Conference on Multimedia
October 27 - 31, 2025
Dublin, Ireland

Conference Sponsors:
SIGMM

https://doi.org/10.1145/3746027.3754901
ISBN: 9798400720352

Bimodal Debiasing for Text-to-Image Diffusion: Adaptive Guidance in Textual and Visual Spaces

Liu Yu*
liu.yu@std.uestc.edu.cn
University of Electronic Science and Technology of China
Chengdu, Sichuan, China
The University of Auckland
Auckland, New Zealand

Jiajun Sun*
async@std.uestc.edu.cn
University of Electronic Science and Technology of China
Chengdu, Sichuan, China

Ping Kuang†
kuangping@uestc.edu.cn
University of Electronic Science and Technology of China
Chengdu, Sichuan, China

Rui Zhou†
ruizhou@uestc.edu.cn
University of Electronic Science and Technology of China
Chengdu, Sichuan, China

Fan Zhou
fan.zhou@uestc.edu.cn
University of Electronic Science and Technology of China
Chengdu, Sichuan, China

Zhikun Feng†
202411090917@uestc.edu.cn
University of Electronic Science and Technology of China
Chengdu, Sichuan, China

Abstract

Social biases in text-to-image models have drawn increasing attention, yet existing debiasing efforts often focus solely on either the textual (e.g., CLIP) or visual (e.g., U-Net) space. This unimodal perspective introduces two major challenges: **(i)** Debiasing only the textual space fails to control visual outputs, often leading to pseudo- or over-corrections due to unaddressed visual biases during denoising; **(ii)** Debiasing only the visual space can cause modality conflicts when biases in textual and vision are misaligned, degrading the quality and consistency of generated images.

To address these issues, we propose a **B**imodal **AD**aptive **G**uidance D**E**biasing within Textual and Visual Spaces (**BADGE**). First, BADGE-quantifies attribute-level bias inclination in both modalities, providing precise guidance for subsequent mitigation. Second, to avoid pseudo/over-correction and modality conflicts, the quantified bias degree is used as the debiasing strength for adaptive guidance, enabling fine-grained correction tailored to discrete attribute concepts. Extensive experiments demonstrate that BADGE significantly enhances fairness across intra- and inter-category attributes (e.g., gender, skin tone, age, and their interaction) while preserving high image fidelity. *Our project page is at https://badgediffusion.github.io/

CCS Concepts

• **Security and privacy** → **Social aspects of security and privacy**; • **Social and professional topics** → **User characteristics**.

Keywords

Social fairness, text-to-image, stable diffusion, image diversity

ACM Reference Format:
Liu Yu, Jiajun Sun, Ping Kuang, Rui Zhou, Fan Zhou, and Zhikun Feng. 2025. Bimodal Debiasing for Text-to-Image Diffusion: Adaptive Guidance in Textual and Visual Spaces. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746027.3754901

*Both authors contributed equally to this research.
†Corresponding author.

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3754901

1 Introduction

The growing interest in vision-language models, such as CLIP [41] and diffusion models (DMs) [23, 35, 42, 47, 71] has heralded the era of AI-generated content [9, 27, 28, 70], enabling the application of text-to-images across a diverse range of fields [12, 13, 25, 37, 50, 54]. However, these models often replicate and amplify social biases present in their training data [2, 4, 5, 36, 43, 63–65], disproportionately representing certain groups and marginalizing others. For instance, when prompted with "A photo of a doctor", over 90% of generated images depict men, revealing a strong gender bias (1a).

Substantial progress has been made in debiasing text-to-image models. ***(1)*** One approach [18, 19, 66] *debiases the textual embedding solely*. Take the projection method VL-Debias [18] as an example, while it ensures the balanced distribution across attributes, its synthesized images (in Fig. 1b) exhibit the over-correction deficits , where previously underrepresented groups become overrepresented. ***(2)*** Another line of work [20, 40] *focuses solely on applying the guidance in the latent visual space* (Fig. 1c), if the biases in the textual and visual spaces are misaligned, the guidance process may lead to conflicts, potentially degrading the quality and consistency of synthesized images. For example, in text embedding from CLIP, the "doctor" is semantically close to "female", and attempting attribute guidance for "male" in U-Net can lead to failed guidance and showing a low-quality image that is neither clearly male nor female (gender-ambiguous, synthesizing a woman with a beard).

In response to these challenges, we argue that debiasing solely in textual or visual space is insufficient for fair generation and propose that group diversity needs to simultaneously satisfy unbiasedness

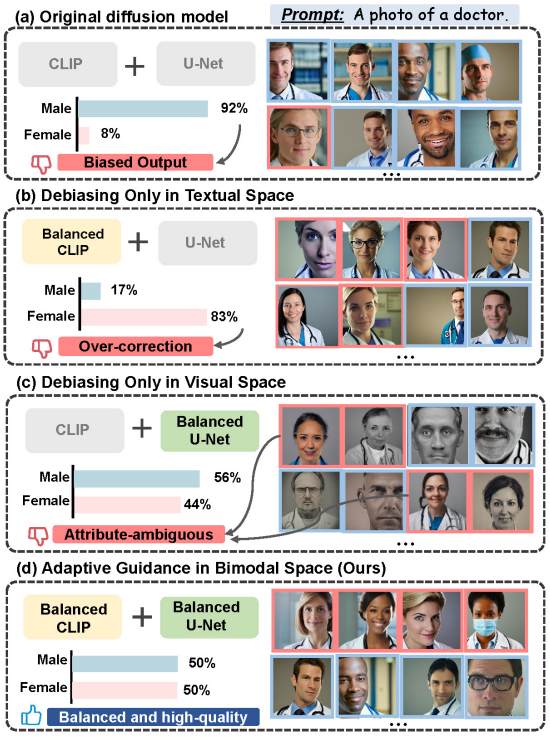

Figure 1: Given a prompt with a neutral profession ("A photo of a doctor"): (a) Original DMs [44] produce biased outputs. (b) Textual-only debiasing [18] leads to pseudo/over-correction. (c) Vision-only debiasing [20] synthesizes low-quality, attribute-ambiguous images. (d) Our bimodal adaptive guidance produce balanced and high-quality results.

in both text and visual spaces, in Fig. (1d). Observing that neither method specifically addresses the biases within the model internally, we raise the question: *can we locate and quantify the biases in both text and visual space, and then use them as guidance?*

To achieve this goal, we introduce (**BADGE**) **B**imodal **AD**aptive **G**uidance D**E**biasing that manipulates biases in both textual and visual spaces for fair text-to-image generation. Concretely, BADGE follows a two-stage pipeline: *(i)* it first locate and quantify bias inclination towards various professions within each modal space, respectively. *(ii)* during inference, BADGE applies adaptive guidance in bimodal space based on the degree of bias inclination as a debiasing strength, controlling the generation towards our desired attribute direction. This process allows for *fine-grained control* over intra- and inter- category attribute correction and offers unbiased text embedding and visual noisy state, avoiding pseudo/over-correction or cross-modal conflicts. The extensive experiments demonstrate that the consideration of bimodal bias facilitates comprehensive bias mitigation throughout the generation process, and achieves considerable group diversity and fidelity across multiple biases.

