Enhancing Diversity in Text-to-Image Generation without Compromising Fidelity

Anonymous authors
Paper under double-blind review

Abstract

Effective text-to-image generation must synthesize images that are both realistic in appearance (sample fidelity) and have sufficient variations (sample diversity). Diffusion models have achieved promising results in generating high-fidelity images based on textual prompts, and recently, several diversity-focused works have been proposed to improve their demographic diversity by enforcing the generation of samples from various demographic groups. However, another essential aspect of diversity, sample diversity—which enhances prompt reusability to generate creative samples that reflect real-world variability—has been largely overlooked. Specifically, how to generate images that have sufficient demographic and sample diversity while preserving sample fidelity remains an open problem because increasing diversity comes at the cost of reduced fidelity in existing works. To address this problem, we first propose a bimodal low-rank adaptation of pretrained diffusion models which decouples the text-to-image conditioning, and then propose a lightweight bimodal guidance method that introduces additional diversity to the generation process using reference images retrieved through a fairness strategy by separately controlling the strength of text and image conditioning. We conduct extensive experiments to demonstrate the effectiveness of our method in enhancing demographic diversity (Intersectional Diversity (Shrestha et al., 2024)) by 2.47× and sample diversity (Recall (Kynkäänniemi et al., 2019)) by 1.45× while preserving sample fidelity (Precision (Kynkäänniemi et al., 2019)) compared to the baseline diffusion model.

1 Introduction

Diffusion models have made significant progress in generating high-fidelity content across various applications (*e.g.*, Text-to-Image (T2I) Generation) (Guo and Chen, 2024; Ho et al., 2020; Rombach et al., 2022; Dhariwal and Nichol, 2021). Despite the advancements in achieving better control over these models to produce high-fidelity content, there remains a lack of sufficient control in generating both high-fidelity and diverse content, particularly in terms of *demographic diversity* and *sample diversity*. Demographic diversity aims to address societal biases and ensure fair representation across diverse demographic groups (Wan et al., 2024a; Xu et al., 2018; Friedrich et al., 2023; Li et al., 2024b) while sample diversity seeks to improve prompt reusability to avoid monotonous or overly similar outputs and ensure that generated samples capture the variability of real-world samples (Zhang and Schomaker, 2022; Miao et al., 2024; Sadat et al., 2023; Li et al., 2021; Xia et al., 2021; Liu et al., 2020; Wang et al., 2024). Recently, several diversity-focused works (Friedrich et al., 2023; Esposito et al., 2023; Li et al., 2023a; Zhang et al., 2023a; Bansal et al., 2022; Luccioni et al., 2023; Perera and Patel, 2023; Bianchi et al., 2023) have been proposed to improve demographic diversity. However, enhancing sample diversity in diffusion models, particularly improving both aspects of diversity, remains overlooked.

Insufficient sample diversity in diffusion models hinders their further advancements in producing more photorealistic large-scale images (Marwood et al., 2023; Zameshina et al., 2023; Rassin et al., 2024; Li et al., 2024a). As shown in Fig. 1 (left, first row), images generated by state-of-the-art (SOTA) diffusion model (Stable Diffusion 3.5-Large (Esser et al., 2024)) appear repetitive, formulaic, and monotonous, resembling typical stock photos (*e.g.*, images generated with the prompt "Photo of a doctor" always feature a white wall background and a similar pose). This sharply contrasts with real-world images, which exhibit greater variation and creativity (Kynkäänniemi et al., 2019; Naeem et al., 2020), thereby limiting their broader

Figure 1: The proposed method (right) enhances both demographic diversity (*e.g.*, generating both female and male images) and sample diversity (*e.g.*, varied poses and backgrounds) while preserving sample fidelity (*e.g.*, sharp details, natural faces, and realistic lighting).

applicability. Furthermore, SOTA models (Betker et al., 2023; Podell et al., 2023) are limited in large-scale image generation, as when generating 50-200 images from the same prompt, they often produce images similar to those already generated images even with different random initialization seeds (Tang et al., 2024; Du et al., 2024). We quantitatively demonstrate that SOTA models and existing diversity-focused methods struggle to capture real-world sample variability in Tab. 1 and Fig. 4, and exhibit limited prompt reusability in Fig. 5.

Since sample diversity is often overlooked in the literature, we propose a method to enhance it while also considering demographic diversity. It is well established that diversity and fidelity exhibit a trade-off (Dhariwal and Nichol, 2021; Ho and Salimans, 2022; Brock, 2018b; Kingma and Dhariwal, 2018), meaning that improving one often reduces the other. For instance, many general methods (Nichol et al., 2021; Blattmann et al., 2022), such as Classifier-Free Guidance (CFG) (Ho and Salimans, 2022), a widely adopted technique in diffusion models, trade off output variability over fidelity, as lower guidance scales introduce more details but reduce control. Consequently, a key challenge in improving diversity is effectively alleviating this diversity-fidelity trade-off, as both are essential for generating high-quality images. In this paper, our primary objective is to enhance both demographic and sample diversity while preserving sample fidelity.

To achieve our objective, we investigate whether *incorporating additional diversity can enhance output diversity without compromising fidelity*. To introduce more diversity, we alternate training between text and image modalities to encode image information, and leverage the rich visual details from multiple retrieved reference images to augment the user prompt during inference. To preserve sample fidelity, we propose Bimodal Classifier-Free Guidance (BCFG), which extends the unified control of text and image modalities in CFG by separately controlling guidance from each modality. This technique allows image modality to fully exploit the added diversity while properly tuning text modality to maintain both sample fidelity and image-text alignment, as empirically verified in Fig. 7. Additionally, existing demographic diversity-focused methods (Shrestha et al., 2024) typically support only a fixed fairness criterion, which may cause overshooting biases (Wan and Chang, 2024). To address this limitation, we reformulate demographic diversity enhancement as a distribution alignment problem (Shen et al., 2024), enabling adaptability to various fairness criteria based on user needs. Finally, our method is implemented in a lightweight and efficient manner. Specifically, we design a low-rank joint text-image adapter architecture on the pre-trained model for bimodal conditioning. Our approach eliminates the computational overhead of re-training or Full Fine-Tuning (FFT) (Blattmann et al., 2022; Chen et al., 2022) by freezing all text modality parameters to preserve

pre-trained knowledge and designing a delicate adapter for the image modality. We summarize our main contributions below:

- Proposing, to our knowledge, the first method to enhance both demographic diversity and sample diversity while preserving sample fidelity and alignment by alleviating the diversity-fidelity trade-off in a lightweight and efficient manner (12M parameters for Stable Diffusion v2.1).
- Highlighting and formulating sample diversity which is overlooked by existing diversity-focused methods.
- Providing an extensive empirical analysis of the proposed method, demonstrating its superior performance in enhancing demographic diversity (Intersectional Diversity (Shrestha et al., 2024)) from 0.19 to 0.47 and sample diversity (Recall) from 0.31 to 0.45 while maintaining sample fidelity (Precision) at 0.54, comparable to 0.52 of the baseline.

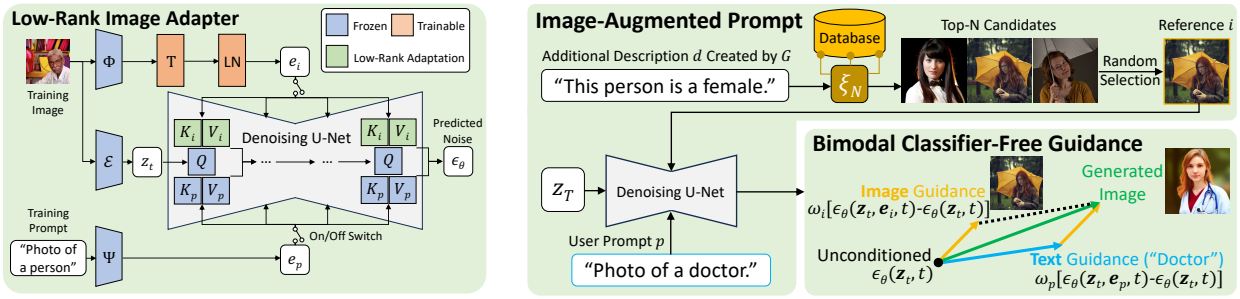
2 Related Work

Enhancing Diversity in Diffusion Models. While diffusion models can generate high-fidelity images (Ho et al., 2020; Rombach et al., 2022; Dhariwal and Nichol, 2021), they exhibit insufficient demographic (Wan et al., 2024a; He et al., 2024; Esposito et al., 2023) and sample diversity (Marwood et al., 2023; Cao et al., 2024; Naeem et al., 2020). To enhance demographic diversity, several methods directly augment the user prompt with demographic description (Bansal et al., 2022; Ding et al., 2021; Friedrich et al., 2023), while others apply Parameter-Efficient Fine-Tuning (Zhang et al., 2023a; Teo et al., 2024; Shen et al., 2024). Building on Retrieval-Augmented Generation (RAG) (Lewis et al., 2020; Cai et al., 2022), FairRAG (Shrestha et al., 2024) uses the user prompt (*e.g.*, "Photo of a doctor") to retrieve relevant images from external databases and boosts minority group sampling. Despite efforts to enhance demographic diversity (Luo et al., 2024; Chuang et al., 2023; Fraser et al., 2023; Gandikota et al., 2024), few methods improve sample diversity or both diversity in diffusion models.

