FairImagen: Post-Processing for Bias Mitigation in Text-to-Image Models

Zihao Fu
The Chinese University of Hong Kong
zihaofu@cuhk.edu.hk

Ryan Brown
University of Oxford
blac0977@ox.ac.uk

Shun Shao
University of Cambridge
ss3047@cam.ac.uk

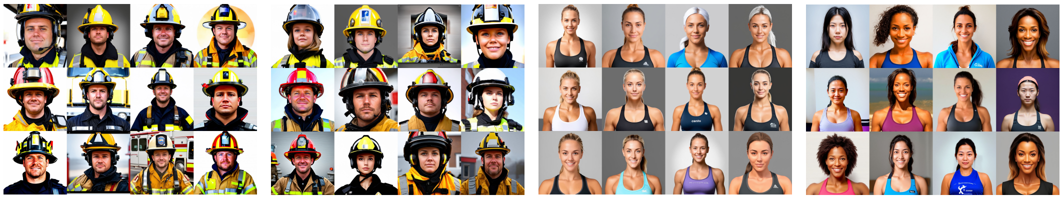
Kai Rawal
University of Oxford
kaivalyarawal45@gmail.com

Eoin Delaney
Trinity College Dublin
eoin.delaney@tcd.ie

Chris Russell
University of Oxford
chris.russell@oii.ox.ac.uk

Abstract

Text-to-image diffusion models, such as Stable Diffusion, have demonstrated remarkable capabilities in generating high-quality and diverse images from natural language prompts. However, recent studies reveal that these models often replicate and amplify societal biases, particularly along demographic attributes like gender and race. In this paper, we introduce **FairImagen**[1], a post-hoc debiasing framework that operates on prompt embeddings to mitigate such biases without retraining or modifying the underlying diffusion model. Our method integrates Fair Principal Component Analysis to project CLIP-based input embeddings into a subspace that minimizes group-specific information while preserving semantic content. We further enhance debiasing effectiveness through empirical noise injection and propose a unified cross-demographic projection method that enables simultaneous debiasing across multiple demographic attributes. Extensive experiments across gender, race, and intersectional settings demonstrate that FairImagen significantly improves fairness with a moderate trade-off in image quality and prompt fidelity. Our framework outperforms existing post-hoc methods and offers a simple, scalable, and model-agnostic solution for equitable text-to-image generation.



Firefighter (Base) Gender Debias (FairImagen) Yoga Instructor (Base) Race Debias (FairImagen)

Figure 1: FairImagen mitigates demographic biases in text-to-image generation. Compared to baseline Stable Diffusion (Base), our method (FairImagen) produces more balanced representations across demographic attributes such as gender and race, while preserving visual quality and semantic fidelity.

1 Introduction

Recent advances in text-to-image generation have led to the widespread adoption of models such as Stable Diffusion [1, 2], DALL·E [3, 4], Imagen [5], and Parti [6]. These can produce photorealistic

[1] https://github.com/fuzihaofzh/FairImagen

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

and diverse images from natural language prompts. These models leverage powerful vision-language encoders such as CLIP [7] to align textual and visual modalities, enabling open-ended image generation from arbitrary input. Due to their flexibility, scalability, and accessibility, these systems are increasingly integrated into applications across design, content creation, and interactive media [8, 9].

However, studies have shown that these generative models often replicate and even amplify social biases present in the training data [10–18]. For example, prompts such as: "a photo of a CEO" or "a nurse" typically yield images depicting white males and females, respectively, reflecting gender and racial stereotypes. These biases pose serious concerns regarding fairness, representation, and downstream harms, particularly as generative models are integrated into public-facing systems.

To mitigate such biases, researchers have proposed a variety of debiasing techniques. Methods fall into three main categories: **Prompt-based**, **Fine-tuning**, and **Post-hoc editing** (Table 1). Prompt-based approaches [10, 19, 20] modify the input to influence the model's output, but often require per image heuristic rewriting and manually-curated prompts. Fine-tuning methods [21, 13, 22] retrain or adapt parts of the model to encode fairness objectives, but they are computationally intensive and require access to model internals. Post-hoc editing methods [23, 24, 13] modify prompt embeddings at inference without updating model weights, offering a lightweight and deployment-friendly alternative. Each category exhibits differing trade-offs in terms of fidelity, interpretability, and generalizability.

We focus on **post-hoc editing** methods due to their simplicity and compatibility with a wide range of off-the-shelf diffusion models. Approaches such as SDID [23] and TBIE [24] demonstrated the feasibility of manipulating prompt embeddings to mitigate demographic bias. SDID identified a gender direction by subtracting CLIP embeddings of pairs of hand-crafted prompts; adding or subtracting this vector in generation. This heavily relies on the group bias being linearly separable and correctable via modifying embeddings in a single direction. This is inappropriate for debiasing involving more than two demographic groups, e.g., ethnicity. TBIE improves over SDID by applying PCA on CLIP embeddings of gender-related words to identify bias directions in a data-driven way. However, PCA is performed on simple gendered words without an explicit optimization criterion for content alignment. As a result, the debiasing process is often overly aggressive, removing not just demographic cues but also key semantic information. This results in semantic drift, loss of prompt fidelity, and unnatural image generation. Both methods offer limited control over the trade-off between fairness and fidelity, and neither generalizes well to intersectional prompts or unseen groups.

To overcome these limitations, we propose a novel post-hoc debiasing framework we call **FairImagen** (Fair Image Generation). It explicitly integrates Fair Principal Component Analysis (FairPCA) [25] into the Stable Diffusion pipeline and optimizes for semantic preservation while minimizing group-dependent variance. FairImagen operates in three stages: first, it extracts CLIP-based prompt embeddings; second, it applies a fairness-aware projection using FairPCA to remove group-dependent directions from both pooled and token-level embeddings; finally, it synthesizes images from the transformed embeddings using a modified Stable Diffusion decoder. To further enhance performance, we incorporate an empirical noise injection scheme to avoid overly neutralized outputs, and propose a unified cross-demographic debiasing formulation to jointly mitigate intersectional bias. Unlike existing post-hoc approaches, FairImagen offers precise control over the trade-off between fairness and content alignment. It is fully compatible with off-the-shelf diffusion models, supports multiple demographic attributes simultaneously, and preserves visual quality while effectively reducing bias.

Our contributions are as follows: (1) We introduce a post-hoc fairness framework that integrates FairPCA with diffusion-based text-to-image generation, enabling bias mitigation without model retraining. (2) We propose empirical noise injection to obscure residual demographic signals and improve fairness-performance trade-offs. (3) We develop a cross-demographic debiasing formulation that handles multiple protected attributes in a unified manor, avoiding over-pruning from sequential projections. (4) We conduct extensive quantitative and qualitative evaluations across gender, race, and joint debiasing tasks, demonstrating that our method outperforms existing post-hoc baselines.

2 Related Works

Existing debiasing methods for text-to-image generation can be categorized into three types: Prompt-based, Fine-tuning-based, and Training-free methods. As summarized in Table 1, no single category is universally superior; the choice of method often depends on the specific application scenario and deployment constraints.

Criteria	Prompt-based	Fine-tuning	Post-hoc editing
Training-free	✓	✗	✓
Black-box compatible	✓	✗	✓
Low human effort	✗	✓	✓
Low computational cost	✓	✗	✓
Generalizable to new prompts	✗	✓	✓
Strong bias mitigation	✗	✓	✓
Preserves prompt fidelity	✓	✓	✗
Easy deployment	✗	✗	✓

Table 1: Comparison of prompt-based, fine-tuning-based, and post-hoc editing methods for debiasing text-to-image generation.