In summary, our work offers three contributions:

- To the best of our knowledge, BADGE is the first work to address the conflict of biases between textual and visual spaces for fair text-to-image generation.
- We propose adaptive guidance for debiasing based on the quantified degree of bias inclination, which is to steer attributes towards our desired balanced direction.
- Our BADGE is simple and efficient, requiring no extra training costs or corpora, which only directly manipulate the textual and visual spaces during the inference process.

2 Related Works

Text-only Debiasing in Text-to-Image Generation. Many works in this direction [6, 7, 18, 19, 39, 55, 58, 66, 69] mitigate the unfair biases from solely text embedding in CLIP. Some try to add a soft prompt outside the backbone [21, 32, 46], which includes a linear projection layer that covers diversity [51] or adds inclusive tokens [66]. The drawback is that they require training based on external corpora and designing balanced sampling strategies. Another needs no training but applies uniform debiasing toward any profession without distinguishing different attributes. Take VL-Debias [18] as a typical method. It achieves debiasing by projecting the prompt embedding into the orthogonal space of , making the embedding orthogonal to the binary-sensitive attributes. However, the generated images sometimes display pseudo- or over-correction (*cf.* Fig. 1 (b)), as fair text condition fails to control visual space.

Vision-only Debiasing in Text-to-Image Generation. This line of work [1, 3, 14, 22, 29, 40, 48, 56, 57, 60] enhances diversity by directly applying balance guidance in the latent diffusion space, without considering text conditions or measuring bias inclination in both spaces. For example, Balanced Act [40] correct bias from the *h*-space [33] in U-Net, while Fair Diffusion [20] uses semantic guidance within U-Net. They provide effective guidance even when text conditions are insufficient [17, 38, 59, 62]. However, a significant drawback is that biases in visual space can conflict with implicit biases in the text conditions, leading to modality conflicts and degraded image quality [26, 61], as shown in Fig. 1(c).

Although existing methods show promising results, they often fail to fully address the inherent biases in text-to-image models, as they focus only on CLIP or U-Net. A recent work [49] fine-tunes both components for debiasing, but it requires additional training. We address these limitations by jointly considering bias in both textual and visual spaces. In particular, our proposed BADGE framework does not require additional data or parameter updates.

3 Our Debiasing Framework BADGE

Problem Statement: Let a set $\mathbf{A}$ denote M *inter-categories* of sensitive attributes (e.g., gender, age, skin tone), each *intra-category* contains binary or multiple attributes (i.e., (*female*, *male*)). Let $a_{m,k}$ represent the k-th specific-attribute in the m-th category (e.g., $m = 1$ is the gender case; $a_{1,1} = male$, $a_{1,2} = female$): $\mathbf{A} = \{\mathcal{A}_m \mid 1 \leq m \leq M\}$; $\mathcal{A}_m = \{a_{m,k} \mid 1 \leq k \leq K_m\}$, K_m denotes attributes number in $\mathcal{A}_m$. When given a textual user-input prompt c^i including the neutral profession word i, the diffusion model [24, 45] aims to learn a conditional distribution:

$$\boldsymbol{e}^{\boldsymbol{i}} = f(c^i) \tag{1}$$

$$P\left(\mathbf{z}^i \mid \boldsymbol{e}^i\right) = \prod_{t=1}^{T} P\left(\mathbf{z}_{t-1}^i \mid \mathbf{z}_t^i, t, \boldsymbol{e}^i\right) \tag{2}$$

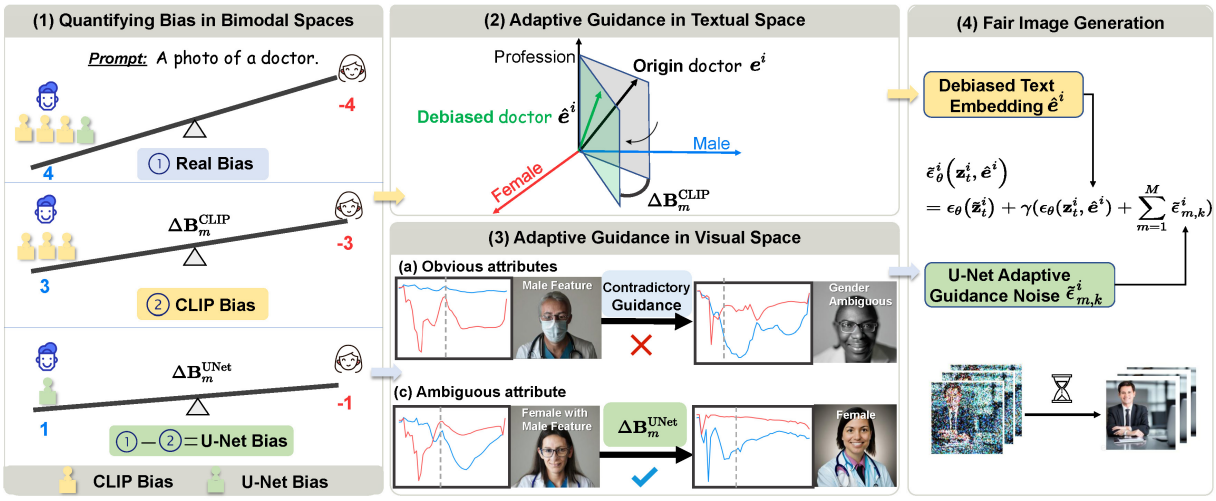

Figure 2: The overall framework of BADGE. (1) We firstly obtain the quantified bias inclination in bimodal spaces (*Insight.1* and *Insight.2*); (2) BADGE employs adaptive guidance based on quantified CLIP bias degree for obtaining a debiased text embedding (*Insight.3*); (3) Based on quantified U-Net bias degree, BADGE adaptively guide the samples with ambiguous attribute for obtaining a unbiased noise (*Insight.4*); (4) Fair image generation is achieved by removing the estimated noise from the latent noisy state based on unbiased textual embedding.

where $\boldsymbol{e}^i$ is the embedding of user-input prompt c^i from the CLIP text encoder, and $\mathbf{z}_t^i$ is the visual representation of U-Net at time step t. The inherent bias in the training datasets of the CLIP and U-Net models can distort the distribution P, leading to a pronounced bias toward specific sensitive attributes. Prior works only consider bias encoded in the text semantic space in Eq. (1) or only in the latent diffusion space in Eq. (2). However, our work highlights that biases arise in both bimodal spaces, and that these biases constrain each other, leading to suboptimal debiasing results. Exploring the synergistic debiasing in bimodal spaces is an important issue.