Diversity-Fidelity Trade-Off. Diffusion models inherently exhibit the diversity-fidelity trade-off (Nichol et al., 2021; Blattmann et al., 2022), which limits high-quality data generation, as both aspects contribute to overall quality. Many general techniques (Dhariwal and Nichol, 2021; Ho and Salimans, 2022; Brack et al., 2023; Friedrich et al., 2023; Chen et al., 2022; Bansal et al., 2023; Epstein et al., 2023) enhance fidelity at the expense of diversity (Dhariwal and Nichol, 2021; Ho and Salimans, 2022; Brack et al., 2023; Friedrich et al., 2023; Chen et al., 2022; Bansal et al., 2023; Epstein et al., 2023). For instance, Classifier Guidance (CG) (Dhariwal and Nichol, 2021) increases fidelity by combining the original score estimate with the gradient from an auxiliary classifier to better align the sampling process with the conditioning information, albeit at the cost of reduced diversity. Classifier-Free Guidance (CFG) (Ho and Salimans, 2022) mathematically interprets the classifier gradient in CG as a combination of unconditional and conditional score estimates, removing the need for an explicit classifier. However, effectively enhancing diversity without compromising fidelity remains an open challenge.

3 Methods

We propose a lightweight and efficient method to enhance both demographic and sample diversity while preserving sample fidelity. Our method leverages the additional diversity introduced by retrieved reference images to alleviate the diversity-fidelity trade-off. During training, we design Low-Rank Image Adapter (LoRIA), a lightweight adapter architecture that extracts image information into visual token and efficiently fuses text and image modalities to enable pre-trained models to incorporate reference images as additional conditioning. During inference, we introduce Image-Augmented Prompt (IAP), which retrieves reference images based on generated demographic description to augment the original user prompt, thereby introducing additional sample diversity and enhancing demographic diversity in a manner adaptable to various fairness criteria. Furthermore, we propose Bimodal Classifier-Free Guidance (BCFG), which independently controls the guidance strength of text and image modalities, better aligning the sampling process with conditioning information from each modality.



(a) During training, we alternate between text and image modalities to encode image information, and freeze all text modality parameters while training a linear projector T with Layer Normalization (LN) and applying LoRA for the image modality to incorporate additional image conditioning.

(b) During inference, we first retrieve Top-N candidates using Diverse Retrieval Strategy ξ based on specific fairness criteria, and then randomly select one or multiple images as reference images to construct Image-Augmented Prompt. Lastly, we use Bimodal Classifier-Free Guidance to leverage reference images to introduce additional diversity while applying stronger text guidance to preserve sample fidelity and text-image alignment.

Figure 2: The framework of the proposed Low-Rank Image Adapter, Image-Augmented Prompt, and Bimodal Classifier-Free Guidance.

3.1 Low-Rank Image Adapter (LoRIA)

To incorporate additional conditioning on reference images i, several methods (Blattmann et al., 2022; Chen et al., 2022; Ramesh et al., 2022) re-train or fine-tune pre-trained SD models which are originally conditioned solely on text prompts p. However, these methods are resource-intensive and time-consuming, and may compromise image-text alignment due to unrelated content in i retrieved from limited external databases, as studied in Fig. 6. To address this and ensure that the image modality introduces variation without dominating generation, we freeze all text-modality parameters to retain pre-trained knowledge (Li et al., 2023c), while training a linear projector and applying Low-Rank Adaptation (LoRA)(Hu et al., 2021) to selectively update specific cross-attention weights in the modality fusion module (*e.g.*, U-Net(Ronneberger et al., 2015)), as illustrated in Fig. 2a. This lightweight design enables efficient image generation during both training and inference. Additionally, it helps prevent catastrophic forgetting (Kirkpatrick et al., 2017; Zhai et al., 2023; Smith et al., 2023) for the training prompt "Photo of a person", as analyzed in Appendix 3.2.

Extracting Image Information into the Visual Token. We utilize a frozen CLIP image encoder Φ followed by a trainable linear layer T and Layer Normalization (LN) (Ba, 2016) to obtain the visual token embedding $\mathbf{e}_i$. To ensure the model effectively extracts image information into the visual token, including demographic information, we train it on cases where only reference images are provided without any text prompt (Algorithm 1), which forces the model to rely maximally on the image for noise removal. As evidence, our ablation study (Tab. 3) demonstrates that omitting model training on cases with only image modality leads to a notable reduction in demographic diversity. Thus, during inference, the visual token is expected to retain demographic information from the retrieved reference images.

Fusing Text and Image Modalities. The bimodal embeddings $(\mathbf{e}_p, \mathbf{e}_i)$ are fed into SD for conditioning, with the text token embedding $\mathbf{e}_p$ obtained from a frozen CLIP text encoder Ψ. To fuse two modalities while considering their distinct roles, we introduce a *low-rank decoupled cross-attention layer*. The query features, shared across both modalities, are computed as $\mathbf{Q} = \mathbf{z}\mathbf{W}^q$, where $\mathbf{W}^q$ is the corresponding weight matrix. For text modality, the key and value features are obtained as $\mathbf{K}_p = \mathbf{e}_p\mathbf{W}_p^k$ and $\mathbf{V}_p = \mathbf{e}_p\mathbf{W}_p^v$, respectively. For image modality, the key and value features are calculated using weight matrices derived from the text weight matrices via LoRA, *i.e.*, $\mathbf{W}_i^k = \mathbf{W}_p^k + \mathbf{B}^k\mathbf{A}^k$ and $\mathbf{W}_i^v = \mathbf{W}_p^v + \mathbf{B}^v\mathbf{A}^v$, where $\mathbf{B}^k \in \mathbb{R}^{m\times r}$, $\mathbf{A}^k \in \mathbb{R}^{r\times n}$, $\mathbf{B}^v$, and $\mathbf{A}^v$ are low-rank matrices that approximate fine-tuning adjustments, with rank $r \ll \min(m, n)$. Overall, the layer output is formulated as:

$$\mathbf{z}' = \text{Attention}(\mathbf{Q}, \mathbf{K}_p, \mathbf{V}_p) + \text{Attention}(\mathbf{Q}, \mathbf{K}_i, \mathbf{V}_i), \tag{1}$$

where $\mathbf{Q} = \mathbf{z}\mathbf{W}^q$, $\mathbf{K}_p = \mathbf{e}_p\mathbf{W}_p^k$, $\mathbf{V}_p = \mathbf{e}_p\mathbf{W}_p^v$, $\mathbf{K}_i = \mathbf{e}_i(\mathbf{W}_p^k + \mathbf{B}^k\mathbf{A}^k)$, $\mathbf{V}_i = \mathbf{e}_i(\mathbf{W}_p^v + \mathbf{B}^v\mathbf{A}^v)$. During training, only $\mathbf{B}^k$, $\mathbf{A}^k$, $\mathbf{B}^v$, $\mathbf{A}^v$, and linear layer T with LN are learnable while all other parameters remain frozen. The training objective is to minimize the following loss function:

$$\mathcal{L} = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\mathbf{I}),\mathbf{z}_0,\mathbf{e}_p,\mathbf{e}_i,t\sim[1,T]}\left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, \mathbf{e}_p, \mathbf{e}_i, t)\|^2\right], \tag{2}$$

where ϵ is the sampled noise during the diffusion process; $\mathbf{z}_0 = \mathcal{E}(i)$ is the original latent representation of the training image i encoded by a variational autoencoder (VAE) (Kingma, 2013) $\mathcal{E}$; $\mathbf{z}_t = \alpha_t\mathbf{z}_0 + \sigma_t\epsilon$ is the noisy version of $\mathbf{z}_0$ at timestep t, with α_t and σ_t defining the diffusion process (Rombach et al., 2022); ϵ_θ is the noise predicted by the denoising diffusion model parameterized by θ; and p is the fixed training prompt (*e.g.*, "Photo of a person"). To enhance the model's ability to extract information from each modality, we randomly discard conditioning during training. Specifically, we replace the training prompt with an empty sequence with probability $\pi_p = 0.1$, following (Saharia et al., 2022), and replace the image embedding with the all-zero embeddings of the same size with probability $\pi_i = 0.1$. The training and inference algorithms are shown in Algorithm 1 and Algorithm 2. We present the implementation details in Appendix 6.

Algorithm 1 Training a Diffusion Model with Bimodal Classifier-Free Guidance

Require: π_p: the probability of training without text conditioning
Require: π_i: the probability of training without image conditioning
repeat
  $(\mathbf{z}_0, p, i) \sim \mathbf{p}(\mathbf{z}_0, p, i)$ ▷ Sample data $\mathbf{z}_0$, training prompt p, and reference images i from the dataset
  $p \leftarrow \emptyset$ with probability π_p ▷ Discard text conditioning with probability π_p
  $\mathbf{e}_i \leftarrow \emptyset$ with probability π_i ▷ Discard image conditioning with probability π_i
  $t \sim [1, T]$ ▷ Sample a timestep uniformly from the range
  $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ ▷ Generate Gaussian noise
  $\mathbf{z}_t = \alpha_t\mathbf{z}_0 + \sigma_t\epsilon$ ▷ Corrupt the latent representation by adding noise
  Take a gradient step on $\nabla_\theta\left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, \mathbf{e}_p, \mathbf{e}_i, t)\|^2\right]$ ▷ Optimize the denoising model
until converged

Comparison of Design Aspects between LoRIA and Existing Image Adapters (Ramesh et al., 2022; Ye et al., 2023; Zhang et al., 2023b; Mou et al., 2024). The main design difference between existing methods and ours stems from different objectives: while they aim to edit the input image based on the user prompt (where the input image acts as the primary subject and its details need to be maximally preserved) (Ye et al., 2023), our method intends to leverage the image to introduce additional variation, hence enhancing diversity in T2I generation (where the input image serves as supportive context (Zhou et al., 2024)). As compared in Fig. 6, our method better aligns generated images with the user prompt by addressing potential unrelated content in reference images. To control the image modality strength, existing methods use an image scale parameter λ (*e.g.*, scaling Attention$(\mathbf{Q}, \mathbf{K}_i, \mathbf{V}_i)$ in Eq. (1) (Ye et al., 2023)). However, adjusting this parameter during inference can introduce discrepancy between training and inference, which may induce flaws (*e.g.*, overexposure or lower resolution) (Lin et al., 2024b). Aligning it across training and inference requires retraining, which is time-consuming and limits flexibility for user control. To address this, we employ ω_i in BCFG, a training-free hyper-parameter, to control the strength of image modality while preserving image quality, as shown in Fig. 3. We present the detailed comparison between control using the image scale parameter λ and control via BCFG in Sec. 4.6.