Prompt-based methods mitigate bias by modifying the input prompts. Friedrich et al. [10] proposed Fair Diffusion using fairness-guided prompts constructed from demographic opposites. Sakurai and Sato [19] utilized LLMs to automatically detect and revise biased prompts. Bansal et al. [20] and Chuang et al. [26] examined ethical interventions and latent direction projection. Kim et al. [27] and Al Sahili et al. [28] developed learned fairness prompts, and Bianchi et al. [29] assessed the impact of biased prompts at scale. These methods are flexible but often rely on heuristic or external guidance for every single image. This can be somewhat opaque and laborious [20, 13].

Fine-tuning based methods update model parameters to enforce fairness. Li et al. [21] introduced Fair Mapping by training a linear projection layer. Zhang et al. [13] aligned prompt embeddings with fair visual examples. Shen et al. [22] applied a distributional alignment loss for fairness. Kim et al. [27], Orgad et al. [30], and Gandikota et al. [31] proposed fine-tuning specific modules or applying concept editing. Parihar et al. [32] incorporated interpretable latent directions and population-level optimization, respectively. These methods provide effective bias mitigation but often require costly model access and retraining.

Post-hoc editing methods avoid parameter updates and modify inference behavior. Zhang et al. [13] and Li et al. [23] manipulated prompt embeddings with CLIP-based or interpretable directions. Tanjim et al. [24] used PCA to subtract biased components. Friedrich et al. [10] employed classifier-free guidance alternations. Sadat et al. [33] explored sampling noise perturbation and conditioning annealing to reveal underrepresented concepts. Post-hoc filtering is also employed in some commercial systems [3]. These methods are deployment-friendly, take advantage of both prompt- and model-based strategies, and avoid extensive retraining or heavy prompt engineering.

3 Method

We propose a fairness-aware text-to-image generation framework, **FairImagen**. This framework integrates FairPCA [25] into Stable Diffusion [1, 2]. Our goal is to reduce social bias in image generation by modifying prompt embeddings prior to synthesis, while preserving semantic fidelity.

To estimate and remove demographic information from prompt embeddings, we begin with a small training set of natural language prompts, each annotated with protected attributes such as gender or race (e.g., “a lady playing computer,” “an Asian man holding a phone”). These prompts are used to construct a FairPCA projection matrix that suppresses group-specific directions while retaining core semantic content. The learned projection is then applied at inference time to unseen prompts, making FairImagen entirely training-free with respect to the diffusion model.

The framework consists of three main components: (1) Prompt Embedding Extractor, (2) Fair Representation Transformer, and (3) Image Generator. In addition, we propose an empirical noise injection scheme to prevent overly neutralized outputs and a unified cross-demographic debiasing formulation to jointly mitigate intersectional bias.

3.1 Prompt Embedding Extraction

To construct the FairPCA projection, we begin with a small training set of natural language prompts $\mathcal{P} = \{p_1, \ldots, p_n\}$, each annotated with a protected attribute label $a_i \in \mathcal{A}$. These prompts are designed to be demographically informative yet semantically neutral (e.g., “a lady playing computer,” “a Black man riding a bike”), and serve as the foundation for identifying group-dependent components in the embedding space.

Given a prompt $p \in \mathcal{P}$, we first encode it using a pre-trained CLIP model [7]. Let $\{w_1, \ldots, w_T\}$ be the tokenized prompt, where T is the number of tokens. The encoder outputs a token-level embedding matrix $E_p \in \mathbb{R}^{T \times D}$, where D is the embedding dimension, and a pooled embedding $\bar{E}_p \in \mathbb{R}^D$. The pooled embedding is computed as the mean of the token embeddings: $\bar{E}_p = \frac{1}{T} \sum_{t=1}^{T} E_p[t]$. These representations are extracted from the Stable Diffusion text encoder. Let $\mathcal{P} = \{p_1, \ldots, p_n\}$ denote a set of prompts, each associated with protected attribute labels $a_i \in \mathcal{A}$. For each attribute a, we organize the pooled embeddings by group:

$$X = \{\bar{E}_{p_i}\}_{i=1}^{n} \in \mathbb{R}^{n \times D}, \quad Z = \{z_i\}_{i=1}^{n} \in \{0,1\}^{n \times G},$$

where z_i is a one-hot group indicator for the attribute a_i, and $G = |\mathcal{A}|$. These grouped embeddings are used to estimate the bias direction and define fairness-aware projections.

3.2 Fair Representation Transformer

We use Principal Component Analysis (PCA) to approximate the original prompt embedding space with a lower-dimensional subspace that preserves semantic information. Specifically, we seek a projection matrix that can faithfully reconstruct the original embeddings from a reduced set of basis directions.

Let $P \in \mathbb{R}^{D \times d}$ be a projection matrix, where $d < D$. Classical PCA solves the following reconstruction objective:

$$\underset{P \in \mathbb{R}^{D \times d} : P^\top P = I}{\arg\min} \sum_{i=1}^{n} \left\| \mathbf{x}_i - PP^\top \mathbf{x}_i \right\|_2^2 \equiv \underset{P \in \mathbb{R}^{D \times d} : P^\top P = I}{\arg\max} \operatorname{Tr}(P^\top X X^\top P), \tag{1}$$

where $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^\top \in \mathbb{R}^{n \times D}$ is the matrix of pooled prompt embeddings. The left-hand side minimizes the total squared reconstruction error of projecting the data onto a d-dimensional subspace, while the right-hand side expresses the equivalent trace maximization formulation.

To further ensure fairness, we also require that the projection removes demographic signals by aligning with the null space of group-dependent variation. Specifically, we define the group-dependent feature matrix $B = Z^\top X \in \mathbb{R}^{G \times D}$, where $Z \in \{0,1\}^{n \times G}$ is the group indicator matrix. Each row of B captures the mean embedding direction for a specific demographic group. By constraining the projection matrix P to lie in the null space $\mathcal{N}(B)$, we ensure that the resulting representations are orthogonal to any direction that separates groups in the embedding space, thereby eliminating linear demographic signals.

Building on this intuition, FairPCA [25] incorporates a fairness regularization term into the PCA objective, yielding the following formulation:

$$\min_{P^\top P = I} -\operatorname{Tr}(P^\top \Sigma_X P) + \lambda \|BP\|_F^2, \tag{2}$$

where λ is a hyperparameter controlling the trade-off between reconstruction quality and fairness. The first term ensures that the projection subspace remains a good approximation of the original feature space. The second term penalizes the degree to which the projected embeddings retain group-specific components, thereby reducing demographic separability. Once the projection matrix is obtained, we apply it during inference to both the pooled and token-level prompt embeddings:

$$\bar{E}'_p = PP^\top \bar{E}_p, \quad E'_p = E_p PP^\top.$$

3.3 Empirical Noise Injection

To further enhance diversity and realism, we introduce an empirical noise injection mechanism that perturbs representations along estimated group-dependent directions. It prevents the generated output from becoming overly neutral (e.g., generating a man who appears feminine; see further discussion in Appendix E). Let $\mathcal{G}$ denote the set of protected groups (e.g., $\mathcal{G} = \{\text{Male}, \text{Female}\}$), and let $g \in \mathcal{G}$ be a particular group. For each group g, we compute its empirical bias direction as

$$v_g = \frac{1}{|X^{(g)}|} \sum_{\bar{E}_p \in X^{(g)}} \bar{E}_p - \bar{E},$$

where $X^{(g)}$ is the set of pooled embeddings belonging to group g, and $\bar{E}$ is the overall mean embedding across all groups. We define an empirical distribution $\mathcal{D}_g$ as the set of scalar projections of group-specific embeddings onto the bias direction:

$$\mathcal{D}_g = \left\{ v_g^\top \bar{E}_p : \bar{E}_p \in X^{(g)} \right\}.$$

Each value $\delta \in \mathcal{D}_g$ represents the magnitude of projection of an embedding onto the bias direction v_g, quantifying how strongly that embedding aligns with group-specific attributes. To inject noise in the inference stage, we sample $\delta \sim \mathcal{D}_g$ and apply the perturbation:

$$\bar{E}_p'' = \bar{E}_p' + \epsilon \cdot \delta \cdot v_g,$$

where ϵ is a tunable noise scale parameter. This procedure introduces controlled variability along biased directions to obscure protected group information while preserving semantic structure. We also conduct experiments with additional noise injection strategies, including mean-based, Gaussian, fixed-directional, and deterministic shift variants. A detailed comparison of these methods and their impact on fairness and image quality is provided in Appendix D. Among these strategies, empirical noise demonstrates the best overall performance.