Task Objective: Given a diffusion model $\mathcal{M}$ with potential social bias and a human-written neutral prompt c^i (e.g., "a photo of a doctor"), we aim to obtain a fair generative model $\mathcal{M}'$ that generates diverse images ensuring each sensitive attribute $a_{m,k}^i \in \mathcal{A}_m$ has an equal probability. Alternatively, fair prediction can be formalized as approaching a discrete uniform distribution:

$$\hat{P}_{\mathcal{M}'} \sim \mathcal{U}\{0, \frac{1}{|\mathcal{A}_m|}\} \tag{3}$$

Next, we proceed to our methodology. For each component, we initiate our analysis by formulating a fundamental **Research Question** (RQ), followed by presenting our key theoretical insights (denoted as *Insight 1-4*) that address the question, which subsequently guide the corresponding technical details. Note that BADGE is generic to various biases and professions, with the *gender* bias and *doctor* profession serving as just our example in this section.

The overall framework of BADGE is shown in Fig. 2.

3.1 Quantifying Bias in Bimodal Space

RQ1: How can we systematically locate and quantify biases embedded in both textual and visual spaces? This foundational inquiry is critical as precise bias measurement establishes the basis for analyzing cross-modal bias conflicts, and informs the design of targeted joint debiasing strategies.

Insight.1 *Attribute-specific prompts help measure bias inclination in textual space (i.e., CLIP). Removing CLIP's bias from synthesized image distribution exposes biases in the visual space (i.e., U-Net).*

Bias Quantification from the Textual Space. Let i denote the index of i-th professional word, for a given user-input original prompt c^i (e.g., "a photo of a doctor") with a neutral professional word w^i ("doctor") and the sensitive attribute set $\mathbf{A}$, we execute a pre-processing step to obtain the attribute-specific prompt via:

$$c_{m,k}^i = \Phi(a_{m,k}, w^i \mid c^i)$$

where Φ denotes the enumeration function – "a photo of a $[a_{m,k}]$ $[w^i]$". For example, $c_{1,1}^1$ is "a photo of a [male] [doctor]" and $c_{1,2}^1$ is "a photo of a [female] [doctor]" when given c_1 as "a photo of a doctor". All attribute-specific prompts are formed as a set:

$$\mathbf{C} = \{C_m \mid 1 \le m \le M\}; C_m = \{c_{m,k}^i \mid 1 \le k \le K_m\}.$$

Accordingly, we obtain all attribute-specific text embeddings $\mathbf{E}$ from the CLIP text encoder:

$$\mathbf{E} = \{\mathcal{E}_m \mid 1 \le m \le M\}; \mathcal{E}_m = \{\boldsymbol{e}_{m,k}^i \mid 1 \le k \le K_m\}.$$

Then, we quantify the bias inclination of a particular profession w^i within the textual space. This is achieved by computing the embedding similarity between the neutral prompt $\boldsymbol{e}^i$ and attribute-specific prompts $\boldsymbol{e}_{m,k}^i$:

$$s_{m,k}^i = \frac{\exp(\text{sim}(\boldsymbol{e}_{m,k}^i, \boldsymbol{e}^i))}{\sum_{j=1}^{K_m} \exp(\text{sim}(\boldsymbol{e}_{m,j}^i, \boldsymbol{e}^i))} \tag{4}$$

$$\text{sim}(\boldsymbol{v}_1, \boldsymbol{v}_2) = \frac{\boldsymbol{v}_1 \cdot \boldsymbol{v}_2}{\|\boldsymbol{v}_1\| \|\boldsymbol{v}_2\|} \tag{5}$$

where the $\text{sim}(\cdot) \in [-1, 1]$, closer to 1 indicating a higher similarity, and $s^i_{m,k} \in [0, 1]$, where values approaching 1 indicate that the particular profession w^i has a stronger inclination toward attribute k. Finally, the attribute-specific biases of profession w^i from the CLIP text encoder are quantified as follows:

$$\mathbf{B}^{i,CLIP}_m = \{s^i_{m,k} \mid 1 \le k \le K_m\} \tag{6}$$

where $\mathbf{B}^{i,CLIP}_m$ denotes the set of biases on m-th inter-category attribute for the i-th profession word in CLIP space.

Bias Quantification from the Visual Space. As proposed in Insight 1, subtracting the bias present in CLIP from the distribution of synthesized images can reveal the biases within the U-Net model. The straight way to obtain the distribution of synthesized images is using an attribute classifier [18, 67]. However, it requires (a) training dedicated attribute classifiers, and (b) generating large-scale image batches – both imposing computational costs.

RQ2: How can we develop an efficient bias quantification framework that circumvents the computational overhead?

Insight.2 *The semantic trajectory of predicted noise during diffusion denoising enables* ***early attribute prediction****, bypassing the need for complete image generation and explicit classification models.*

Thanks to semantics inherent in U-Net latent space [8] , we propose *early attributes prediction* to accelerate attribute acquisition at an iterative denoising step without fully generating images. Specifically, the training objective of the diffusion model on conditioned on $\boldsymbol{e}^i$ from CLIP text encoder is:

$$\mathbb{E}_{\boldsymbol{e}^i, \epsilon \sim \mathcal{N}(\mathbf{0},\mathbf{I}), t}\left[\|\epsilon - \epsilon_\theta(\mathbf{z}^i_t, t, \boldsymbol{e}^i)\|^2_2\right] \tag{7}$$

where $\mathbf{z}^i_t$ is the latent noisy state at each time step t. To determine the most likely attribute during the denoising process, we compare the neutral prompt noisy state $\mathbf{z}^i_t$ with the attribute-specific noisy state $\mathbf{z}_t^{i,a_{m,k}}$ via:

$$h(\mathbf{z}^i_t) = \begin{cases} a_{m,p}, & \text{if } \text{sim}(\mathbf{z}^i_t, \mathbf{z}_t^{i,a_{m,p}}) > \tau > \text{sim}(\mathbf{z}^i_t, \mathbf{z}_t^{i,a_{m,q}}) \\ \varnothing, & \text{Otherwise} \end{cases} \tag{8}$$

where τ is the pre-defined threshold, and $0 < p < q < k$. This attribute decision process in Eq. (8) is repeated for each time step t from T to 0. Once the attribute's similarity exceeds τ the decision process is terminated, and the attribute of the current image is extracted. Based on the predictable attributes determined above, we obtain the real bias from the conditional DM via:

$$\mathbf{B}^{i,Real}_m = \{P_{m,k} | 1 \le k \le K_m\} \tag{9}$$
$$P_{m,k} = \mathbb{E}_{\mathbf{z}^i_t \sim \hat{P}_\mathcal{M}}\left[\mathbb{1}_{h(\mathbf{z}^i_t)=a_{m,k}}\right]$$

where $P_{m,k}$ is estimated by calculating the proportion of each predictable attribute $a_{m,k}$. Finally, the discrepancy of bias inclination between the synthesizing image and the CLIP text encoder is considered as bias from the U-Net model:

$$\mathbf{B}^{i,Unet}_m = \mathbf{B}^{i,Real}_m - \mathbf{B}^{i,CLIP}_m \tag{10}$$

where if $\mathbf{B}^{i,Unet}_m \to \mathbf{0}$, indicating that the bias distribution of synthesizing images on i-th profession matches that from the text encoder, signifying that bias arises from the CLIP model. Conversely, if $\mathbf{B}^{i,Unet}_m$ deviates significantly from $\mathbf{0}$, it suggests the presence of external bias in the U-Net model of i-th profession.

