Harnessing Diffusion-Generated Synthetic Images for Fair Image Classification

Abhipsa Basu[1], **Aviral Gupta**[2], **Abhijnya Bhat**[3], **Venkatesh Babu Radhakrishnan**[1]

[1]Indian Institute of Science, Bangalore
[2]BITS Pilani
[3]Stanford University

Abstract

Image classification systems often inherit biases from uneven group representation in training data. For example, in face datasets for hair color classification, blond hair may be disproportionately associated with females, reinforcing stereotypes. A recent approach leverages the Stable Diffusion model to generate balanced training data, but these models often struggle to preserve the original data distribution. In this work, we explore multiple diffusion-finetuning techniques, e.g., LoRA and DreamBooth, to generate images that more accurately represent each training group by learning directly from their samples. Additionally, in order to prevent a single DreamBooth model from being overwhelmed by excessive intra-group variations, we explore a technique of clustering images within each group and train a DreamBooth model per cluster. These models are then used to generate group-balanced data for pretraining, followed by fine-tuning on real data. Experiments on multiple benchmarks demonstrate that the studied finetuning approaches outperform vanilla Stable Diffusion on average and achieve results comparable to SOTA debiasing techniques like Group-DRO, while surpassing them as the dataset bias severity increases.

Code —
https://github.com/abhipsabasu/harnessing_diff_models
Extended version — https://arxiv.org/pdf/2511.08711

1 Introduction

Image classification models often exhibit harmful biases, posing significant risks for real-world deployment (Wang, Liu, and Wang 2021; Zhao et al. 2017; Metaxa et al. 2021). These biases arise from imbalances in training data; e.g., in CelebA (Celeba), blond female faces considerably outnumber blond males, leading to misclassification of the latter. While numerous debiasing techniques have been proposed (Sagawa et al. 2019; Kirichenko, Izmailov, and Wilson 2022; Nam et al. 2020), mitigating bias becomes increasingly tough when dataset imbalances become severe. With the recent breakthroughs in image generation using models like Stable Diffusion (Ramesh et al. 2022), we pose a critical question: *Can we harness the generative power of such models to create images that facilitate the training of fair classification systems, even in presence of extreme dataset bias?*

A recent work FFR (Qraitem, Saenko, and Plummer 2023) leverages images generated by Stable Diffusion (SD) to train fair classification systems. However, due to the stochastic nature of diffusion models (Shin, Kang, and Park 2023), SD-generated images often diverge from the original data distribution and may not follow prompt instructions accurately. Consider the Waterbirds dataset (Sagawa et al. 2019), which contains 'waterbird' and 'landbird' classes. Spurious correlations arise as classification models associate water backgrounds with waterbirds and land backgrounds with landbirds, thus relying on background rather than the bird features for predictions. Attempts to prompt SD to generate 'waterbird on land' images often produce water backgrounds, even when explicitly instructed otherwise. FFR attempts to mitigate this by using highly specific prompts, such as '`photo of a flamingo on pavement`', relying on exact bird names and background details. However, without precise domain knowledge, such prompts may yield irrelevant or out-of-distribution images. In this work, we explore fine-tuning generative models directly on the dataset, enabling them to better capture the data distribution and significantly improve classification fairness.

To generate images closely aligned with the training set, we fine-tune Stable Diffusion using LoRA (Hu et al. 2021a) on each dataset group (e.g., Non-Blond Females, Blond Males in CelebA (Celeba)). Prior work on synthetic data augmentation (Shin, Kang, and Park 2023) trains personalized diffusion models on individual classes to produce class-consistent, in-distribution images. Building on this, we further explore DreamBooth (Ruiz et al. 2022), which introduces a special token ('`[V]`') to represent the subject, improving resemblance to real samples. While such models typically target visually similar objects (e.g., a specific dog breed), our training groups contain diverse images sharing only a common attribute (e.g., hair color). To address this intra-group variation, we explore a simple extension—Clustered DreamBooth—which clusters each group into visually similar subsets and fine-tunes a separate DreamBooth model per cluster. These different generation methods are illustrated in Fig. 1. Using each strategy, we generate equal number of images per group.