3.4 Cross-Demographics Debiasing

The FairPCA framework [25] debiases multiple demographics by jointly encoding them into a multi-dimensional attribute matrix. It minimizes group-specific information across all attributes simultaneously via a single projection derived from a stacked group indicator matrix. However, when applied to image generation, this approach fails to adequately represent all demographics, as it forces features to be orthogonal to each group direction. Consequently, the model tends to preserve information aligned with only one group at a time, resulting in degraded contextual fidelity and the loss of important visual details in the generated images.

To overcome this, we propose a unified cross-demographic debiasing method that constructs a single attribute space based on the Cartesian product of all group combinations. For example, if the gender attribute has two groups $\{\text{Male}, \text{Female}\}$ and the race attribute has three groups $\{\text{White}, \text{Asian}, \text{Black}\}$, we define a joint attribute space with six composite groups: $\mathcal{A}_{\text{joint}} = \{\text{White Male}, \text{White Female}, \text{Asian Male}, \text{Asian Female}, \text{Black Male}, \text{Black Female}\}$. We then apply Fair Representation Transformer once over this joint attribute space. Therefore, our cross-demographic debiasing approach can debias all demographics simultaneously. We have also conducted experiments with alternative strategies for handling multiple demographic attributes, including stacking and sequential projection. For a comprehensive comparison of these cross-demographic debiasing methods, please refer to Appendix G.

3.5 Image Generator

After debiasing, we pass the transformed embeddings into a customized Stable Diffusion pipeline [1, 2], which supports external prompt embeddings. Specifically, we generate the image as: $I_p = \mathcal{G}(\bar{E}_p'', E_p')$, where $\mathcal{G}(\cdot)$ denotes the generation function, and $\bar{E}_p''$ and E_p' are the pooled and token-level debiased embeddings.

4 Experiments

4.1 Dataset

The Winobias dataset consists of 46 professions, collected from the US Bureau of Labor Statistics, that are stereotypically considered "male biased" or "female biased" [31, 30, 34, 35]. In our experiments, we extend this list to 120 professions using publicly available lists[2]. Our list covers professions that have been found to be biased towards men (e.g., Janitor or CEO) and women (e.g., Nurse or Librarian) in generative AI systems [36].

[2] The full list is included in our supplementary material with the code. We manually extended the winobias list using a publicly available list of occupations from Wikipedia: `https://en.wikipedia.org/wiki/Lists_of_occupations`

4.2 Experimental Settings

Our modified pipeline extends HuggingFace's StableDiffusion3Pipeline to accept external embeddings and apply FairImgen at inference time. We generate images using classifier-free guidance with scale $w = 7.0$ and $T = 28$ diffusion steps. Images are generated in batches (12 per prompt), stitched, and evaluated with fairness and perceptual quality metrics. We split the dataset into a development set of 20 samples and use the remaining 100 samples as the test set. We tune all models on the development set to maximize the average (AVG) score and report their performance on the test set. We run all the models on a NVIDIA A100 GPU with 80 GB memory.

4.3 Evaluation

We report four scalar metrics: Fairness, Accuracy, MUSIQ, and their average.

Fairness is monitored using a lightweight pretrained facial attribute classifier from DeepFace [37]. This classifier detects and counts members of each group in every generated image. Next the distribution of counts is scored with

$$1 - \frac{\sum_i |p_i - \frac{1}{k}|}{2(1 - \frac{1}{k})}$$

following the normalized-deviation formulation of Teo et al. [38]. Here p is a vector of group proportions and k is the number of groups. A score of 1 indicates that all groups are generated at the same rate,while a score of 0 indicates that only one group is generated (e.g., all men when considering gender as the protected attribute and prompted to generate images of a CEO). We measure gender-based groups of male and female individuals alongside ethnicity-based groups of asian, black, latino hispanic, middle eastern, and white individuals. We also consider intersectional groups of gender and ethnicity throughout the paper.

Accuracy is measured using CLIPScore [39], which quantifies how closely a generated image matches its text prompt. Specifically, we compute the cosine similarity between the prompt embedding and the image embedding produced by CLIP (ViT-B/16 backbone). Following best practices reported by Hessel el al. [39], we multiply the CLIPScore values by 2.5 for scaling as the original scores typically fall between 0 and 0.4.

To assess the visual quality of an image, we use MUSIQ [40], which is a no-reference perceptual-quality model trained on millions of aesthetic ratings. This metric was selected due to its flexibility, as it can work with native resolution images with varying sizes and aspect ratios.

4.4 Comparison Models

Base is the vanilla Stable Diffusion model, which directly generates images from the prompt without any fairness intervention.

FairPrompt follows [10, 19] by using human-designed prompts for each image. We evenly apply different prompts corresponding to protected groups for each individual image. This serves as an upper-bound performance baseline, as each prompt is specifically tailored for fairness.

ForcePrompt explicitly includes fairness-related instructions (e.g. "Please avoid gender bias.") in the prompt, directing the Stable Diffusion model to generate fair representations.

SAL [41, 42] uses Singular Value Decomposition to project the input representations into directions with reduced covariance with the biases.

CDA (Counterfactual Data Augmentation) [43, 44] replaces gendered words with their counterfactual counterparts, such as replacing "man" with "woman." We follow the CDA methodology to construct counterfactual samples and augment the dataset.

TBIE (Text-Based Image Editing) [24] applies PCA to gender-related words and performs debiasing along the identified principal components.

SDID (Self-Discovering Interpretable Diffusion) [23] computes a gender vector using the difference between gender-specific and gender-neutral embeddings, and injects this vector into the prompt embedding.

	Gender Fairness	Accuracy	MUSIQ	Avg
Base	0.167	0.785	0.574	0.509
FairPrompt	0.732	0.766	0.586	0.695
ForcePrompt	0.292	0.755	0.601	0.549
SAL	0.217	0.779	0.602	0.533
CDA	0.547	0.772	0.549	0.623
TBIE	0.35	0.782	0.567	0.566
SDID	0.507	0.776	0.553	0.612
SDID-AVG	0.315	0.783	0.562	0.553
ITI	0.27	0.769	0.528	0.522
FairQueue	0.197	0.809	0.621	0.542
FairImagen	0.56	0.771	0.541	0.624
FairImagen-T5	0.572	0.768	0.533	0.624
FairImagen-OC	0.573	0.767	0.534	0.625

(a) Gender Debias.

	Race Fairness	Accuracy	MUSIQ	Avg
Base	0.193	0.785	0.574	0.517
FairPrompt	0.444	0.752	0.566	0.587
ForcePrompt	0.266	0.761	0.574	0.534
SAL	0.262	0.788	0.607	0.552
CDA	0.358	0.772	0.537	0.556
TBIE	0.366	0.762	0.532	0.553
SDID	0.37	0.77	0.537	0.559
SDID-AVG	0.361	0.769	0.544	0.558
ITI	0.214	0.77	0.53	0.504
FairQueue	0.118	0.736	0.631	0.495
FairImagen	0.389	0.76	0.536	0.562
FairImagen-T5	0.386	0.76	0.537	0.561
FairImagen-OC	0.387	0.76	0.536	0.561

(b) Race Debias.