3.2 Adaptive Guidance Debiasing in Bimodal Space

Adaptive Debiasing in Textual Space. Different professions have different attribute bias inclination in textual space, such as "doctor" tending towards *males* and "nurses" tending towards *females*. Prior textual-only debiasing strategies [18, 19] that do not distinguish the bias inclination during projection, but only blur sensitive attributes by applying a uniform debiasing strength towards different professions. However, an inappropriate choice of correction strength can lead to pseudo-debiasing or over-correction (see Fig. (1 b)).

RQ3: Based on the located and quantified bias in the above, how can we effectively mitigate the bias in the textual space?

Insight.3 *Adaptive guidance based on quantified bias degree avoids bias inclination towards specific attributes.*

By comparing the quantified bias with a fair distribution, we can determine the bias inclination via:

$$\Delta\mathbf{B}^{i,CLIP}_m = \mathbf{B}^{i,CLIP}_m - 1/|\mathcal{A}_m| \tag{11}$$

where $\mathbf{B}^{i,CLIP}_m$ is the quantified bias in Eq. (6), and $1/|\mathcal{A}_m|$ denotes uniform distributions on m-th inter-category. Eq. (11) is the debiasing strength for subsequent adaptive guidance.

A fair textual space should satisfy the criterion that the embedding of the debiased input prompts $\hat{\boldsymbol{e}}^i$ maintain an equivalent similarity to each attribute-specific prompt $\boldsymbol{e}^i_{m,k}$. Instead of the limited addressing biases in binary attributes [18], our BADGE extends to handle biases across multiple attributes. The adaptive guidance optimum for obtaining a debiased text embedding $\hat{\boldsymbol{e}}^i$ is:

$$\begin{aligned} \mathcal{L}_{\text{bias}} = & \frac{\mathcal{H}(\Delta\mathbf{B}^{i,CLIP}_m)}{\sum_{m=1}^{M}|A_m|}\sum_{m=1}^{M}\sum_{0<p<q<K_m}(\boldsymbol{e}^{iT}\boldsymbol{e}^i{}_{m,p} - \boldsymbol{e}^{iT}\boldsymbol{e}^i{}_{m,q})^2 \\ & - \lambda(\boldsymbol{e}^{iT}\boldsymbol{e}^i{}_{m,p} + \boldsymbol{e}^{iT}\boldsymbol{e}^i{}_{m,q}) \\ \mathcal{H}(\Delta\mathbf{B}^{i,CLIP}_m) := & \max\{\Delta\mathbf{B}^{i,CLIP}_m\} \times 1000 \times \alpha \end{aligned} \tag{12}$$

where $\mathcal{H}(\cdot)$ controls the debiasing strength of adaptive guidance, and α is a pre-defined hyper-parameters. $\mathcal{L}_{\text{bias}}$ enables the combination of attributes across M inter-categories, and each intra-category contains K_m attributes, encompassing binary or multiple attributes. The overall optimization goal of the CLIP model is as follows:

$$\begin{aligned} \mathcal{L}_{i,CLIP} &= \mathcal{L}_{\text{bias}} + \mathcal{L}_{\text{sem}} \\ \mathcal{L}_{\text{sem}} &= \|\hat{\boldsymbol{e}}^i - \boldsymbol{e}^i\|^2 \end{aligned} \tag{13}$$

where $\mathcal{L}_{\text{sem}}$ used to preserve the similarity between the debiased text embedding $\hat{\boldsymbol{e}}^i$ and original embedding $\boldsymbol{e}^i$, thus maintaining the original semantic information. Finally, we derive the closed-form solution for Eq. (13):

$$\begin{aligned} \hat{\boldsymbol{e}}^i = & \left(I + \frac{\mathcal{H}(\Delta\mathbf{B}^{i,CLIP}_m)}{\sum_{m=1}^{M}|A_m|}\sum_{m=1}^{M}\sum_{0<p<q<K_m}\boldsymbol{e}^i_{\text{diff}}\boldsymbol{e}^{iT}_{\text{diff}}\right)^{-1} \\ & \times (\lambda\boldsymbol{e}^i_{\text{same}} + \boldsymbol{e}^i) \end{aligned} \tag{14}$$

where I is the identity matrix, and $\boldsymbol{e}^i_{\text{diff}} = \boldsymbol{e}^i_{m,p} - \boldsymbol{e}^i_{m,q}$, $\boldsymbol{e}^i_{\text{same}} = \boldsymbol{e}^i_{m,p} + \boldsymbol{e}^i_{m,q}$ for brevity.

Adaptive Debiasing the Visual Space. While textual space achieves a balance across sensitive attributes, this manipulation is primarily restricted to the static textual space. In contrast, the dynamic visual

space presents additional challenges, as the random noise at different denoising steps encodes distinct attribute-related concepts. Consequently, despite achieving fairness in the text-conditioned setting, this balance does not necessarily translate to the final synthesized images due to the uncertain attributes. Therefore, it is crucial to jointly consider the visual space even under a balanced textual condition.

However, (a) not all samples are suitable for guidance, and (b) the success rate of guidance is not high at any time steps. Existing method [20] overlook the semantics inherently embedded in the noise, which can lead to ineffective or even contradictory guidance. For instance, if the initial noise distribution exhibits a bias toward *female* attributes, applying guidance to enforce a *male* attribute may result in gender ambiguity (as illustrated in Fig. (1c)).

RQ4: Which samples are suitable for adaptive guidance, and when is it effective to conduct the guidance? ***Insight.4*** *Images with ambiguous attribute tendencies during the early stage are easier to guide toward a specified attribute in the visual space, whereas those with obvious attributes do not require guidance and should be generated naturally.*

Specifically, based on the debiased CLIP text embedding $\hat{\boldsymbol{e}}^i$ as a condition, the objective of DM can be rewritten as:

$$\mathbb{E}_{\hat{\boldsymbol{e}}^i,\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I}),t}\left[\|\epsilon-\epsilon_\theta^i(\mathbf{z}_t^i,t,\hat{\boldsymbol{e}}^i)\|_2^2\right] \tag{15}$$

where $\epsilon\sim\mathcal{N}(0,\mathcal{I})$ sampled from a Gaussian distribution. The DM takes unbiased prompt vectors $\hat{\boldsymbol{e}}^i$, the time step t, and the latent noisy state $\mathbf{z}_t^i$ as input to predict the noise ϵ_θ^i. During the training process, the condition $\hat{\boldsymbol{e}}^i$ is randomly replaced with null values with a fixed probability, resulting in a joint model that accommodates both conditional and unconditional scenarios. The semantic and arithmetic principles [8] of noise space enable us to extract the attribute concept of synthesizing images via:

$$\epsilon_{m,k}^i=\epsilon_\theta^i\left(\mathbf{z}_t^i,\boldsymbol{e}_{m,k}^i\right)-\epsilon_\theta^i(\tilde{\mathbf{z}}_t^i) \tag{16}$$