3.2 Image-Augmented Prompt (IAP)

To introduce demographic diversity, conventional methods such as Text-Augmented Prompt (TAP) (Bianchi et al., 2023; Bansal et al., 2022; Ding et al., 2021) augment user prompt p (*e.g.*, "Photo of a doctor") by appending it with fixed demographic descriptions d (*e.g.*, "This person is a female."). However, TAP still exhibits limited sample diversity and struggles to substantially enhance demographic diversity due to linguistic ambiguity (*e.g.*, many *skin tone* categories are difficult to express using text alone) (Zhang et al., 2023a; Wan et al., 2024a; Shrestha et al., 2024). To address these limitations, we propose Image-Augmented Prompt (IAP), which enriches p with the visual token encoded from one or multiple reference

Algorithm 2 Inference with Bimodal Classifier-Free Guidance

Require: p: user prompt
Require: i: reference images
Require: w_p: text guidance scale
Require: w_i: image guidance scale
$\mathbf{z}_T \sim \mathcal{N}(0, \mathbf{I})$ ▷ Initialize latent representation with Gaussian noise
for $t = T, \ldots, 1$ **do**
  $\tilde{\epsilon}_\theta(\mathbf{z}_t, \mathbf{e}_p, \mathbf{e}_i, t) = \epsilon_\theta(\mathbf{z}_t, t) + \omega_p[\epsilon_\theta(\mathbf{z}_t, \mathbf{e}_p, t) - \epsilon_\theta(\mathbf{z}_t, t)] + \omega_i[\epsilon_\theta(\mathbf{z}_t, \mathbf{e}_i, t) - \epsilon_\theta(\mathbf{z}_t, t)]$ ▷ Compute BCFG score
  $\mathbf{z}_{t-1} = (\mathbf{z}_t - \sigma_t \tilde{\epsilon}_\theta(\mathbf{z}_t, \mathbf{e}_p, \mathbf{e}_i, t))/\alpha_t$ ▷ Update latent representation for the next timestep
end for
return $\mathbf{z}_0$ ▷ Return the final denoised latent representation

images i retrieved based on textual demographic description d. Incorporating visual token not only conveys demographic information to enhance demographic diversity but also introduces additional sample diversity through rich visual details (*e.g.*, appearance, environment, and lighting conditions) that text alone cannot provide, thereby increasing sample diversity. We mathematically prove (Theorem 1 in Appendix 1.1) and empirically verify (Tab. 1) that IAP generates more diverse images than TAP.

Diverse Retrieval Strategy (DRS). To further improve diversity, we introduce DRS, a two-step retrieval strategy ξ, as shown in Fig. 2b. First, we perform nearest neighbor search (Borgeaud et al., 2022) with demographic information d to retrieve Top-N candidates (Blattmann et al., 2022) based on cosine similarity between their CLIP-encoded text embeddings $\mathbf{e}_d$ and image embeddings $\mathbf{e}_i$ (Radford et al., 2021). Next, we randomly select one or more samples as reference images i for each generation query q. This random selection contributes to generating diverse samples since various image candidates representing d, rather than a fixed text description d, can reduce biases introduced by deterministic selection. Compared with TAP (p, d), IAP is defined as (p, i), where $i = \xi(d)$. With IAP (p, i) integrating both text p and image i, we explore a method to separately control the strength of each modality in Sec. 3.3.

Adaptable Description Generator (ADG). Existing demographic diversity-focused methods (Bansal et al., 2022; Friedrich et al., 2023; Zhang et al., 2023a; Teo et al., 2024; Shen et al., 2024; Shrestha et al., 2024) are designed for a specific fairness criterion and limit sufficient control over demographic distribution of generated images, which may cause overshooting biases (Wan and Chang, 2024; Wan et al., 2024b) since perceptions of fairness vary across contexts. To address this, we formulate demographic diversity enhancement as a distributional alignment problem (Shen et al., 2024) and propose ADG, which generates d for image retrieval based on a target distribution $\mathcal{D}$ specified by user-defined fairness criteria, making it adaptable to various fairness criteria. For a generation query q with prompt p, we use the target distribution $\mathcal{D}_p(a_1, ..., a_j)$ of demographic attributes $A_1, ..., A_j$ to direct a description generator $G(\mathcal{D}_p, \Lambda)$ in creating demographic description d_q by filling a template Λ (*e.g.*, "This person is a [AGE], [SKIN TONE] [SEX]."") with respective demographic qualifiers[1]. Note that $\mathcal{D}_p$ and Λ can be adapted to various fairness criteria such as disparate impact (Esposito et al., 2023), demographic factuality (Wan et al., 2024b), counter-stereotypes (Bianchi et al., 2023), or other user-defined fairness criteria; implementations are in Appendix 2. In the main paper, following (Friedrich et al., 2023; Xu et al., 2018; Shrestha et al., 2024), we consider a well-known fairness criterion, demographic parity (DP) (Hardt et al., 2016), to define $\mathcal{D}_p$. DP requires the independence between the label Y (*e.g.*, occupation) and demographic attribute A (*e.g.*, sex) for a (generated/synthetic) dataset, *i.e.*, $P(Y|A) = P(Y)$. Thus, to achieve DP and generate an overall balanced dataset, $\mathcal{D}_p$ is a uniform distribution[2] for any p. This DP-based augmentation forms the foundation for the proposed IAP.

3.3 Bimodal Classifier-Free Guidance (BCFG)

Classifier-Free Guidance (CFG) (Ho and Salimans, 2022) is commonly used in diffusion-based models (Ho et al., 2020; Rombach et al., 2022; Nichol et al., 2021) to direct the inverse diffusion process towards condi-

[1]We present the values for demographic qualifiers in Appendix 2.1.
[2]We present the proof in Appendix 2.2.

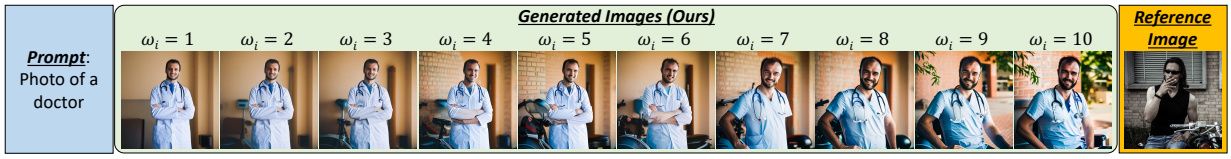


Figure 3: Visualization of varying image guidance scale ω_i from 1 to 10 in BCFG (Eq. (4)), with text guidance scale ω_p fixed at 7.5. ω_i controls how much the inverse diffusion process path will be pulled towards features of the reference image (*e.g.*, increasing ω_i shortens the doctor's sleeves to match the reference image). This allows users to easily choose their desired degree of diversity by varying ω_i.

tioning features by incorporating the predicted noise during sampling:

$$\tilde{\epsilon}_\theta(\mathbf{z}_t, \mathbf{e}_p, t) = \epsilon_\theta(\mathbf{z}_t, t) + \omega[\epsilon_\theta(\mathbf{z}_t, \mathbf{e}_p, t) - \epsilon_\theta(\mathbf{z}_t, t)], \tag{3}$$

where $\epsilon_\theta(\mathbf{z}_t, \mathbf{e}_p, t)$ and $\epsilon_\theta(\mathbf{z}_t, t)$ are the conditional and unconditional score estimate, respectively; $\omega > 0$ is the guidance scale; $\mathbf{z}_t$ is the intermediate representation of the U-Net (Ronneberger et al., 2015) at timestep t; and $\mathbf{e}_p$ is the text embedding of user prompt p. However, CFG is well known for exhibiting the diversity-fidelity trade-off, where reducing ω enhances diversity at the cost of fidelity (Dhariwal and Nichol, 2021; Ho and Salimans, 2022; Nichol et al., 2021). In this work, we propose two modifications to CFG that alleviate this trade-off, thereby enabling diffusion models to achieve greater diversity while preserving sample fidelity.

First, we extend CFG to incorporate reference images i (introduced by IAP) by replacing $\epsilon_\theta(\mathbf{z}_t, \mathbf{e}_p, t)$ with $\epsilon_\theta(\mathbf{z}_t, \mathbf{e}_p, \mathbf{e}_i, t)$ in Eq. (3) where $\mathbf{e}_i$ is the image embedding of i. By introducing image modality, rich visual details and variations from different reference images can introduce additional diversity beyond the textual description d, even though the user prompt p remains unchanged.