	Gender Fairness	Race Fairness	Accuracy	MUSIQ	Avg
Base	0.163	0.193	0.785	0.574	0.508
FairPrompt	0.69	0.478	0.747	0.574	0.671
ForcePrompt	0.287	0.304	0.764	0.591	0.547
SAL	0.182	0.214	0.776	0.599	0.519
CDA	0.362	0.27	0.779	0.557	0.566
TBIE	0.40	0.286	0.776	0.546	0.574
SDID	0.223	0.256	0.782	0.556	0.52
SDID-AVG	0.352	0.28	0.778	0.553	0.561
ITI	0.32	0.235	0.747	0.467	0.511
FairQueue	0.0567	0.34	0.773	0.606	0.478
FairImagen	0.537	0.32	0.753	0.544	0.611
FairImagen-T5	0.48	0.31	0.766	0.532	0.593
FairImagen-OC	0.482	0.311	0.766	0.532	0.593

(c) Gender + Race Debias.

Table 2: Quantitative evaluation results for debiasing text-to-image generation across three settings: (a) gender, (b) race, and (c) both gender and race. FairImagen achieves the best overall performance among post-hoc methods, striking a strong balance between fairness, accuracy, and perceptual quality.

SDID-AVG extends the SDID [23] model by computing neutral embeddings through averaging the embeddings within each protected group.

ITI-GEN [13] extracts gender-related CLIP embeddings from images and adds them to the prompt embeddings prior to image generation.

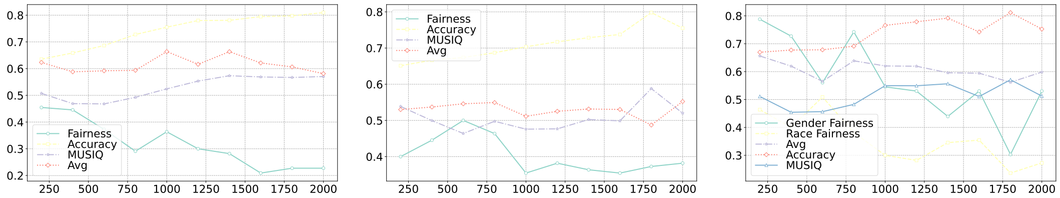
FairQueue [45] rethinks prompt learning approaches by identifying abnormalities in early denoising steps. It proposes Prompt Queuing (using base prompts without sensitive attribute tokens in initial steps) and Attention Amplification (enhancing attribute representation in later steps) to modify cross-attention maps during generation, achieving competitive fairness while improving image quality.

FairImagen-T5/OC: To evaluate if our method is compatible with encoders beyond CLIP, we evaluate using alternative text encoders. **FairImagen-T5** replaces the CLIP text encoder with T5 [46], while **FairImagen-OC** uses OpenCLIP [47] instead of the original CLIP encoder. These variants demonstrate the generalizability of our fairness framework across different text encoding architectures.

4.5 Experimental Results

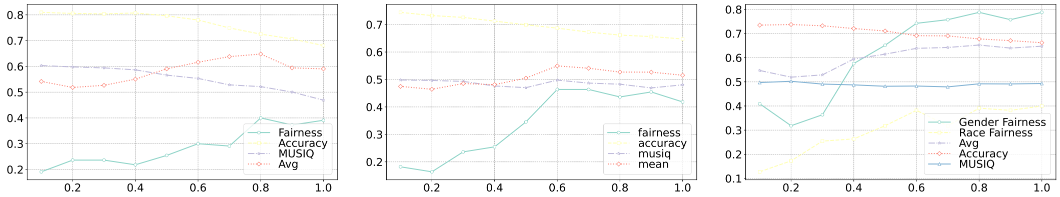
4.5.1 Main Experiments. We apply our model, as well as other baseline models, to generation tasks involving debiasing with respect to gender (§4.5.1), race (§4.5.1), and both gender and race simultaneously (§4.5.1). The results show that: (1) our proposed FairImagen model outperforms all postprocessing baseline models in terms of fairness scores across all three scenarios, demonstrating its effectiveness in mitigating bias in various contexts; (2) our proposed FairImagen model also outperforms all postprocessing models in terms of average (AVG) scores, indicating that it achieves the best balance among fairness, accuracy, and image quality; and (3) our model consistently outperforms all postprocessing baselines when debiasing both gender and race simultaneously, highlighting its strong capability in addressing multi-attribute bias mitigation. (4) Our proposed FairImagen model slightly lags behind other models in terms of Accuracy and MUSIQ. However, given the substantial improvement in Fairness, this trade-off is justified and considered worthwhile. (5) FairPrompt achieves the best performance across all experiments. It should be noted, however, that this model relies on manually designed prompts tailored to each individual image, which is both time-consuming and labor-intensive. As such, it serves primarily as an upper bound to illustrate the best possible performance a model can achieve on this task.

4.5.2 Effect of Hidden Dimension on FairImagen Performance. To investigate the impact of dimensionality reduction on the effectiveness of FairImagen, we vary the number of retained principal components (i.e., hidden dimensions) from 200 to 2000. Figure 2 shows performance trends across three scenarios: (a) gender debiasing, (b) race debiasing, and (c) joint gender and race debiasing. The results demonstrate a clear trade-off between fairness and other metrics as dimensionality varies. Notably, reducing the number of components tends to improve fairness scores, particularly in gender and race separately, but at the cost of reduced Accuracy and MUSIQ. In contrast, larger hidden dimensions preserve visual and semantic fidelity better, but may reintroduce bias. The joint debiasing setting (Figure 2c) further reveals the challenge of balancing fairness across multiple attributes simultaneously, with Fairness metrics for gender and race sometimes diverging. Overall, these



(a) Gender Debiasing (b) Race Debiasing (c) Gender + Race Debiasing

Figure 2: Effect of hidden dimension size on FairImagen performance across different debiasing settings: (a) gender debiasing, (b) race debiasing, and (c) joint gender and race debiasing. Reducing the number of retained dimensions improves fairness but may reduce Accuracy and MUSIQ, highlighting the trade-off between fairness and semantic or visual fidelity.



(a) Gender Debiasing (b) Race Debiasing (c) Gender + Race Debiasing

Figure 3: Effect of e-noise on FairImagen performance across different debiasing settings: (a) gender debiasing, (b) race debiasing, and (c) joint gender and race debiasing. Increasing e-noise enhances fairness, especially in the joint setting, but introduces a moderate drop in Accuracy and MUSIQ. This highlights the need to balance fairness improvements with fidelity preservation when tuning noise.

results indicate that non-linear biases are more likely to occur in higher-dimensional spaces, and points towards the importance of selecting an appropriate dimensionality to achieve a desirable fairness-fidelity trade-off.

4.5.3 Effect of Empirical Noise on FairImagen Debiasing. We further examine the influence of the empirical noise parameter (e-noise) in the FairImagen framework, which controls the magnitude of perturbation added to simulate empirical distributions aligned with different protected attributes. As e-noise increases, the sampled directions more closely follow gender- or race-specific variations, enabling stronger debiasing effects. Figure 3 presents the performance across gender, race, and joint gender+race debiasing tasks under varying noise levels from 0.0 to 1.0. Higher e-noise values significantly improve fairness metrics, particularly in the joint debiasing setting, where both gender and race fairness steadily increase. However, this comes with a modest trade-off in Accuracy and MUSIQ, suggesting a balance must be struck between fairness gains and fidelity preservation. These results validate the effectiveness of controlled empirical noise injection in sampling representative latent directions that better approximate the real distribution of protected attributes.