To apply the adaptive guidance in the visual space using attribute concepts $\epsilon_{m,k}^i$, we offer ***attribute prominence analysis*** based on noisy latent:

$$\begin{aligned}\mu_{m,k}=&\max_k\int_{t_1}^{t_2}\mathrm{sim}(\mathbf{z}_t,\mathbf{z}_t^{i,a_{m,p}})\,dt\\&-\min_k\int_{t_1}^{t_2}\mathrm{sim}(\mathbf{z}_t,\mathbf{z}_t^{i,a_{m,q}})\,dt\end{aligned} \tag{17}$$

where the larger $\mu_{m,k}$ is, the higher the prominence of the attribute. Next, based on the measured bias inclination $\Delta\mathbf{B}_m^{i,UNet}$ and attribute prominence $\mu_{m,k}$, we distinguish the predicted time-varying noise with attribute concept into *obvious* and *ambiguous* two types via:

$$\Delta\mathbf{B}_m^{i,UNet}=\mathbf{B}_m^{i,UNet}-1/|\mathcal{A}_m| \tag{18}$$

$$\tilde{\epsilon}_{m,k}^i=\begin{cases}\mathcal{G}(\Delta\mathbf{B}_{m,k}^{i,UNet})\cdot\epsilon_{m,k}^i, & \mu_{m,k}\le\eta\\ \mathbf{0}, & \mu_{m,k}>\eta\end{cases} \tag{19}$$

where η is the pre-set threshold. The adaptive function $\mathcal{G}(\Delta\mathbf{B}_m^{i,UNet}):=\max\left\{|\Delta\mathbf{B}_m^{i,UNet}|\right\}\times\beta$ enables *fine-grained* debiasing over discrete attributes, meaning it can modulate the debiasing strength β proportionally to the measured bias degree, rather than applying a fixed correction. If $\mu_{m,k}>\eta$, denote noises with *obvious* attributes, which are difficult to guide toward other attributes. In this case, the guidance condition is set to $\mathbf{0}$, indicating that no guidance is applied, and the generation proceeds naturally. If $\mu_{m,k}\le\eta$, noises with *ambiguous* attributes, which are more easily guided toward any attribute. We primarily focus on guiding this type of samples, and the unbiased noise $\tilde{\epsilon}_{m,k}^i$ is used to synthesize images via:

$$\tilde{\epsilon}_\theta^i\left(\mathbf{z}_t^i,\hat{\boldsymbol{e}}^i\right)=\epsilon_\theta\left(\tilde{\mathbf{z}}_t^i\right)+\gamma(\epsilon_\theta(\mathbf{z}_t^i,\hat{\boldsymbol{e}}^i)+\sum_{m=1}^{M}\tilde{\epsilon}_{m,k}^i) \tag{20}$$

where γ is the guidance scale. Finally, by removing the estimated noise from the latent noisy state, we obtain the final debiased synthesized image:

$$\begin{aligned}\mathbf{z}_{t-1}^i&=\mathrm{SchedulerStep}\left(\mathbf{z}_t^i,\tilde{\epsilon}_\theta^i\left(\mathbf{z}_t^i,\hat{\boldsymbol{e}}^i\right),t\right)\\\mathbf{x}^i&=\mathrm{Decoder}(\mathbf{z}_0^i)\end{aligned} \tag{21}$$

where SchedulerStep$(\cdot)$ denotes a generic update rule for the diffusion process, and the conditional debiased embedding $\hat{\boldsymbol{e}}^i$ from CLIP exhibits reduced bias towards specific attributes, resulting in fewer conflicts during the guidance process, thereby enhancing the diversity and quality of the generated images.

4 Experiments

4.1 Experimental Setup

Attribute Sets. Referring to FairFace [30], we construct the attribute-specific prompt using following sets: $a_1\in$ {Male, Female} for gender case; $a_2\in$ {East Asian, Indian, Middle Eastern, White, Black, Southeast Asian} for the source of skin tone; and $a_3\in$ {0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70+} for the age case.

Quantitative Metrics. Following [16, 40, 53], we quantify image distribution diversity and image quality using: (1) *Fairness Discrepancy* (FD) measures image diversity via L2 norm distance between the attribute distribution of the generated images and a uniform distribution. (2) *Fréchet Inception Distance* (FID) score is computed by comparing our generated images to the FFHQ [31] dataset to assess the image quality. The lower both values, the better.

Experimental Setting. We use Stable Diffusion v2.1 [44] as the backbone for all methods. We set the generated images with the resolution of 768 by 768 pixels and employ DDIM [52] noise scheduler. We use OpenCLIP as the CLIP model, which is included within Stable Diffusion v2.1. We use the CLIP text encoder and set the dimension of the text embeddings as 1024, $\alpha=0.5$, $\beta=2$, and $\eta=0.75$. The default value of γ is 7.5.

4.2 Baselines

Stable Diffusion (SD) [44] is the baseline model. We split our comparison methods into three categories: (1) *Debiasing in CLIP space*: **VL-Debias** [18] uses an orthogonal projection matrix to eliminate the gender bias in text embedding from CLIP; **ITI-GEN** [66] trains a series of inclusive tokens adding after CLIP embedding to control sensitive attributes. (2) *Debiasing in U-Net space*: **Fair Diffusion (FairDiff)** [20] employs Semantic Guidance (SEGA) [8] to enhance diversity in image generation. **InterpretDiff** [34] introduces semantic vectors into the U-Net to enable attribute-level

Table 1: The FD score comparisons with benchmarks on (a) intra-categories attribute and (b) inter-categories attributes. The lower ↓ the FD value, the better. We use CLIP [41] and pre-trained classifiers [30] as attribute classifier [15, 18]. † denotes evaluation on binary classes, as its strategy does not support multi-class classification or report inter-category results (-). ∗ indicates they only support 4 "Skin" classes and 3 "Age" classes.

Methods	(a) Intra-categories Attribute			(b) Inter-categories Attributes		
	Gender (2 classes)	Age (8 classes)	Skin (6 classes)	Gender × Skin (2 × 6 classes)	Gender × Age (2 × 8 classes)	Gender × Age × Skin (2 × 6 × 8 classes)
SD [44]	0.281	0.528	0.547	0.432	0.406	0.315
VL-DEBIAS† [18]	0.321	0.473	0.703	–	–	–
BALANCING ACT* [40]	0.109	0.789	0.224	–	–	–
INTERPRETDIFF* [34]	0.017	0.198	0.327	–	–	–
FINETUNING* [49]	0.204	-	0.119	0.224	–	–
FAIRDIFF [20]	0.087	0.111	0.219	0.168	0.141	0.171
ITI-GEN [66]	**0**	0.264	0.104	0.140	0.217	**0.091**
BADGE w/o CLIP	5×10^{-5}	0.160	0.101	0.156	0.145	0.162
BADGE w/o U-NET	0.305	0.423	0.576	0.675	0.661	0.634
BADGE (FULL)	**0**	**0.107**	**0.051**	**0.114**	**0.138**	0.166