Second, to ensure that incorporating diversity does not compromise sample fidelity, we extend the unified control of text and image modalities in CFG by separately controlling guidance from two modalities. As empirically demonstrated in Fig. 7, using a unified guidance scale ω for both modalities leads to suboptimal sample fidelity and text-image alignment (Appendix 3.1) due to potentially conflicting content in i (*e.g.*, a skier) that contradicts p (*e.g.*, "Photo of a doctor") (Teo et al., 2024). However, in T2I generation, the generated image should primarily reflect user-provided prompt p (Li et al., 2019; Qiao et al., 2019; Ding et al., 2021). Thus, in our proposed IAP (p, i), the user prompt p serves as the primary role, while reference images i serve as the supportive role, introducing additional variations rather than imposing all details that could dominate generated images. To achieve this, we propose Bimodal Classifier-Free Guidance (BCFG), which enables separate control over the text prompt and reference image modalities by leveraging the predicted noise during sampling:

$$\begin{aligned}\tilde{\epsilon}_\theta(\mathbf{z}_t, \mathbf{e}_p, \mathbf{e}_i, t) = \epsilon_\theta(\mathbf{z}_t, t) &+ \omega_p[\epsilon_\theta(\mathbf{z}_t, \mathbf{e}_p, t) - \epsilon_\theta(\mathbf{z}_t, t)] \\ &+ \omega_i[\epsilon_\theta(\mathbf{z}_t, \mathbf{e}_i, t) - \epsilon_\theta(\mathbf{z}_t, t)],\end{aligned} \tag{4}$$

where $\omega_p > 0$ and $\omega_i > 0$ are guidance scales for user prompt p and reference images i. We set ω_p as 7.5 following (Ho and Salimans, 2022) and choose $\omega_i < \omega_p$ to prioritize text modality in image generation. In Fig. 7, we empirically compare BCFG with CFG on the diversity-fidelity trade-off, showing that BCFG can enhance diversity while preserving fidelity. Besides, we compare it with other guidance methods (Ho and Salimans, 2022; Brack et al., 2023; Chen et al., 2022) and analyze ω_i at different scales in Appendix 3.1.

The Corresponding Sampling Distribution of Bimodal Classifier-Free Guidance (BCFG). As proved below, the predicted noise in BCFG yields approximate samples from the following distribution[3],

$$\tilde{p}_\theta(\mathbf{z}_t|\mathbf{e}_p, \mathbf{e}_i) \propto p_\theta(\mathbf{z}_t) p_\theta(\mathbf{e}_p|\mathbf{z}_t)^{\omega_p} p_\theta(\mathbf{e}_i|\mathbf{z}_t)^{\omega_i}. \tag{5}$$

Intuitively, the effect of BCFG is to increase the sampling probability of data points with a higher likelihood of matching the user prompt p and reference image i by the implicit classifier (Ho and Salimans, 2022).

[3]For brevity, we omit the timestep variable t.

In Sec. 4, we apply BCFG to the proposed bimodal conditioning module (LoRIA) to control the strength of each modality. Notably, BCFG can be applied independently of the conditioning module, as empirically verified with other conditioning modules in Sec. 4.7.

Proof of the Corresponding Sampling Distribution. In this subsection, we provide a mathematical interpretation of the sample distribution of Bimodal Classifier-Free Guidance. Recall that the diffusion score (Ho and Salimans, 2022)

$$\begin{aligned}
\epsilon_\theta(\mathbf{z}_t, \mathbf{e}_p, \mathbf{e}_i) &= -\sigma_t \nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t, \mathbf{e}_p, \mathbf{e}_i) && (6)\\
&= -\sigma_t \nabla_{\mathbf{z}_t} \log[p(\mathbf{e}_i, \mathbf{e}_p) p(\mathbf{z}_t|\mathbf{e}_i, \mathbf{e}_p)] && (7)\\
&= -\sigma_t \nabla_{\mathbf{z}_t} [\log p(\mathbf{e}_i, \mathbf{e}_p) + \log p(\mathbf{z}_t|\mathbf{e}_i, \mathbf{e}_p)] && (8)\\
&\approx -\sigma_t \nabla_{\mathbf{z}_t} \log p(\mathbf{z}_t|\mathbf{e}_i, \mathbf{e}_p) && (9)
\end{aligned}$$

since $p(\mathbf{e}_i, \mathbf{e}_p)$ is not a function of $\mathbf{z}_t$.

Suppose we have two auxiliary implicit classifier models $p_\theta(\mathbf{e}_p|\mathbf{z}_t) \propto \frac{p(\mathbf{z}_t|\mathbf{e}_p)}{p(\mathbf{z}_t)}$ and $p_\theta(\mathbf{e}_i|\mathbf{z}_t) \propto \frac{p(\mathbf{z}_t|\mathbf{e}_i)}{p(\mathbf{z}_t)}$. Assume exact estimate $\epsilon^*(\mathbf{z}_t, \mathbf{e}_p)$ of $p(\mathbf{z}_t|\mathbf{e}_p)$, $\epsilon^*(\mathbf{z}_t, \mathbf{e}_i)$ of $p(\mathbf{z}_t|\mathbf{e}_i)$, and $\epsilon^*(\mathbf{z}_t)$ of $p(\mathbf{z}_t)$. The gradient of the resulting classifier can be written as $\nabla_{\mathbf{z}_t} \log p(\mathbf{e}_p|\mathbf{z}_t) = -\frac{1}{\omega_p}[\epsilon^*(\mathbf{z}_t, \mathbf{e}_p) - \epsilon^*(\mathbf{z}_t)]$ and $\nabla_{\mathbf{z}_t} \log p(\mathbf{e}_i|\mathbf{z}_t) = -\frac{1}{\omega_i}[\epsilon^*(\mathbf{z}_t, \mathbf{e}_i) - \epsilon^*(\mathbf{z}_t)]$ respectively. Since the exact scores $\epsilon^*(\mathbf{z}_t, \mathbf{e}_p)$, $\epsilon^*(\mathbf{z}_t, \mathbf{e}_i)$, and $\epsilon^*(\mathbf{z}_t)$ are not available, we use their estimates $\epsilon_\theta(\mathbf{z}_t, \mathbf{e}_p)$, $\epsilon_\theta(\mathbf{z}_t, \mathbf{e}_i)$, and $\epsilon_\theta(\mathbf{z}_t)$ respectively. The modified score function in Eq. (4) can thus be rewritten as

$$\begin{aligned}
&\epsilon_\theta(\mathbf{z}_t, t) + \omega_p[\epsilon_\theta(\mathbf{z}_t, \mathbf{e}_p, t) - \epsilon_\theta(\mathbf{z}_t, t)] + \\
&\qquad\qquad \omega_i\ [\epsilon_\theta(\mathbf{z}_t, \mathbf{e}_i, t) - \epsilon_\theta(\mathbf{z}_t, t)] && (10)\\
&\approx \epsilon_\theta(\mathbf{z}_t, t) - \sigma_t \omega_p \nabla_{\mathbf{z}_t} \log p_\theta(\mathbf{e}_p|\mathbf{z}_t) - \\
&\qquad\qquad \sigma_t \omega_i \nabla_{\mathbf{z}_t} \log p_\theta(\mathbf{e}_i|\mathbf{z}_t) && (11)\\
&= -\sigma_t \nabla_{\mathbf{z}_t} [\log p_\theta(\mathbf{z}_t) + \omega_p \log p_\theta(\mathbf{e}_p|\mathbf{z}_t) + \omega_i \log p_\theta(\mathbf{e}_i|\mathbf{z}_t)] && (12)\\
&= -\sigma_t \nabla_{\mathbf{z}_t} \log[p_\theta(\mathbf{z}_t) p_\theta(\mathbf{e}_p|\mathbf{z}_t)^{\omega_p} p_\theta(\mathbf{e}_i|\mathbf{z}_t)^{\omega_i}] && (13)\\
&= -\sigma_t \nabla_{\mathbf{z}_t} \log[p_\theta(\mathbf{z}_t) p_\theta(\mathbf{e}_p|\mathbf{z}_t) p_\theta(\mathbf{e}_i|\mathbf{z}_t) \\
&\qquad\qquad p_\theta(\mathbf{e}_p|\mathbf{z}_t)^{\omega_p - 1} p_\theta(\mathbf{e}_i|\mathbf{z}_t)^{\omega_i - 1}] && (14)\\
&= -p(\mathbf{e}_p, \mathbf{e}_i) \sigma_t \nabla_{\mathbf{z}_t} [p_\theta(\mathbf{z}_t|\mathbf{e}_p, \mathbf{e}_i) p_\theta(\mathbf{e}_p|\mathbf{z}_t)^{\omega_p - 1} p_\theta(\mathbf{e}_i|\mathbf{z}_t)^{\omega_i - 1}] && (15)
\end{aligned}$$

where $p(\mathbf{e}_p, \mathbf{e}_i)$ is a positive constant. In the last step, we used the Bayes Rule and the independence of random variable $\mathbf{e}_p$ and $\mathbf{e}_i$ given $\mathbf{z}_t$ to derive $p(\mathbf{z}_t|\mathbf{e}_p, \mathbf{e}_i) = \frac{p(\mathbf{z}_t, \mathbf{e}_p, \mathbf{e}_i)}{p(\mathbf{e}_p, \mathbf{e}_i)} = \frac{p(\mathbf{e}_p, \mathbf{e}_i|\mathbf{z}_t) p(\mathbf{z}_t)}{p(\mathbf{e}_i, \mathbf{e}_p)} = \frac{p(\mathbf{z}_t) p(\mathbf{e}_p|\mathbf{z}_t) p(\mathbf{e}_i|\mathbf{z}_t)}{p(\mathbf{e}_p, \mathbf{e}_i)}$. This proves Eq. (5).

4 Experimental Evaluation

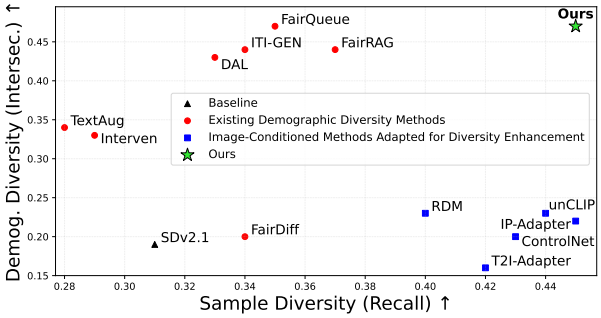
4.1 Experiment Setup

Datasets and Evaluation Prompts Following (Shrestha et al., 2024), we construct non-overlapping training and reference image datasets from human images in MSCOCO (Lin et al., 2014) and OpenImages-v6 (Krasin et al., 2017), and store the pre-computed CLIP image embeddings (Radford et al., 2021) to speed up the inference process by bypassing the usage of the image encoder during inference. For evaluation, we use prompts of 80 occupations that exhibit bias toward specific demographic groups (Shrestha et al., 2024; Friedrich et al., 2023) and generate 10,000 images (*i.e.*, 125 per prompt). Specifically, we employ the template "Photo of a [OCCUPATION]" to create prompts such as "Photo of a doctor". Additional details are provided in Appendix 5.