4.5.4 Qualitative Case Study. To qualitatively assess our debiasing strategies, we examine the prompt "Generate a photo of a face of a CEO" under four FairImagen settings: Base, FairImagen-Gender, FairImagen-Race, and FairImagen-Gender+Race. As shown in Figure 4, the Base model predominantly generates white male faces, reflecting societal bias. FairImagen-Gender (Figure 4b) increases female representation while maintaining visual coherence. FairImagen-Race (Figure 4c) introduces more racial diversity, including Black and Asian CEOs. Combining both dimensions (Figure 4d) yields broader diversity but also introduces visual inconsistencies in background and style, due to stronger empirical noise. This also highlights the trade-off: enhancing fairness can compromise visual stability and prompt fidelity.

4.5.5 Evaluation on Occupations with Man/White Dominance. To evaluate debiasing performance, we analyze occupation prompts that exhibit strong male and white biases in the baseline model. Figure 5 shows that the Base model produces predominantly male outputs, while FairImagen and FairPrompt significantly improve gender balance. Similarly, Figure 6 reveals strong white dominance

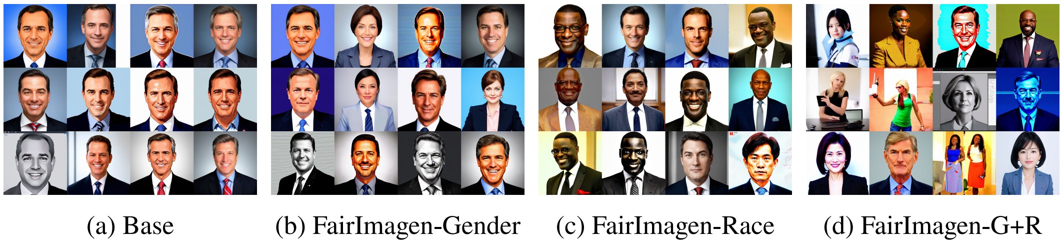

(a) Base (b) FairImagen-Gender (c) FairImagen-Race (d) FairImagen-G+R

Figure 4: Generated results for the prompt “Generate a photo of a face of a CEO” under four FairImagen settings: (a) Base (no debiasing), (b) FairImagen-Gender, (c) FairImagen-Race, and (d) FairImagen-Gender+Race. Debiasing increases demographic diversity across gender and race dimensions. However, stronger debiasing—especially under intersectional settings—can introduce variation in background and style, reflecting a trade-off between fairness and visual consistency.

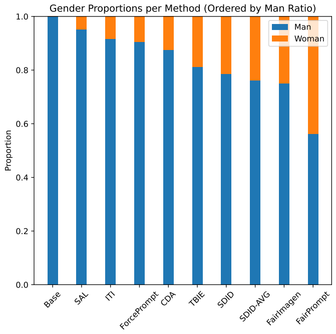


Figure 5: Gender proportions for male-dominated occupations. Each bar shows the proportion of male and female outputs generated by different methods, sorted by male ratio. FairImagen and FairPrompt substantially reduce male overrepresentation compared to baselines.

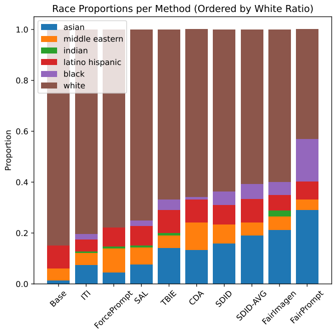


Figure 6: Race proportions for white-dominated occupations. Each bar shows the proportion of racial groups in the generated outputs, sorted by white ratio. FairImagen and FairPrompt noticeably reduce white overrepresentation and enhance racial diversity.

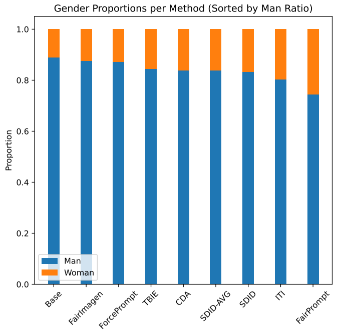
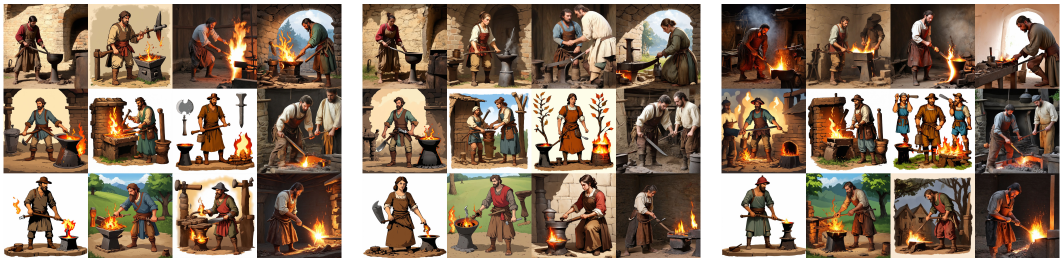


Figure 7: Gender distribution for prompts with historically male-associated roles. While most methods retain male-dominant outputs, FairPrompt introduces more females, contradicting historical facts. In contrast, FairImagen preserves the intended gender associations.

in the Base model, with FairImagen and FairPrompt increasing representation of Black, Asian, and Latino Hispanic individuals. These results demonstrate that FairImagen effectively mitigates demographic bias in skewed prompts, achieving fairness comparable to FairPrompt while remaining model-agnostic and training-free.

4.5.6 Robustness to Demographically Determined Prompts. A major challenge in fairness-aware generative modeling is to ensure that debiasing methods do not compromise semantic fidelity, particularly when prompts inherently reflect justified demographic attributes. In real-world use cases, certain prompts—such as those referencing historical figures or culturally specific roles—are expected to yield outputs with a specific gender association. Overcorrecting in such cases may lead to semantically incongruent or historically inaccurate generations, undermining user trust and model reliability.[3] Therefore, it is essential for fairness interventions to be context-aware and capable of preserving prompt intent when the bias is grounded in legitimate semantics. To this end, we evaluate whether FairImagen can maintain semantic alignment when prompts exhibit strongly determined gender associations. We focus on examples such as “a middle ages blacksmith”, “the Pope”, and “the King of France”, which traditionally imply male representations. Figure 7 shows that across these historically gender-fixed prompts, most models continue to generate predominantly male outputs. Notably, FairPrompt slightly increases the proportion of female representations, even in male-dominant

[3] See, e.g., `https://www.nytimes.com/2024/02/22/technology/google-gemini-german-uniforms.html` and `https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical`.



(a) Base (b) FairPrompt (c) FairImagen

Figure 8: Generated results for the prompt “a middle ages blacksmith” using three different methods: (a) Base, (b) FairPrompt, and (c) FairImagen. FairPrompt introduces female representations even when the prompt implies a male role, while FairImagen preserves the intended gender semantics, yielding historically aligned outputs.

contexts. Figure 8 presents qualitative comparisons of images generated for the blacksmith prompt using the Base, FairPrompt, and FairImagen models. While FairPrompt introduces female depictions regardless of the prompt’s semantics, FairImagen respects the gender bias encoded in the original embedding and yields predominantly male outputs. This demonstrates a key strength of FairImagen: when a prompt conveys a strong and contextually justified gender preference, FairImagen does not override it unnecessarily. As such, FairImagen adapts to prompt intent while still being effective in mitigating bias in less explicitly gendered scenarios.

5 Conclusion

We present FairImagen, a novel post-hoc debiasing framework for text-to-image generation that integrates FairPCA into the Stable Diffusion pipeline. Our method modifies prompt embeddings to mitigate demographic biases without requiring model retraining or prompt rewriting. Through a fairness-aware projection, empirical noise injection, and a unified cross-demographic formulation, FairImagen achieves strong bias reduction results while preserving visual fidelity and prompt alignment. Extensive experiments across gender, race, and intersectional attributes demonstrate that our approach outperforms existing post-hoc baselines on both fairness and utility metrics. By offering a training-free, model-agnostic, and extensible solution, FairImagen paves the way for more equitable and controllable generative systems.