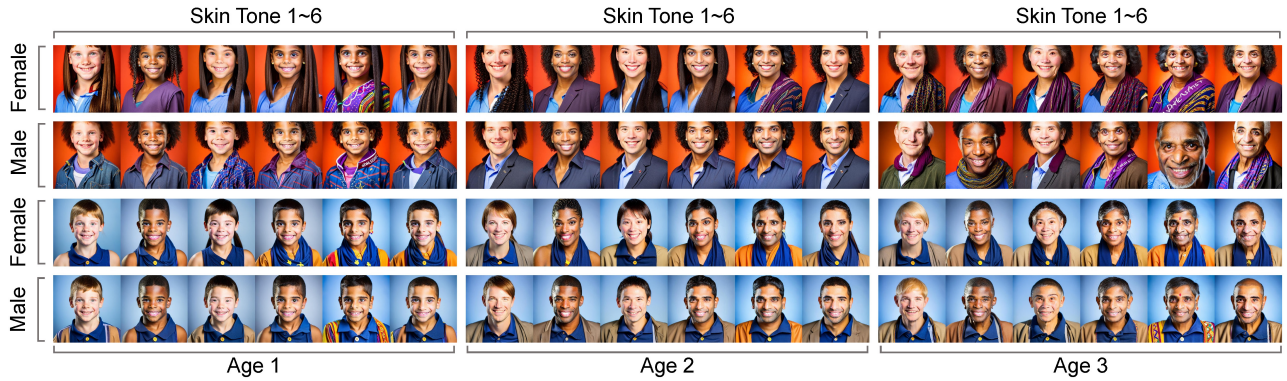

Figure 3: Qualitative results of inter-categories Gender × Age × Skin in Table 1 ("a headshot of a person"). BADGE achieves precise fine-grained control via bimodal adaptive guidance, showing 36 variations of one person, a total of two people.

control. **Balancing Act** [40] performs distribution guidance in h-space to achieve debiasing. (3) *Debiasing in CLIP and U-Net spaces*: **Finetuning** [49], a training-required model, supports simultaneous fine-tuning in both spaces.

4.3 Main Results

Intra-categories Attribute. To show that BADGE can mitigate various types of biases, we perform adaptive guidance on multiple single attributes. As observed the FD score in Table 1, BADGE achieves superior diversity compared to debiasing-only in the CLIP or U-Net methods. Compared to the current CLIP-only debiasing SOTA model ITI-GEN, BADGE improves gender and skin tone diversity by 59.4% and 50.9%, respectively. Visual-only methods like Balancing Act and InterpretDiff perform poorly due to representation conflicts caused by indiscriminate attribute guidance in the U-Net. Although Finetuning[49] targets both modalities, its results are suboptimal. Its original paper also notes that fine-tuning the U-Net leads to classifier overfitting and quality degradation. VL-Debias shows the worst performance as the *fair text condition fails to control visual space* to synthesize balanced images as bias also exists in U-Net. This verifies our motivation: considering biases in both textual and visual spaces, which is conducive to mitigating bias throughout the generation process, thereby increasing image diversity. Moreover, the bias location and quantification are general across different attributes, thus ensuring robustness.

Inter-categories Attributes. *When given multiple attributes, can BADGE's intersection of these attributes vary continuously and linearly like single attributes?* BADGE provides an affirmative answer, as shown in Table 1 (b). BADGE improves diversity in Gender×Age and Gender×Skin Tone intersections by 18.6% and 36.4%, respectively, compared to ITI-GEN, while keeping good image fidelity in almost all cases. The performance of ITI-GEN falls short of ours due to its reliance on external corpora for discrete training, which prevents it from including attributes not present in the dataset. However, this kind of training is also good in some situations, for Gender×Age×Skin Tone case, ITI-Gen slightly outperforms

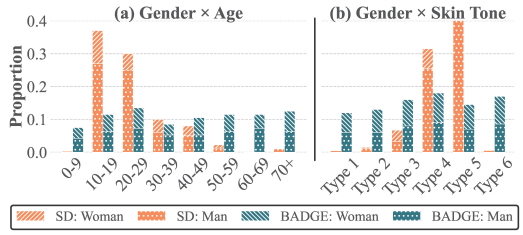

Figure 4: Inter-category distribution with "a headshot of a person". The images generated by BADGE are distributed more uniformly across various sub-groups compared to the original Stable Diffusion. Qualitative results *cf.* Fig. 5.

Figure 5: Results of BADGE on inter-category attributes for Gender × Age (Fig. 4 (a)) and Gender × Skin Tone (Fig. 4 (b)). Examples are randomly picked with "a headshot of a person".

BADGE in diversity, as well as in the fidelity of Gender×Skin Tone in Table 2, mainly due to its better attribute representations learned from a dataset. In contrast, our adaptive guidance is without training. We also visualize the inter-category distribution under two settings in Fig. 4: (a) Gender × Age, and (b) Gender × Skin Tone. Interestingly, we observe that the original SD model generated more proportion of Type 5 examples featuring Black individuals. We speculate that the model includes an internal debiasing mechanism, but it may be prone to over-correction. BADGE achieves inclusiveness across all setups, particularly in extremely underrepresented categories for ages (<10 and > 50 years old), skin tone for Type 1 (East Asian) and Type 6 (Southeast Asian). This inclusiveness is attributed to the integrated effect of the reduced bias impact in text embedding and the adaptive guidance in visual space.

Beyond Specific Domains. In addition to demographic biases related to gender, age, and skin tone, since BADGE does not require training on specific attributes, it can be easily extended to any other domains. For instance, the natural scenery in Fig. 6 demonstrates our generalizability across different domains.

Fine-Grained Visualization from BADGE. As shown in Fig. 8, the synthesized images regarding intra-categories gender and age cases, BADGE linearly and gradually increases male and age characteristics. The arithmetic benefit of the noisy state within the U-Net, enabling the linear attribute changes and control a much broader fine-grained by several discrete prompts. Moreover, BADGE not only demonstrates fine-grained control within a single intra-category

Figure 6: Fine-grained control of natural scenery.

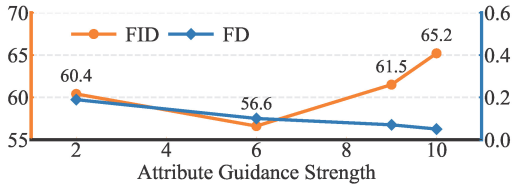

Figure 7: The trade-off between image fidelity and diversity of attribute guidance strength γ in Gender×Age case.

Table 2: The balance between image quality and diversity.

Method	Gender×Skin		Gender×Age		Gender×Skin×Age	
	FD ↓	FID ↓	FD ↓	FID ↓	FD ↓	FID ↓
SD	0.432	76.18	0.355	70.11	0.315	71.86
FairDiff [20]	0.168	74.22	0.141	74.37	0.171	68.33
ITI-Gen [67]	0.140	**60.68**	0.217	57.76	**0.091**	62.68
BADGE (ours)	**0.114**	68.88	**0.015**	**56.65**	0.166	**62.61**

but also achieves comparable performance across inter-category scenarios. Intuitive examples over Gender × Age , and Gender × Skin in Fig. 5, Fig. 9, and the more complex inter-categories Gender × Age × Skin case in Fig. 3 illustrate its precise fine-grained control.