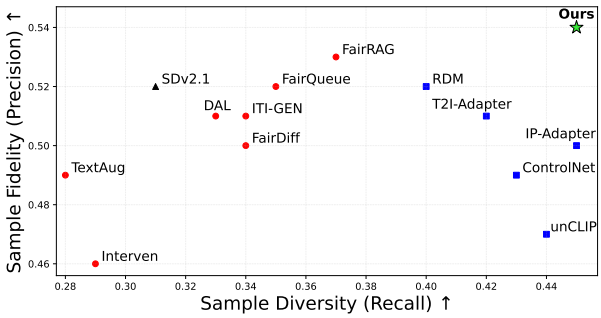
Metrics For demographic diversity, we use intersectional diversity (Shrestha et al., 2024), calculated as the normalized entropy of unique demographic groups categorized by sex, age, and skin tone, and individual diversity (Shrestha et al., 2024) to measure each attribute separately. For sample diversity, we use Recall (Kynkäänniemi et al., 2019) and W1KP (Tang et al., 2024). We use a fixed set of samples from

Table 1: Comparison of our method with existing methods. **Bold** indicates the best results, and underline indicates the second-best results.

Category	Method	Demographic Diversity ↑				Sample Diversity		Sample Fidelity		Sample Quality		Alignment
		Sex	Age	Skin Tone	Intersec.	W1KP ↓	Recall ↑	IS ↑	Precision ↑	FID ↓	$F1_{PR}$ ↑	CLIP ↑
Baseline	SDv2.1 Rombach et al. (2022)	0.27	0.22	0.22	0.19	0.88	0.31	22.81	0.52	27.87	0.39	23.63
Existing Demographic Diversity Methods	Interven Bansal et al. (2022)	0.45	0.44	0.36	0.33	0.85	0.29	19.93	0.46	32.71	0.36	23.19
	TextAug Ding et al. (2021)	0.77	0.43	0.33	0.34	0.80	0.28	16.03	0.49	30.81	0.36	23.02
	FairDiff Friedrich et al. (2023)	0.37	0.23	0.22	0.20	0.83	0.34	22.40	0.50	27.77	0.40	**23.97**
	ITI-GEN Zhang et al. (2023a)	0.80	0.56	0.38	0.44	0.70	0.34	20.56	0.51	34.02	0.41	21.44
	FairQueue Teo et al. (2024)	0.82	0.56	0.41	**0.47**	0.72	0.35	21.23	0.52	30.43	0.42	22.65
	DAL Shen et al. (2024)	**0.83**	**0.57**	**0.42**	0.43	0.79	0.33	**23.50**	0.51	28.77	0.40	23.12
	FairRAG Shrestha et al. (2024)	0.80	0.56	**0.42**	0.44	0.61	0.37	20.88	0.53	26.82	0.44	23.63
Image-Conditioned Methods Adapted for Diversity Enhancement	RDM Blattmann et al. (2022)	0.43	0.22	0.38	0.23	0.57	0.40	18.52	0.52	25.24	0.45	23.78
	unCLIP Ramesh et al. (2022)	0.50	0.27	0.26	0.23	0.72	0.44	18.09	0.47	41.33	0.45	21.22
	IP-Adapter Ye et al. (2023)	0.55	0.22	0.27	0.22	0.65	**0.45**	19.88	0.50	26.38	0.47	22.28
	T2I-Adapter Mou et al. (2024)	0.37	0.24	0.16	0.16	0.70	0.42	18.55	0.51	62.55	0.46	22.54
	ControlNet Zhang et al. (2023b)	0.43	0.29	0.18	0.20	0.65	0.43	18.49	0.49	58.25	0.46	22.60
	Ours	0.82	**0.57**	**0.42**	**0.47**	**0.54**	**0.45**	22.45	**0.54**	**23.18**	**0.49**	23.05



(a) Existing methods fail to enhance both demographic and sample diversity.



(b) Image-conditioned methods increase sample diversity but lack fidelity.

Figure 4: Our method (top-right) enhances both demographic diversity and sample diversity **simultaneously** while preserving sample fidelity, whereas existing methods can only improve either demographic diversity or sample diversity while sacrificing sample fidelity.

MSCOCO that do not overlap with the reference image dataset as real samples to compute metrics involving real samples (Brock, 2018a). Additionally, we use W1KP, a metric independent of real samples, to evaluate sample diversity and prompt reusability, as it is more effective than other diversity metrics (*e.g.*, LPIPS (Zhang et al., 2018) and ST-LPIPS (Ghildyal and Liu, 2022)) (Tang et al., 2024). We use Precision (Kynkäänniemi et al., 2019) and Inception Score (IS) (Salimans et al., 2016) to evaluate sample fidelity. We use FID (Heusel et al., 2017) to assess overall sample quality, as it captures both diversity and fidelity (Dhariwal and Nichol, 2021; Karras et al., 2019; 2020). Additionally, we calculate the harmonic mean of Precision and Recall, denoted as $F1_{PR}$, to comprehensively evaluate both sample diversity and sample fidelity. We use CLIP Score (Radford et al., 2021) to evaluate the alignment between the user prompt and the actual content of the generated image.

4.2 Comparison with Existing Demographic Diversity Methods

We compare our method with an extensive list of existing demographic diversity methods in Tab. 1 and Figs. 4 and 5. Since most compared methods are built on SDv2.1, we use it as the backbone in this comparison, and later evaluate our method on other backbones in Fig. 8, including the SOTA SDv3.5-L (Esser et al., 2024). In Tab. 1, our method outperforms all demographic diversity methods in improving sample diversity, while having on-par or better demographic diversity and sample fidelity. For instance, compared to FairRAG (Shrestha et al., 2024), which retrieves reference images using p, our method achieves high sample diversity by using description d, as more candidates match d than p. Interestingly, our method also surpasses TAP-based methods (Bansal et al., 2022; Ding et al., 2021; Friedrich et al., 2023) in demographic diversity,

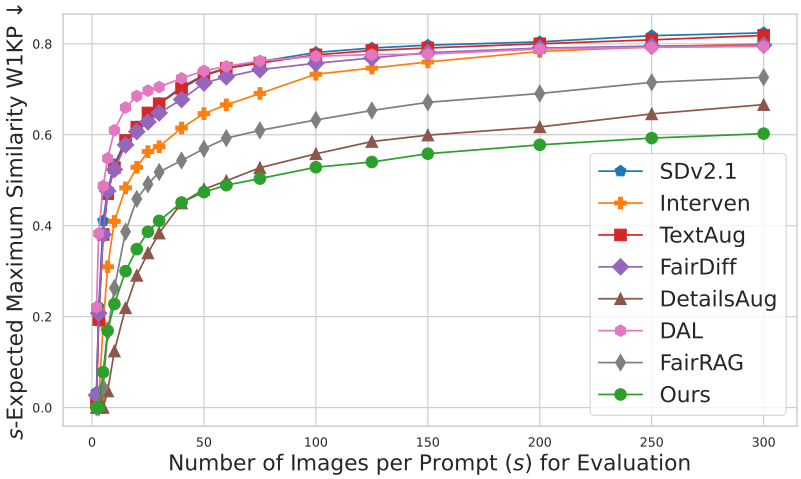


Figure 5: Prompt reusability evaluated by W1KP Tang et al. (2024) (lower is better). Our method performs the best compared to other methods.

particularly in skin tone, possibly because skin tone is difficult to specify in text but easily conveyed through images (Zhang et al., 2023a). Moreover, our method achieves comparable performance in sample fidelity to the baseline (*e.g.*, precision 0.54 *vs.* 0.52 and FID 23.18 *vs.* 27.87). We attribute the FID improvement to increased sample diversity, as FID captures both diversity and fidelity. In Fig. 5, where we evaluate prompt reusability across representative methods with greater sample diversity, our method performs the best.

4.3 Adapting Image-Conditioned Methods for Enhancing T2I Diversity

Despite a lack of existing solutions for enhancing sample diversity in T2I generation, it might be possible to construct such a solution by simply combining existing image-conditioned methods with a retrieval strategy. We implement such a combination as follows: for any given user prompt, we first retrieve reference images using a diverse pre-existing dataset and then provide the reference images as prompts to image-conditioned generative methods to synthesize the final image. We consider several SOTA image-conditioned methods, including RDM (Blattmann et al., 2022), unCLIP (Ramesh et al., 2022), IP-Adapter (Ye et al., 2023), T2I-Adapter (Mou et al., 2024) and ControlNet (Zhang et al., 2023b)), under this pipeline and study their effectiveness in enhancing sample diversity while preserving sample fidelity. In Fig. 4 and Tab. 1, we observe that they can increase sample diversity, but it comes at the cost of reducing fidelity, whereas our proposed method enhances both demographic and sample diversity without compromising fidelity. To elucidate the source of their limitation, in Fig. 6 we show how these methods rely strongly on the relevance between user prompts and retrieved reference images, which can degrade sample quality in generating using large-scale diverse datasets due to limited relevant references. In contrast, our method uses BCFG to separately handle text and image modalities and prioritize text guidance for greater robustness to unrelated content in reference images.