Acknowledgements

This work was supported through research funding provided by an EPSRC Doctoral Scholarship, the Wellcome Trust (grant no. 223765/Z/21/Z), Sloan Foundation (grant no. G-2021-16779), Department of Health and Social Care, EPSRC (grant no. EP/Y019393/1), Origin Investments, and Luminate Group.

References

[1] 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.

[2] 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.

[3] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

[4] 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.

[5] 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.

[6] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

[7] 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.

[8] 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.

[9] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. Text-to-image diffusion models in generative ai: A survey. *arXiv preprint arXiv:2303.07909*, 2023.

[10] 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 at arXiv:2302.10893*, 2023.

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

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

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

[14] Pushkar Shukla, Aditya Chinchure, Emily Diana, Alexander Tolbert, Kartik Hosanagar, Vineeth N Balasubramanian, Leonid Sigal, and Matthew A Turk. Biasconnect: Investigating bias interactions in text-to-image models. *arXiv preprint arXiv:2503.09763*, 2025.

[15] Wenxuan Wang, Haonan Bai, Jen-tse Huang, Yuxuan Wan, Youliang Yuan, Haoyi Qiu, Nanyun Peng, and Michael Lyu. New job, new gender? measuring the social bias in image generation models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3781–3789, 2024.

[16] Aadi Chauhan, Taran Anand, Tanisha Jauhari, Arjav Shah, Rudransh Singh, Arjun Rajaram, and Rithvik Vanga. Identifying race and gender bias in stable diffusion ai image generation. In *2024 IEEE 3rd International Conference on AI in Cybersecurity (ICAIC)*, pages 1–6. IEEE, 2024.

[17] Shah Prerak. Addressing bias in text-to-image generation: A review of mitigation methods. In *2024 Third International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN)*, pages 1–6. IEEE, 2024.

[18] Eunji Kim, Siwon Kim, Minjun Park, Rahim Entezari, and Sungroh Yoon. Rethinking training for de-biasing text-to-image generation: Unlocking the potential of stable diffusion. *arXiv preprint arXiv:2408.12692*, 2024.

[19] Jinya Sakurai and Issei Sato. Fair2i: Mitigating social bias in text-to-image generation via large language model-assisted detection and attribute rebalancing. *arXiv preprint arXiv:2502.03826*, 2025.

[20] Hritik Bansal, Da Yin, Masoud Monajatipoor, and Kai-Wei Chang. How well can text-to-image generative models understand ethical natural language interventions? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1358–1370, 2022.

[21] 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*, 2023.

[22] Xudong Shen, Chao Du, Tianyu Pang, Min Lin, Yongkang Wong, and Mohan Kankanhalli. Finetuning text-to-image diffusion models for fairness. *arXiv preprint arXiv:2311.07604*, 2023.

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

[24] Md Mehrab Tanjim, Krishna Kumar Singh, Kushal Kafle, Ritwik Sinha, and Garrison W Cottrell. Discovering and mitigating biases in clip-based image editing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2984–2993, 2024.

[25] Matthäus Kleindessner, Michele Donini, Chris Russell, and Muhammad Bilal Zafar. Efficient fair pca for fair representation learning. In *International Conference on Artificial Intelligence and Statistics*, pages 5250–5270. PMLR, 2023.

[26] 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.

[27] Eunji Kim, Siwon Kim, Chaehun Shin, and Sungroh Yoon. De-stereotyping text-to-image models through prompt tuning. 2023.

[28] Zahraa Al Sahili, Ioannis Patras, and Matthew Purver. Equiprompt: Debiasing diffusion models via iterative bootstrapping in chain of thoughts. *arXiv e-prints*, pages arXiv–2406, 2024.

[29] 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.

[30] Hadas Orgad, Bahjat Kawar, and Yonatan Belinkov. Editing implicit assumptions in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7053–7061, 2023.

[31] 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.

[32] Rishubh Parihar, Abhijnya Bhat, Abhipsa Basu, Saswat Mallick, Jogendra Nath Kundu, and R Venkatesh Babu. Balancing act: distribution-guided debiasing in diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6668–6678, 2024.

[33] 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.

[34] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*, 2018.

[35] U.S. Bureau of Labor Statistics. Table a-11. employed persons by detailed occupation, sex, race, and hispanic or latino ethnicity. `https://www.bls.gov/cps/cpsaat11.htm`, 2025. Accessed: 2025-05-15.

[36] Nouar AlDahoul, Talal Rahwan, and Yasir Zaki. Ai-generated faces influence gender stereotypes and racial homogenization. *Scientific Reports*, 15(1):14449, Apr 2025. ISSN 2045-2322. doi: 10.1038/s41598-025-99623-3. URL `https://doi.org/10.1038/s41598-025-99623-3`.

[37] 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. doi: 10.17671/gazibtd.1399077. URL `https://dergipark.org.tr/en/pub/gazibtd/issue/84331/1399077`.

[38] Christopher T. H. Teo and Ngai-Man Cheung. Measuring fairness in generative models, July 2021. URL `http://arxiv.org/abs/2107.07754`. ICML 2021 Workshop.

[39] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning, March 2022. URL `http://arxiv.org/abs/2104.08718`.

[40] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer, August 2021. URL `http://arxiv.org/abs/2108.05997`. ICCV 2021.

[41] Shun Shao, Yftah Ziser, and Shay B. Cohen. Gold doesn't always glitter: Spectral removal of linear and nonlinear guarded attribute information. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1611–1622, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.118. URL `https://aclanthology.org/2023.eacl-main.118`.

[42] Shun Shao, Yftah Ziser, and Shay B. Cohen. Erasure of Unaligned Attributes from Neural Representations. *Transactions of the Association for Computational Linguistics*, 11:488–510, 05 2023. ISSN 2307-387X. doi: 10.1162/tacl_a_00558. URL `https://doi.org/10.1162/tacl_a_00558`.

[43] Ran Zmigrod, Sabrina J Mielke, Hanna Wallach, and Ryan Cotterell. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. *arXiv preprint arXiv:1906.04571*, 2019.

[44] Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032*, 2020.

[45] Christopher T.H Teo, Milad Abdollahzadeh, Xinda Ma, and Ngai-Man Cheung. Fairqueue: Rethinking prompt learning for fair text-to-image generation. In *Advances in Neural Information Processing Systems*, 2024.

[46] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[47] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. URL `https://doi.org/10.5281/zenodo.5143773`.

[48] Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Evaluating societal representations in diffusion models. *Advances in Neural Information Processing Systems*, 36:56338–56351, 2023.

Appendix. Supplementary Material

A Impact Statement

This work contributes to the development of fair and inclusive generative AI systems by introducing FairImagen, a training-free and model-agnostic framework for mitigating demographic bias in text-to-image diffusion models. FairImagen offers a practical and scalable solution for enhancing fairness in image generation without requiring access to model internals or manual prompt design. Its ability to jointly address multiple demographic attributes while preserving visual fidelity makes it particularly well-suited for deployment in real-world applications such as digital media, design, and educational content. By reducing the social harms associated with biased generation and enabling more representative outputs, FairImagen supports the broader goal of responsible and equitable AI deployment.

B Ethical Statement

This research aims to address ethical concerns surrounding bias and representation in text-to-image generative models. Our proposed framework, FairImagen, is designed to reduce the amplification of demographic stereotypes without compromising image quality or user intent. We acknowledge that fairness is a multi-faceted and context-dependent concept, and our method focuses primarily on gender and racial representation, which may not capture the full spectrum of social identities or cultural nuances.

We do not collect any personal or sensitive user data in our experiments. All generated images are produced from synthetic prompts, and demographic groupings are based on commonly used protected attributes in fairness research. While FairImagen mitigates certain biases, we caution against interpreting it as a complete solution to fairness in generative models. Ongoing monitoring, inclusive evaluation, and engagement with affected communities remain essential for ensuring responsible deployment.