4.4 Ablations and Applications

To verify the necessity of considering biases both from textual and visual space, we design two ablation versions and the experimental results are reported in Table 1:

- **BADGE** w/o CLIP: We observe that removing CLIP module leads to a decrease in both diversity and quality compared to the full version. This suggests that while text-space bias has a smaller impact on the final image, conflicts with the visual space significantly reduce image quality and complicate guidance in U-Net.
- **BADGE** w/o U-Net: Notably, when removing the U-Net debiasing module, we observe a significant decrease in both image quality and diversity. This highlights the critical role of adaptive attribute guidance within the U-Net model (Eq. (20)) in debiasing, as it affects the predicted noise to control the image attribute.

Figure 8: Visualization of fine-grained control over Gender from *female* to *male* (top), and Age from *young* to *old* (bottom).

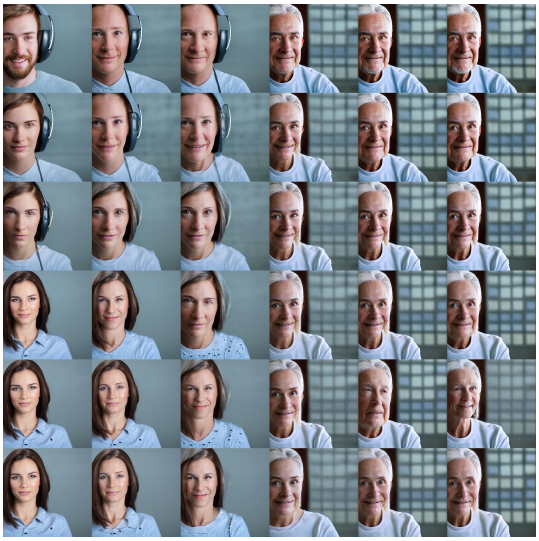

Figure 9: Fine-grained control over intersectional attributes.

Additionally, we conduct an ablation study on the attribute guidance strength γ (Eq. (20)) in visual space. The trend reported by FID and FD in Fig. 7 illustrates that as γ increases, diversity improves significantly, while image fidelity first improves and then declines. This highlights the appropriate selection of the guidance strength, which our quantified bias degree can adaptively determine. Table 2 reports the FD and FID scores for inter-category attributes. Without requiring any finetuning or additional training, BADGE achieves superior performance in both image fidelity and diversity compared to existing SOTA approaches.

Compatibility with ControlNet [68]. BADGE enhances diversity without additional training or modifications to the original text-to-image model, thereby supporting a variety of downstream visual tasks. In Fig. 10 and Fig. 11, we illustrate BADGE's compatibility with ControlNet, allowing for the manipulation of inputs beyond text prompts. The latent-based guidance may introduce another type of inherent bias, such as clothing style preferences, which are beyond the scope of this work. However, we focus on adaptive guidance to mitigate biases that pose risks to social well-being.

5 Conclusion and Future Works

In this paper, we present BADGE, a simple yet effective adaptive guidance framework for mitigating social biases in text-to-image generation. Unlike prior methods that require retraining or operate

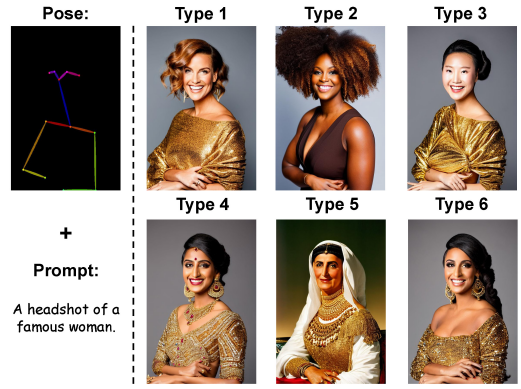

Figure 10: Compatibility with pose [11] condition. BADGE enhances the diversity of ControlNet [68] by adaptive guidance.

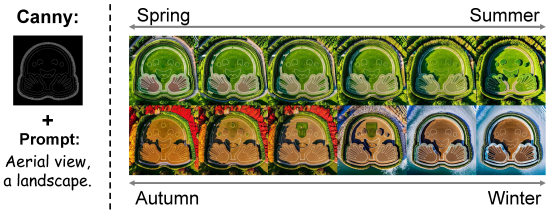

Figure 11: Compatibility with Canny [10] condition. BADGE-controls the season's transition by adaptive guidance.

in a single modality, BADGE manipulates biases in both textual and visual spaces, enabling training-free, bimodal debiasing. Our key insights include: (1) the first use of quantified bias inclination as debiasing strength for adaptive correction across attributes and professions; (2) early attribute prediction to efficiently estimate visual bias without full image generation; and (3) the use of attribute prominence to determine whether guidance is necessary or generation should proceed naturally. Extensive experiments demonstrate that BADGE achieves robust fairness across individual and intersectional attributes, and generalizes well beyond human-centric prompts to domains such as animals and landscapes.

We believe our approach offers a practical step toward building fair, accountable, and responsible generative AI systems. In our future work, we plan to enhance the interpretability of debiasing decisions and establishing standardized, large-scale benchmarks for evaluating fairness in diffusion models, or extend BADGE to more complex tri-modal setting debiasing (e.g., audio or 3D context).

Acknowledgements

This work was supported by Science and Technology Program of Chengdu (Grant No. 2024YF0501231SN), Sichuan Provincial Science and Technology Achievements Transfer and Transformation Project (Grant No. 2025ZHCG0012), Central-guided Local Science and Technology Development Special Fund Project (Grant No. 2024ZYD0265), Sichuan Natural Science Foundation (Grant No. 2025ZNSFSC0478), and China Scholarship Council program (Grant No. 202506070076).

References

[1] Mohsan Alvi, Andrew Zisserman, and Christoffer Nellåker. 2018. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *Proceedings of the European conference on computer vision (ECCV) workshops*. 0–0.

[2] Hritik Bansal, Da Yin, Masoud Monajatipoor, and Kai-Wei Chang. 2022. How well can text-to-image generative models understand ethical natural language interventions? *arXiv preprint arXiv:2210.15230* (2022).

[3] Hugo Berg, Siobhan Hall, Yash Bhalgat, Hannah Kirk, Aleksandar Shtedritski, and Max Bain. 2022. A Prompt Array Keeps the Bias Away: Debiasing Vision-Language Models with Adversarial Learning. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang (Eds.). Association for Computational Linguistics, Online only, 806–822. https://aclanthology.org/2022.aacl-main.61

[4] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2023. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1493–1504.

[5] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963* (2021).

[6] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of" bias" in nlp. *arXiv preprint arXiv:2005.14050* (2020).

[7] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems* 29 (2016).

[8] Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. 2023. Sega: Instructing text-to-image models using semantic guidance. *Advances in Neural Information Processing Systems* 36 (2023), 25365–25389.

[9] Manuel Brack, Felix Friedrich, Katharia Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. 2024. Ledits++: Limitless image editing using text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8861–8870.