4.4 Alleviating the Diversity-Fidelity Trade-Off

We study BCFG in alleviating the diversity/fidelity trade-off by comparing it with Classifier-Free Guidance (CFG) (Ho and Salimans, 2022) across guidance scales ω from 1 to 20. For BCFG, we set $\omega_p = 7.5$ following the setting in (Ho and Salimans, 2022) and vary ω_i from 1 to 20. For CFG, we use two settings: (1) CFG (*Text Only*) (Eq. (3)), which excludes image conditioning, and (2) CFG (*Text & Image*), which modifies Eq. (3) to include image conditioning and a unified guidance scale for both modalities. In Fig. 7, methods incorporating reference images improve sample diversity over CFG (Text Only). BCFG achieves higher diversity than CFG (Text & Image) while preserving fidelity, thereby alleviating the trade-off.

Figure 6: Example outputs from our method and general image-conditioned methods across various occupation categories. Unlike existing methods often produce images that misalign with the user prompt due to unrelated content in reference images, our method generates prompt-aligned samples, thereby enhancing diversity while maintaining text-image alignment. Refer to Tab. 1 for quantitative comparison.

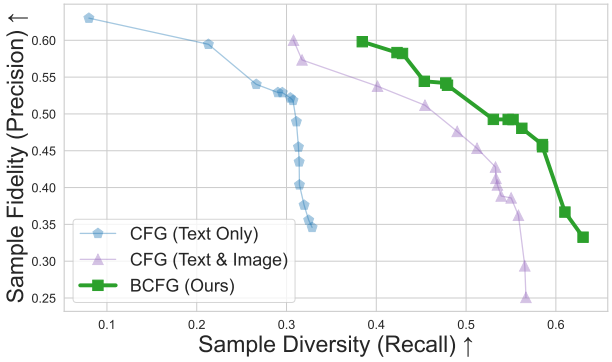


Figure 7: Comparison between guidance methods across different guidance scales for **alleviating the diversity-fidelity trade-off**. Our method achieves higher diversity while preserving fidelity.

4.5 Application to Various Pre-Trained Models

We demonstrate the compatibility of our method by applying it to various pre-trained T2I backbones. In Fig. 8, our method achieves the most significant diversity improvement on SDv3.5-Large (Esser et al., 2024), the latest SD version which exhibits the lowest diversity, highlighting its necessity. Furthermore, as newer SOTA T2I models (Esser et al., 2024; DeepFloyd Lab at StabilityAI, 2023; Black Forest Labs, 2024) tend to yield reduced diversity due to increasing model size (Rassin et al., 2024), we demonstrate the effectiveness of our method on the more challenging task of enhancing diversity in models with better original sample diversity (*e.g.*, SDv1.4 and SDv1.1).

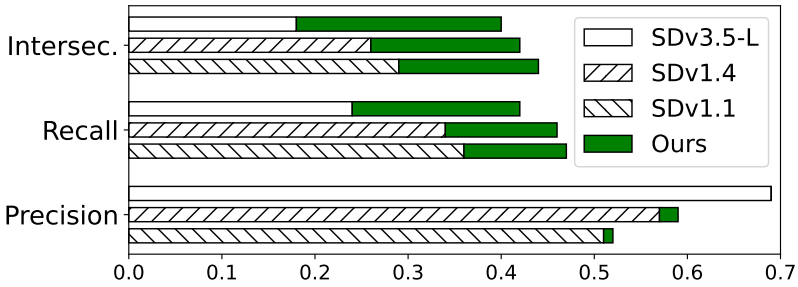
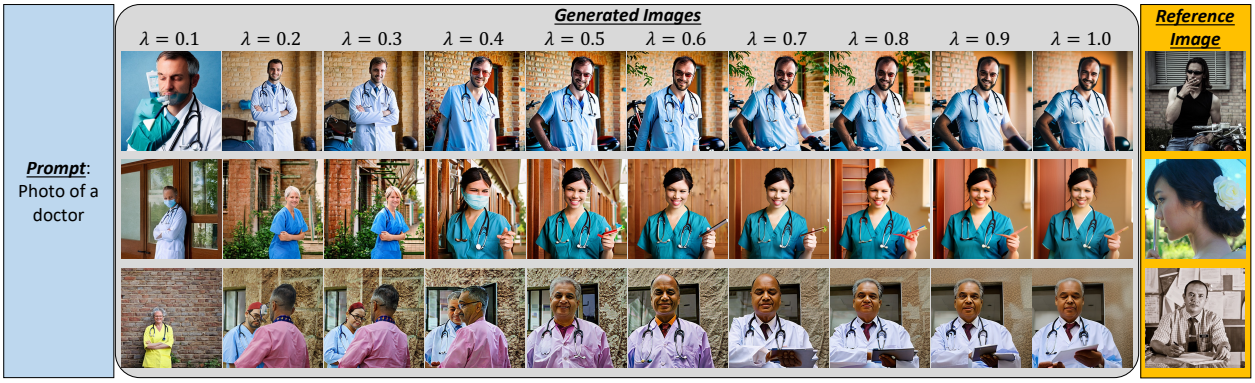
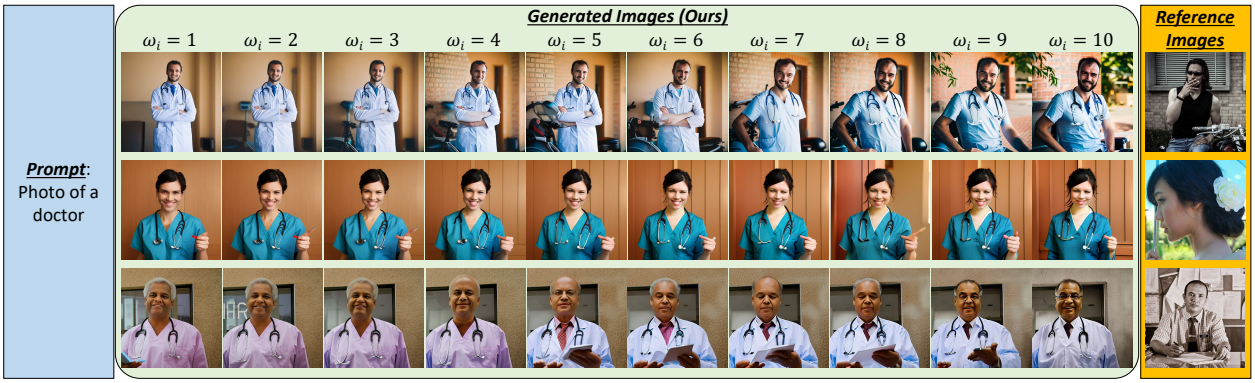


Figure 8: **Application to Various Pre-Trained Models.** Our method achieves the most significant diversity improvement on the latest SD model and even enhances diversity on models with better original sample diversity, all while preserving sample fidelity.



(a) Varying image scale parameter $\lambda^{\text{inference}}$ during inference while keeping image scale parameter $\lambda^{\text{training}}$ fixed at 1.0 during training.



(b) Varying image guidance scale ω_i in Bimodal Classifier-Free Guidance.

Figure 9: Comparison between control using the image scale parameter λ in Decoupled Cross-Attention Ye et al. (2023) and control using the image guidance scale ω_i in Bimodal Classifier-Free Guidance.

4.6 Comparison with Other Methods for Controlling Image Modality Strength

In this section, we compare the control of image modality strength using the image guidance scale ω_i in the proposed BCFG with other potential methods. Specifically, aside from our proposed method, we examine two alternative approaches: 1) adjusting the image scale parameter λ (*e.g.*, scaling Attention$(\mathbf{Q}, \mathbf{K}_i, \mathbf{V}_i)$ in Eq. (1)) in the original Decoupled Cross-Attention (DCA) (Ye et al., 2023), and 2) applying simple scalar multiplication α to the image embedding $\mathbf{e}_i$. Note that aligning the settings of these approaches between