Our code and findings will be released to the research community to promote transparency and further development of fair and accountable generative AI.

C Limitations

Despite the strengths of FairImagen as a post-hoc, training-free debiasing framework, several limitations remain. First, the method currently focuses on a limited set of protected attributes—primarily binary gender and a coarse categorization of race. As frequently noted [48], such simplifications may overlook more nuanced or intersectional demographic identities, such as non-binary gender expressions or multi-ethnic backgrounds. Second, as FairImagen operates on CLIP-based prompt embeddings, it inherits any intrinsic biases present in the CLIP encoder, which itself is trained on large-scale web data with limited curation. While FairPCA reduces group-dependent variance, it cannot fully disentangle bias that is deeply entangled with semantic meaning. Third, although empirical noise injection and projection dimensionality offer tunable fairness-utility trade-offs, determining the optimal balance often requires empirical tuning and may vary across tasks. Additionally, while the framework performs robustly across a wide range of prompts, its effectiveness may degrade for prompts that are strongly tied to cultural or historical contexts, where bias removal risks semantic distortion. Lastly, our evaluation focuses on a specific benchmark of occupational prompts; broader testing across domains, cultures, and creative settings is needed to fully validate generalizability and uncover edge cases where the method may fail.

D Effect of Noise Injection Variants

To investigate the role of noise in fairness-aware generation, we compare several noise injection schemes within the FairImagen framework. The **Empirical Noise** method (enoise) samples both direction and magnitude from real group-specific embedding distributions, introducing realistic and data-driven perturbations. **Mean Empirical Noise** uses the average projection magnitude instead

	Gender Fairness	Accuracy	MUSIQ	Avg
Empirical Noise	0.455	0.808	0.567	0.61
Mean Empirical Noise	0.432	0.788	0.472	0.564
Fixed Directional Noise	0.258	0.76	0.542	0.52
Fixed Directional Noise (b=1)	0.197	0.761	0.547	0.502
Random Gaussian Noise	0.167	0.758	0.56	0.495
Fixed Random Gaussian Noise	0.258	0.763	0.538	0.52
Constant Bias Shift	0.242	0.757	0.541	0.514
Bypass Projection	0.136	0.817	0.631	0.528

Table 3: Comparison of different noise injection strategies used in the FairImagen framework for gender debiasing. Each method perturbs prompt embeddings in distinct ways to obscure demographic signals. The results highlight the trade-off between fairness, semantic accuracy, and visual quality across noise types.

of sampling, resulting in more stable but less diverse shifts. **Fixed Directional Noise** adds binary noise (±1) along the bias direction and optionally biases the sign, simulating controlled reversals in group representation. **Random Gaussian Noise** injects direction-agnostic perturbations, using a Dirichlet-weighted average when multiple groups are involved. Its variant, **Fixed Random Gaussian Noise**, reuses a fixed noise vector for consistency. **Constant Bias Shift** applies a deterministic translation in the bias direction to all embeddings, representing a non-random intervention. Lastly, **Bypass Projection** disables the FairPCA projection and uses the original prompt embedding, serving as an ablation to isolate the effect of projection from noise injection.

The results in Table 3 show that noise type substantially influences the fairness-utility trade-off. Empirical Noise performs best overall, balancing improved fairness with minimal degradation in image quality and semantic alignment. Mean-based perturbations offer more stable behavior but slightly compromise fairness effectiveness. In contrast, direction-agnostic and fixed-noise variants underperform due to their limited alignment with demographic structures. Deterministic shifting introduces consistent but ineffective debiasing, and bypassing the projection leads to high image quality but minimal fairness improvement. These findings underscore the importance of designing context-aware, group-sensitive noise to support effective and reliable debiasing in post-hoc settings.

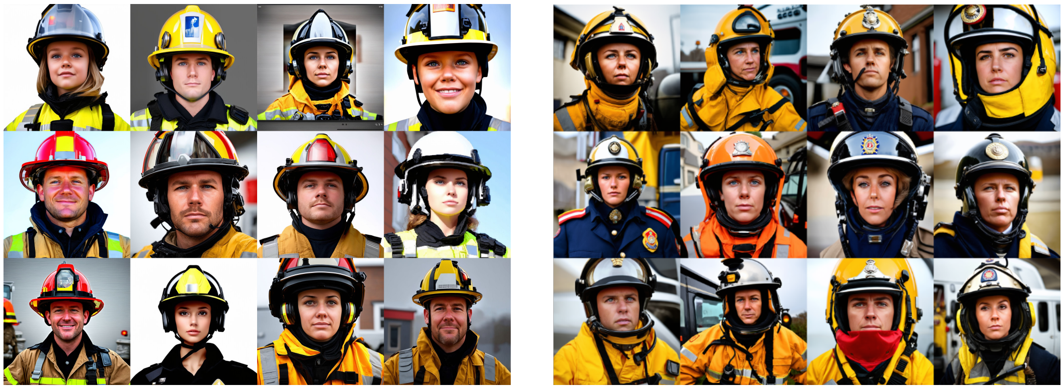
E Qualitative Impact of Empirical Noise

To further understand the role of empirical noise in enhancing fairness without overly neutralizing semantic attributes, we conduct a qualitative case study on firefighter generation. In particular, we compare two male firefighter images produced by FairImagen under different noise configurations. Both images are generated from the same prompt—“Generate a face image of a firefighter”—but differ in whether empirical noise is applied.

As shown in Figure 9, the image generated with empirical noise injection depicts a man with a beard, aligning with natural variations in male appearance. In contrast, the image generated without empirical noise produces a clean-shaven face, which appears overly neutral and lacks realistic diversity. This comparison highlights that without empirical perturbation, FairImagen may suppress group-dependent cues too aggressively, leading to sanitized outputs that obscure important intra-group variation. Empirical noise mitigates this effect by reintroducing sampled group characteristics, such as facial hair, thereby preserving authenticity while maintaining demographic balance. This illustrates the importance of using noise not merely as randomness, but as a mechanism to approximate the true variability within protected groups.

F Effect of Empirical Noise Magnitude

To better understand how empirical noise magnitude influences demographic representation, we conduct a controlled experiment in which we vary the strength of empirical noise injection in FairImagen from -5 to 15. In this setting, empirical noise is applied by perturbing the prompt embedding along the learned demographic bias direction, where the scalar magnitude modulates how strongly group-specific features (e.g., gender-related appearance) are emphasized. We generate firefighter images at each noise level and compute the gender proportions in the outputs.



(a) With Empirical Noise

(b) Without Empirical Noise

Figure 9: Visual comparison of firefighter images generated by FairImagen. (a) With empirical noise: preserves natural male features like a beard. (b) Without noise: produces an overly neutral, clean-shaven appearance.

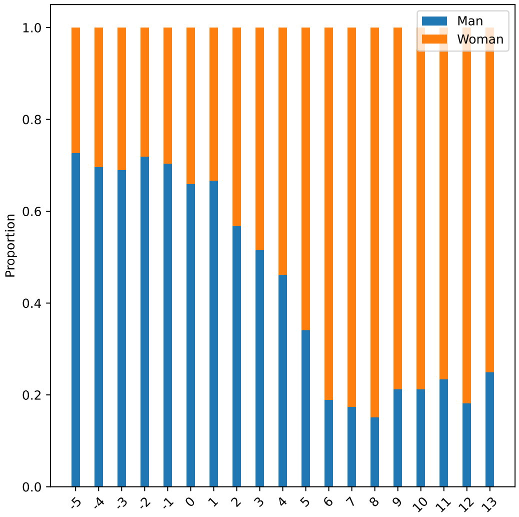


Figure 10: Effect of empirical noise magnitude on gender proportions in firefighter generation. Negative values shift the embedding toward the male-associated direction, while positive values favor female features. The curve is not centered due to male-dominance in the original training prompts.