[10] John Canny. 1986. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence* 6 (1986), 679–698.

[11] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7291–7299.

[12] Bowen Chen, Yun Sing Koh, and Gillian Dobbie. 2024. SSAT-Adapter: Enhancing Vision-Language Model Few-shot Learning with Auxiliary Tasks. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 1004–1013.

[13] Zhangtao Cheng, Jian Lang, Ting Zhong, and Fan Zhou. 2025. Seeing the Unseen in Micro-Video Popularity Prediction: Self-Correlation Retrieval for Missing Modality Generation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*. 142–152.

[14] Jaemin Cho, Abhay Zala, and Mohit Bansal. 2023. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3043–3054.

[15] Jaemin Cho, Abhay Zala, and Mohit Bansal. 2023. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3043–3054.

[16] Kristy Choi, Aditya Grover, Trisha Singh, Rui Shu, and Stefano Ermon. 2020. Fair generative modeling via weak supervision. In *International Conference on Machine Learning*. PMLR, 1887–1898.

[17] Yujin Choi, Jinseong Park, Hoki Kim, Jaewook Lee, and Saerom Park. 2024. Fair sampling in diffusion models through switching mechanism. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 21995–22003.

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

[19] Sepehr Dehdashtian, Lan Wang, and Vishnu Naresh Boddeti. 2024. FairerCLIP: Debiasing CLIP's Zero-Shot Predictions using Functions in RKHSs. *arXiv preprint arXiv:2403.15593* (2024).

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

[21] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618* (2022).

[22] Aditya Grover, Jiaming Song, Ashish Kapoor, Kenneth Tran, Alekh Agarwal, Eric J Horvitz, and Stefano Ermon. 2019. Bias correction of learned generative models using likelihood-free importance weighting. *Advances in neural information processing systems* 32 (2019).

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

[24] Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).

[25] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. 2023. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems* 36 (2023), 78723–78747.

[26] Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. 2022. Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably). In *International conference on machine learning*. PMLR, 9226–9259.

[27] Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, et al. 2024. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8362–8371.

[28] Inbar Huberman-Spiegelglas, Vladimir Kulikov, and Tomer Michaeli. 2024. An edit friendly DDPM noise space: Inversion and manipulations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12469–12478.

[29] Yue Jiang, Yueming Lyu, Ziwen He, Bo Peng, and Jing Dong. 2024. Mitigating Social Biases in Text-to-Image Diffusion Models via Linguistic-Aligned Attention Guidance. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 3391–3400.

[30] Kimmo Karkkainen and Jungseock Joo. 2021. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 1548–1558.

[31] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4401–4410.

[32] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1931–1941.

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

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

[35] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems* 36 (2024).

[36] Yuzhou Mao, Liu Yu, Yi Yang, Fan Zhou, and Ting Zhong. 2023. Debiasing intrinsic bias and application bias jointly via invariant risk minimization (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 16280–16281.

[37] Pietro Melzi, Christian Rathgeb, Ruben Tolosana, Ruben Vera-Rodriguez, Dominik Lawatsch, Florian Domin, and Maxim Schaubert. 2023. GANDiffFace: Controllable Generation of Synthetic Datasets for Face Recognition with Realistic Variations. *arXiv preprint arXiv:2305.19962* (2023).

[38] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. 2024. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 4296–4304.

[39] Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. StereoSet: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456* (2020).

[40] Rishubh Parihar, Abhijnya Bhat, Abhipsa Basu, Saswat Mallick, Jogendra Nath Kundu, and R Venkatesh Babu. 2024. Balancing Act: Distribution-Guided Debiasing in Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer*

Vision and Pattern Recognition. 6668–6678.

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

[42] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* 1, 2 (2022), 3.

[43] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* 1, 2 (2022), 3.

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

[45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

[46] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 22500–22510.

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

[48] Ashish Seth, Mayur Hemani, and Chirag Agarwal. 2023. DeAR: Debiasing Vision-Language Models with Additive Residuals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6820–6829.

[49] Xudong Shen, Chao Du, Tianyu Pang, Min Lin, Yongkang Wong, and Mohan Kankanhalli. 2024. Finetuning Text-to-Image Diffusion Models for Fairness. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=hnrB5YHoYu

[50] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. 2024. Instantbooth: Personalized text-to-image generation without test-time finetuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8543–8552.

[51] Robik Shrestha, Yang Zou, Qiuyu Chen, Zhiheng Li, Yusheng Xie, and Siqi Deng. 2024. FairRAG: Fair human generation via fair retrieval augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11996–12005.

[52] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).

[53] Christopher Teo, Milad Abdollahzadeh, and Ngai-Man Man Cheung. 2024. On measuring fairness in generative models. *Advances in Neural Information Processing Systems* 36 (2024).

[54] Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. 2023. Effective data augmentation with diffusion models. *arXiv preprint arXiv:2302.07944* (2023).

[55] Jialu Wang, Yang Liu, and Xin Eric Wang. 2021. Are gender-neutral queries really gender-neutral? mitigating gender bias in image search. *arXiv preprint arXiv:2109.05433* (2021).

[56] Mei Wang and Weihong Deng. 2020. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9322–9331.

[57] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. 2020. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8919–8928.

[58] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359* (2021).

[59] Chen Henry Wu, Saman Motamed, Shaunak Srivastava, and Fernando D De la Torre. 2022. Generative visual prompt: Unifying distributional control of pre-trained generative models. *Advances in Neural Information Processing Systems* 35 (2022), 22422–22437.

[60] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. 2018. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE international conference on big data (big data)*. IEEE, 570–575.

[61] Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, et al. 2023. mplug-2: A modularized multi-modal foundation model across text, image and video. In *International Conference on Machine Learning*. PMLR, 38728–38748.

[62] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721* (2023).

[63] Liu Yu, Ludie Guo, Ping Kuang, and Fan Zhou. 2024. Biases mitigation and expressiveness preservation in language models: A comprehensive pipeline (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 23701–23702.

[64] Liu Yu, Ludie Guo, Ping Kuang, and Fan Zhou. 2025. Bridging the fairness gap: Enhancing pre-trained models with llm-generated sentences. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[65] Liu Yu, Yuzhou Mao, Jin Wu, and Fan Zhou. 2023. Mixup-based unified framework to overcome gender bias resurgence. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1755–1759.

[66] Cheng Zhang, Xuanbai Chen, Siqi Chai, Chen Henry Wu, Dmitry Lagun, Thabo Beeler, and Fernando De la Torre. 2023. Iti-gen: Inclusive text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3969–3980.

[67] Cheng Zhang, Xuanbai Chen, Siqi Chai, Chen Henry Wu, Dmitry Lagun, Thabo Beeler, and Fernando De la Torre. 2023. Iti-gen: Inclusive text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3969–3980.

[68] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3836–3847.

[69] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.03310* (2019).

[70] Fan Zhou, Yuzhou Mao, Liu Yu, Yi Yang, and Ting Zhong. 2023. Causal-debias: Unifying debiasing in pretrained language models and fine-tuning via causal invariant learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 4227–4241.

[71] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. 2023. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852* (2023).