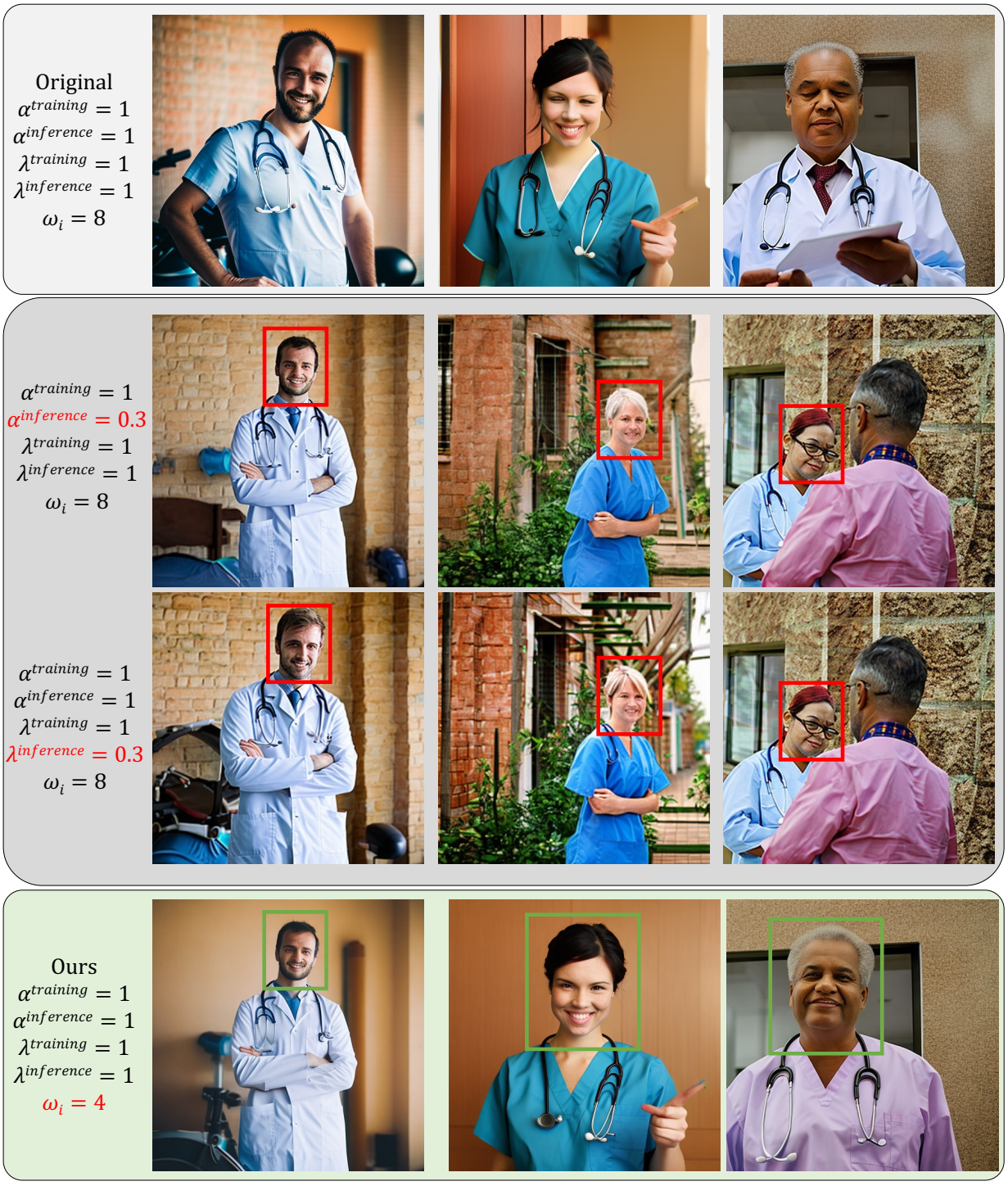


Figure 10: Visual comparison of different approaches for controlling image modality strength during inference: 1) (second row) **applying scalar multiplication α to the image embedding $\mathbf{e}_i$**, 2) (third row) **adjusting the image scale parameter λ in the original DCA Ye et al. (2023)**, and 3) (fourth row) the proposed image guidance scale ω_i in BCFG. Misalignment of α^{training} and $\lambda^{\text{training}}$ between training and inference may lead to flaws (highlighted in the red box), such as overexposure or reduced resolution Lin et al. (2024b). In contrast, the image guidance scale ω_i in our method, a training-free hyper-parameter, effectively controls image modality strength by avoiding the misalignment.

training and inference typically requires retraining, which is time-intensive and limits user control flexibility. To achieve comparable effects to our method, which allows user control over image modality strength without retraining, we fix $\lambda^{\text{training}}$ and α^{training} at 1.0 during training, and vary $\lambda^{\text{inference}}$ and $\alpha^{\text{inference}}$ during inference. As compared in Figs. 9a and 10, adjusting $\lambda^{\text{inference}}$ or $\alpha^{\text{inference}}$ during inference may introduce discrepancies between training and inference, potentially leading to issues such as overexposure or reduced resolution (Lin et al., 2024b). Nevertheless, as shown in Fig. 9b, the image guidance scale ω_i in BCFG, a training-free hyperparameter, effectively controls the strength of the image modality while preserving image quality. Furthermore, it demonstrates robustness across a wide range of guidance scales.

Table 2: Applying BCFG to Various Conditioning Modules. BCFG can enhance demographic and sample diversity across various conditioning modules while preserving sample fidelity.

Method	Demographic Diversity ↑				Sample Diversity		Sample Fidelity		Sample Quality		Alignment
	Sex	Age	Skin Tone	Intersec.	W1KP ↓	Recall ↑	IS ↑	Precision ↑	FID ↓	$F1_{PR}$ ↑	CLIP ↑
SDv2.1 Rombach et al. (2022)	0.27	0.22	0.22	0.19	0.88	0.31	**22.81**	0.52	27.87	0.39	**23.63**
LP (CFG)	0.75	0.51	0.40	0.44	0.60	0.43	17.13	0.42	30.70	0.42	21.33
LP (BCFG)	0.75	0.56	0.40	**0.47**	0.57	0.44	20.88	0.53	25.41	0.48	23.05
DCA (CFG)	0.77	0.54	**0.44**	0.41	**0.51**	0.51	15.57	0.44	27.52	0.47	21.91
DCA (BCFG)	0.76	0.55	0.42	**0.47**	0.59	0.45	17.60	**0.54**	25.44	**0.49**	22.09
LoRIA (CFG)	0.81	0.52	0.41	0.42	0.53	**0.53**	16.61	0.41	26.06	0.46	22.20
LoRIA (BCFG)	**0.82**	**0.57**	0.42	**0.47**	0.54	0.45	22.45	**0.54**	**23.18**	**0.49**	23.05

Table 3: Ablation study of our proposed method.

Method	Intersec. ↑	Recall ↑	Precision ↑	FID ↓	CLIP ↑
SDv2.1	0.19	0.31	0.52	27.87	**23.63**
TextAug	0.34	0.28	0.49	30.81	23.02
DetailsAug	0.43	0.46	0.43	30.86	21.20
Ablated Variants of Our Method					
With TextAug	**0.47**	0.44	0.49	31.98	22.46
W/o LoRIA	**0.47**	0.44	0.53	25.41	22.17
W/o DRS	0.18	0.44	0.53	23.58	23.39
W/o BCFG	0.42	**0.53**	0.41	26.06	22.20
Training w/o only $\mathbf{e}_i$	0.23	0.35	0.52	26.12	23.58
Reference (CelebA)	0.46	0.40	0.52	25.23	23.49
Ours	**0.47**	0.45	**0.54**	**23.18**	23.05

4.7 Applying BCFG to Various Conditioning Modules

In this section, we examine the proposed Bimodal Classifier-Free Guidance (BCFG) as an independent technique and showcase its applicability across different conditioning modules. Specifically, we apply BCFG to our proposed conditioning module (*e.g.*, Low-Rank Image Adapter) and other image conditioning modules such as *Concatenation + Simple Linear Projection* (LP) (Zhao et al., 2024) and *Decoupled Cross-Attention* (DCA) (Ye et al., 2023). In Tab. 2, we observe that BCFG can enhance both demographic and sample diversity across various conditioning modules while maintaining sample fidelity. These results empirically demonstrate that BCFG can be applied independently of the specific conditioning module.

4.8 Ablation Study

In Tab. 3, LoRIA enables bimodal conditioning while maintaining prompt alignment, DRS improves demographic diversity, and BCFG enhances sample diversity while preserving fidelity. In Fig. 3, the image guidance scale ω_i in BCFG, a training-free hyperparameter, can control image modality strength while preserving quality. In contrast, adjusting the image scale in (Ye et al., 2023) or rescaling $\mathbf{e}_i$ during inference may cause discrepancy between training and inference, leading to overexposure or low resolution (Lin et al.,

2024b) (Sec. 4.6). We also evaluate our method combined with TextAug (*With TextAug*) by conditioning on (p, d, i). This setup slightly increases diversity but compromises fidelity, possibly because reference images already encode demographic information, making added dummy information by d suboptimal. To compare with TAP, we construct *Details-Augmented Prompt* (DetailsAug) (Esposito et al., 2023; Datta et al., 2023), where we instruct ChatGPT-o1 (OpenAI, 2024) (see instruction in Appendix 4) to generate a detailed text description dd based on the user prompt, including variations in sex, age, skin tone, location, and camera settings. Our method (p, i) achieves comparable diversity while preserving FID and CLIP, but DetailsAug (p, dd) compromises text-image alignment, likely because added context in DetailsAug shifts focus from the original user prompt, occasionally leading to unrelated outputs (Hao et al., 2024). Moreover, excluding model training on cases with only visual token $\mathbf{e}_i$ (*Training w/o only* $\mathbf{e}_i$) significantly reduces demographic diversity, dropping Intersectional Diversity from 0.47 to 0.23. This confirms that our training strategy effectively encodes demographic information into $\mathbf{e}_i$, as the *only* $\mathbf{e}_i$ case forces the model to rely solely on images for noise removal, requiring maximal extraction of image information into $\mathbf{e}_i$. Additionally, using lower-quality retrieval databases (*e.g.*, CelebA (Liu et al., 2018)) slightly lowers fidelity and quality but maintains competitive diversity, demonstrating the robustness of our method across various reference databases. In Appendix 3, we present a detailed ablation on the number of retrieval and reference images in DRS, additional BCFG guidance scale results, and an analysis of LoRIA regarding LoRA Rank.

5 Conclusion

We propose a lightweight and efficient method to enhance both demographic and sample diversity while preserving fidelity in diffusion models. Extensive empirical results demonstrate its effectiveness in various pre-trained models.

Limitations and Future Directions. While our method shows promising results in enhancing demographic diversity, it relies on the assumption that the bias attribute is known. Thus, a potential direction is to extend it to address unknown biases (Li et al., 2022) by incorporating a bias detection model (*e.g.*, B2T (Kim et al., 2024)) to identify visual biases in T2I models. Moreover, our method leverages IAP to enhance demographic diversity rather than mitigating bias in T2I model itself (Esposito et al., 2023). Another promising direction is to develop T2I models that are less sensitive to biases in the training dataset. Besides, replacing the closed-source database with an open-source one could be beneficial, as it offers more inclusive knowledge to further enhance diversity (Fan et al., 2024).

References

Jimmy Lei Ba. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 843–852, 2023.

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

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *OpenAI Blog*, 2023.

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

Black Forest Labs. FLUX.1-dev Model Documentation. https://huggingface.co/black-forest-labs/FLUX.1-dev, 2024. Accessed: Aug 24, 2024.

Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. Retrieval-augmented diffusion models. *Advances in Neural Information Processing Systems*, 35:15309–15324, 2022.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR, 2022.