As visualized in Figure 10, increasing the empirical noise magnitude leads to a clear transition in output gender. Negative values (corresponding to movement in the male-associated direction) produce predominantly male faces, while large positive values favor female representations. Around zero, the output becomes more balanced. However, the inflection point where the transition occurs is not perfectly centered at zero; instead, the gender ratio begins to shift significantly only at slightly positive values. This asymmetry arises because the original training prompts used to construct the FairPCA projection are themselves male-dominated, which results in a skewed embedding space where zero noise still retains residual male characteristics. Consequently, stronger positive perturbations are needed to counteract this bias and achieve gender balance.

This finding reinforces the notion that the empirical bias direction learned by FairImagen captures semantically meaningful demographic information. Moreover, it illustrates how noise magnitude can be used as a controllable parameter to modulate output demographics, with meaningful behavior emerging even from simple linear perturbations.

	Gender Fairness	Race Fairness	Accuracy	MUSIQ	Avg
Base	0.0758	0.136	0.819	0.616	0.504
FairPrompt	0.652	0.482	0.778	0.60	0.676
FairImagen (Stack)	0.227	0.345	0.799	0.569	0.532
FairImagen (Sequential)	0.106	0.345	0.771	0.525	0.467
FairImagen (Cross)	0.429	0.325	0.792	0.601	0.596

Table 4: Comparison of cross-demographic debiasing methods under joint gender and race settings. FairImagen (Cross) achieves the best fairness-utility trade-off among post-hoc methods.

G Comparison of Cross-Demographic Debiasing Strategies

To evaluate the effectiveness of different strategies for handling multiple protected attributes, we compare three cross-demographic debiasing approaches within the FairImagen framework: **Stack**, **Sequential**, and our proposed **Cross** method. The results are presented in Table 4.

The **Stack** method, originally proposed in the FairPCA paper [25], constructs a single fairness projection by stacking group indicator vectors from all demographic attributes (e.g., gender and race) into a combined group matrix. While this approach is simple and computationally efficient, it tends to over-represent groups with stronger bias signals, leading to suboptimal fairness across multiple dimensions.

The **Sequential** method applies FairPCA in multiple stages, projecting out bias directions for each protected attribute one after another. While this offers a conceptually modular way to remove group-specific information, it suffers from compounded information loss: each projection removes components orthogonal to prior ones, making it difficult to preserve semantic fidelity across multiple passes.

In contrast, our **Cross** method explicitly constructs a joint demographic space based on the Cartesian product of all attribute groups (e.g., *White Male*, *Asian Female*), and learns a unified projection that removes shared group-dependent directions in a single step. This ensures that the projection is optimized for intersectional fairness without over-pruning or repeated reconstruction loss.

As shown in Table 4, the Cross method achieves the best balance between gender and race fairness, while also maintaining competitive Accuracy and MUSIQ scores. Compared to Stack and Sequential variants, Cross substantially improves fairness without significantly compromising generation quality, highlighting its effectiveness in multi-attribute debiasing scenarios.

H Age Debiasing Experiments

To demonstrate the generalizability of FairImagen beyond binary gender and race categories, we extend our evaluation to age debiasing. Age represents a particularly challenging demographic attribute due to the inherent difficulty in distinguishing precise ages (e.g., differentiating between 35 and 38 years old in generated images). Therefore, we categorize age into three distinct groups: **young** (approximately 18-30 years), **middle-aged** (approximately 31-55 years), and **elderly** (approximately 56+ years).

The age debiasing task follows the same experimental setup as our gender and race experiments, using the same set of occupational prompts and evaluation metrics. We apply our FairPCA-based projection to remove age-correlated variance from prompt embeddings and use empirical noise injection to encourage diverse age representations in generated images.

As shown in Table 5, the results demonstrate that: (1) FairImagen and its variants (FairImagen-T5, FairImagen-OC) achieve the highest fairness scores among all compared methods, confirming the effectiveness of our approach in age debiasing; (2) FairImagen-OC achieves the best overall performance among our proposed variants, indicating strong compatibility with alternative text encoders; (3) all FairImagen variants maintain competitive accuracy and visual quality metrics with only modest degradation compared to the baseline, preserving the generation quality while improving fairness; and (4) the average scores show that FairImagen variants achieve the best overall balance between fairness and utility, validating the effectiveness of our post-hoc debiasing framework across different demographic dimensions beyond binary gender and race categories.

	Fairness	Accuracy	MUSIQ	Avg
Base	0.165	0.783	0.572	0.507
FairPrompt	0.21	0.764	0.582	0.519
ForcePrompt	0.202	0.765	0.579	0.515
SAL	0.193	0.785	0.595	0.524
CDA	0.195	0.775	0.564	0.511
TBIE	0.403	0.748	0.557	0.569
SDID	0.253	0.779	0.545	0.526
SDID-AVG	0.368	0.75	0.548	0.556
ITI	0.17	0.772	0.525	0.489
FairQueue	0.318	0.749	0.617	0.562
FairImagen	0.412	0.742	0.558	0.57
FairImagen-T5	0.45	0.738	0.564	0.584
FairImagen-OC	0.45	0.738	0.563	0.584

Table 5: Age debiasing results comparing FairImagen variants against baseline methods. Ages are categorized into young, middle-aged, and elderly groups. FairImagen variants achieve the highest fairness scores while maintaining competitive quality metrics.

	Fairness(%)	Accuracy(%)	MUSIQ(%)	Avg(%)
Base	0.61	0.05	0.39	0.24
FairPrompt	1.85	0.14	0.46	0.66
ForcePrompt	1.01	0.09	0.39	0.28
SAL	1.36	0.07	0.22	0.45
CDA	1.63	0.06	0.33	0.58
TBIE	1.63	0.08	0.29	0.55
SDID	2.17	0.11	0.32	0.78
SDID-AVG	4.3	0.5	0.46	1.71
ITI	0.76	0.11	0.36	0.34
FairQueue	1.21	0.15	0.61	0.26
FairImagen	2.37	0.13	0.28	0.74
FairImagen-T5	2.17	0.13	0.28	0.67
FairImagen-OC	2.27	0.13	0.28	0.72

Table 6: Standard deviation values across 10 random seeds for gender debiasing experiments. The results demonstrate the statistical stability and reliability of our experimental findings.

These results validate that FairImagen can be readily extended to additional protected attributes beyond the binary categories typically studied in fairness literature, opening opportunities for more comprehensive bias mitigation in text-to-image generation.

I Statistical Stability Analysis

To evaluate the statistical reliability of our experimental results, we conduct additional experiments by varying the random seed from 1 to 10 and computing the standard deviation for each evaluation metric. As shown in Table 6, the results demonstrate that: (1) the standard deviation values are relatively modest across all methods, indicating consistent and reliable experimental results; (2) FairImagen variants exhibit stable performance with standard deviations comparable to other baseline methods, confirming the robustness of our approach; (3) the performance differences between FairImagen and baseline methods are substantially larger than the natural variation introduced by random seed changes, validating the statistical significance of our improvements; and (4) the impact of our debiasing approach on image quality metrics remains within acceptable ranges, demonstrating that our method achieves meaningful fairness improvements while introducing quality changes comparable to the inherent stochasticity of the generation process itself.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main text is aligned with the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The experiments part discuss some of the limitations based on the results. Also, we discuss the limitations in Appendix C.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: No theory is is involved in this paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: See §4.5 and the code implementation.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code is included as a supplementary material. We will open-source it on Github after the anonymity period.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: See §4.5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: [NA]

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See §4.5

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: See Appendix A.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See Appendix A. Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate

to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Already cited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: A core aspect of our research is on large generative models.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.