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

Will Brennan. Semanticsegmentation, 2024. Accessed: 2024-09-28.

Andrew Brock. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018a.

Andrew Brock. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018b.

Deng Cai, Yan Wang, Lemao Liu, and Shuming Shi. Recent advances in retrieval-augmented text generation. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 3417–3419, 2022.

Hanqun Cao, Cheng Tan, Zhangyang Gao, Yilun Xu, Guangyong Chen, Pheng-Ann Heng, and Stan Z Li. A survey on generative diffusion models. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022.

Jaemin Cho, Abhay Zala, and Mohit Bansal. 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*, pages 3043–3054, 2023.

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

Kate Crawford. *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press, 2021.

Siddhartha Datta, Alexander Ku, Deepak Ramachandran, and Peter Anderson. Prompt expansion for adaptive text-to-image generation. *arXiv preprint arXiv:2312.16720*, 2023.

DeepFloyd Lab at StabilityAI. DeepFloyd IF: a novel state-of-the-art open-source text-to-image model with a high degree of photorealism and language understanding. https://www.deepfloyd.ai/deepfloyd-if, 2023. Retrieved on 2023-11-08.

Jiankang Deng, Jia Guo, Zhou Yuxiang, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-stage dense face localisation in the wild. In *arxiv*, 2019.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in neural information processing systems*, 34:19822–19835, 2021.

Jiawei Du, Xin Zhang, Juncheng Hu, Wenxin Huang, and Joey Tianyi Zhou. Diversity-driven synthesis: Enhancing dataset distillation through directed weight adjustment. *arXiv preprint arXiv:2409.17612*, 2024.

Dave Epstein, Allan Jabri, Ben Poole, Alexei Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. *Advances in Neural Information Processing Systems*, 36:16222–16239, 2023.

Piero Esposito, Parmida Atighehchian, Anastasis Germanidis, and Deepti Ghadiyaram. Mitigating stereotypical biases in text to image generative systems. *arXiv preprint arXiv:2310.06904*, 2023.

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

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6491–6501, 2024.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

Kathleen C Fraser, Svetlana Kiritchenko, and Isar Nejadgholi. Diversity is not a one-way street: Pilot study on ethical interventions for racial bias in text-to-image systems. *ICCV, accepted*, 2023.

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

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5111–5120, 2024.

Abhijay Ghildyal and Feng Liu. Shift-tolerant perceptual similarity metric. In *European Conference on Computer Vision*, 2022.

Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1820–1828, 2021.

Xu Guo and Yiqiang Chen. Generative ai for synthetic data generation: Methods, challenges and the future. *arXiv preprint arXiv:2403.04190*, 2024.

Yaru Hao, Zewen Chi, Li Dong, and Furu Wei. Optimizing prompts for text-to-image generation. *Advances in Neural Information Processing Systems*, 36, 2024.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proc. of the 30th International Conf. on Neural Information Processing Systems*, page 3323–3331, Red Hook, NY, USA, 2016. Curran Associates Inc.

Ruifei He, Chuhui Xue, Haoru Tan, Wenqing Zhang, Yingchen Yu, Song Bai, and Xiaojuan Qi. Debiasing text-to-image diffusion models. In *Proceedings of the 1st ACM Multimedia Workshop on Multi-modal Misinformation Governance in the Era of Foundation Models*, pages 29–36, 2024.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

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

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

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Kimmo Karkkainen and Jungseock Joo. 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*, pages 1548–1558, 2021.

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

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

Younghyun Kim, Sangwoo Mo, Minkyu Kim, Kyungmin Lee, Jaeho Lee, and Jinwoo Shin. Discovering and mitigating visual biases through keyword explanation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11082–11092, 2024.

Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github. com/openimages*, 2(3):18, 2017.

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

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

Ailin Li, Lei Zhao, Zhiwen Zuo, Zhizhong Wang, Haibo Chen, Dongming Lu, and Wei Xing. Diversified text-to-image generation via deep mutual information estimation. *Computer Vision and Image Understanding*, 211:103259, 2021.

Ailin Li, Lei Zhao, Zhiwen Zuo, Zhizhong Wang, Wei Xing, and Dongming Lu. Specific diverse text-to-image synthesis via exemplar guidance. *IEEE MultiMedia*, 2024a.

Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. Controllable text-to-image generation. *Advances in neural information processing systems*, 32, 2019.

Jia Li, Lijie Hu, Jingfeng Zhang, Tianhang Zheng, Hua Zhang, and Di Wang. Fair text-to-image diffusion via fair mapping. *arXiv preprint arXiv:2311.17695*, 2023a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023b.

Shufan Li, Harkanwar Singh, and Aditya Grover. Popalign: Population-level alignment for fair text-to-image generation. *arXiv preprint arXiv:2406.19668*, 2024b.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023c.

Zhiheng Li, Anthony Hoogs, and Chenliang Xu. Discover and mitigate unknown biases with debiasing alternate networks. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII*, pages 270–288. Springer, 2022.

Kuan Heng Lin, Sicheng Mo, Ben Klingher, Fangzhou Mu, and Bolei Zhou. Ctrl-x: Controlling structure and appearance for text-to-image generation without guidance. *arXiv preprint arXiv:2406.07540*, 2024a.

Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 5404–5411, 2024b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

Steven Liu, Tongzhou Wang, David Bau, Jun-Yan Zhu, and Antonio Torralba. Diverse image generation via self-conditioned gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018.

I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Analyzing societal representations in diffusion models. *arXiv preprint arXiv:2303.11408*, 2023.

Hanjun Luo, Haoyu Huang, Ziye Deng, Xuecheng Liu, Ruizhe Chen, and Zuozhu Liu. Bigbench: A unified benchmark for social bias in text-to-image generative models based on multi-modal llm. *arXiv preprint arXiv:2407.15240*, 2024.

David Marwood, Shumeet Baluja, and Yair Alon. Diversity and diffusion: Observations on synthetic image distributions with stable diffusion. *arXiv preprint arXiv:2311.00056*, 2023.

Zichen Miao, Jiang Wang, Ze Wang, Zhengyuan Yang, Lijuan Wang, Qiang Qiu, and Zicheng Liu. Training diffusion models towards diverse image generation with reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10844–10853, 2024.

Ellis Monk. Monk skin tone scale, 2019.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. 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*, pages 4296–4304, 2024.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, pages 7176–7185. PMLR, 2020.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

OpenAI. Blog. https://openai.com/o1/, 2024.

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

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Mirrorgan: Learning text-to-image generation by redescription. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1505–1514, 2019.

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

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

Royi Rassin, Aviv Slobodkin, Shauli Ravfogel, Yanai Elazar, and Yoav Goldberg. Grade: Quantifying sample diversity in text-to-image models. *arXiv preprint arXiv:2410.22592*, 2024.

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

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M Weber. Cads: Unleashing the diversity of diffusion models through condition-annealed sampling. *arXiv preprint arXiv:2310.17347*, 2023.

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. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

Christoph Schuhmann. Laion-aesthetics: Predicting aesthetic quality of images, 2022. Accessed: 2024-09-27.

Sefik Serengil and Alper Ozpinar. A benchmark of facial recognition pipelines and co-usability performances of modules. *Journal of Information Technologies*, 17(2):95–107, 2024.

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

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

James Seale Smith, Yen-Chang Hsu, Lingyu Zhang, Ting Hua, Zsolt Kira, Yilin Shen, and Hongxia Jin. Continual diffusion: Continual customization of text-to-image diffusion with c-lora. *arXiv preprint arXiv:2304.06027*, 2023.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Raphael Tang, Xinyu Zhang, Lixinyu Xu, Yao Lu, Wenyan Li, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. Words worth a thousand pictures: Measuring and understanding perceptual variability in text-to-image generation. *arXiv preprint arXiv:2406.08482*, 2024.

Christopher T. H Teo, Milad Abdollahzadeh, Xinda Ma, and Ngai man Cheung. Fairqueue: Rethinking prompt learning for fair text-to-image generation, 2024.

Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. Sketch-guided text-to-image diffusion models. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023.

Yixin Wan and Kai-Wei Chang. The male ceo and the female assistant: Probing gender biases in text-to-image models through paired stereotype test. *arXiv preprint arXiv:2402.11089*, 2024.

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

Yixin Wan, Di Wu, Haoran Wang, and Kai-Wei Chang. The factuality tax of diversity-intervened text-to-image generation: Benchmark and fact-augmented intervention. *arXiv preprint arXiv:2407.00377*, 2024b.

Shijian Wang, Linxin Song, Ryotaro Shimizu, Masayuki Goto, et al. Attributed synthetic data generation for zero-shot image classification. In *Synthetic Data for Computer Vision Workshop@ CVPR 2024*, 2024.

Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2256–2265, 2021.

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

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

Mariia Zameshina, Olivier Teytaud, and Laurent Najman. Diverse diffusion: Enhancing image diversity in text-to-image generation. *arXiv preprint arXiv:2310.12583*, 2023.

Yan Zeng, Hanbo Zhang, Jiani Zheng, Jiangnan Xia, Guoqiang Wei, Yang Wei, Yuchen Zhang, Tao Kong, and Ruihua Song. What matters in training a gpt4-style language model with multimodal inputs? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7930–7957, 2024.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. Investigating the catastrophic forgetting in multimodal large language models. *arXiv preprint arXiv:2309.10313*, 2023.

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

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

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023c.

Zhenxing Zhang and Lambert Schomaker. Divergan: An efficient and effective single-stage framework for diverse text-to-image generation. *Neurocomputing*, 473:182–198, 2022.

Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, 2024.

Yucheng Zhou, Xiang Li, Qianning Wang, and Jianbing Shen. Visual in-context learning for large vision-language models. *arXiv preprint arXiv:2402.11574*, 2024.