After generating group-wise images, the key question becomes: *how should we use them for training?* FFR addresses this by first training a classification model on the synthetic

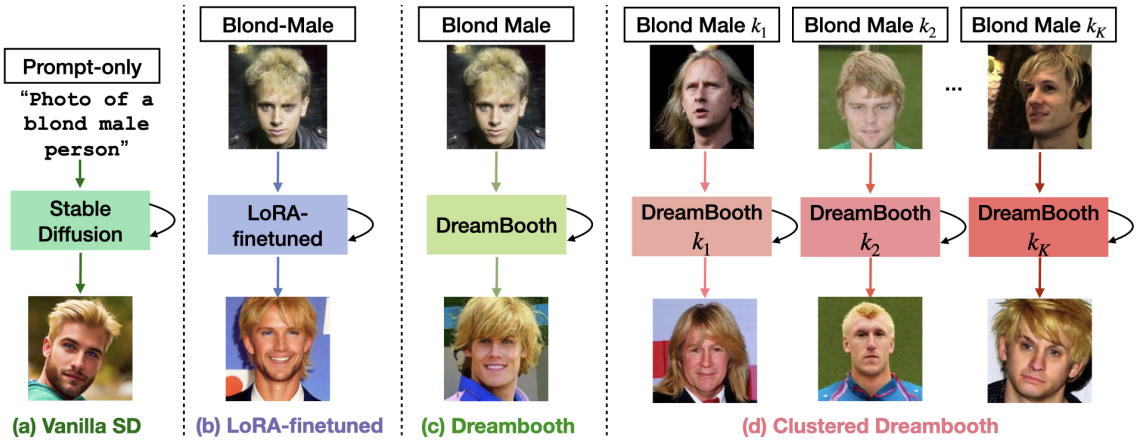

Figure 1: In this paper, we investigate four image generation pipelines for training fairer classifiers: a) *Vanilla SD*, which only accepts prompts, b) *LoRA-finetuning*, which finetunes the diffusion model on the images of a group, c) *Dreambooth*, which finetunes the diffusion model and represents the images of a group using a special token 'V', d) *Clustered Dreambooth*, which finetunes a diffusion model on different clusters k_i present in the training group, representing each of them through a special token 'V'.

data, followed by finetuning on the original dataset. However, since the original data is biased, this step risks reintroducing bias—unless extensive hyperparameter tuning is performed, as in FFR. To avoid this, we propose to finetune only the classifier's softmax layer with the real data (Basu, Mallick et al. 2024). Both the training and finetuning stages are optimized using a weighted combination of Cross-Entropy and Supervised Contrastive Loss (Khosla et al. 2020), which enhances feature representations by promoting better class separation.

We evaluate our proposed pipeline on three fairness benchmarks—Waterbirds (Sagawa et al. 2019), CelebA (Celeba), and UTKFace (Zhang, Song, and Qi 2017). The training on vanilla SD generated images followed by final layer retraining improves the accuracy of the classification systems over FFR indicating the utility of our finetuning approach. The different finetuning strategies discussed in the paper also offer added advantage for all the datasets, performing comparably to traditional debiasing techniques like Group-DRO (Sagawa et al. 2019) and SELF (LaBonte, Muthukumar, and Kumar 2023). Notably, as dataset biases intensify, such methods deteriorate, while our generation-based approaches remain robust, demonstrating their effectiveness in highly imbalanced settings. We summarize the key contributions below:

- We explore the utility of diffusion models and corresponding finetuning mechanisms like LoRA and Dreambooth, which learn to generate images from those in the biased training groups, to develop fair classifiers. We also explore Clustered Dreambooth, that first clusters group images and then trains distinct Dreambooth models on each cluster to better capture intra-group variations.
- We propose a two-stage pipeline for leveraging synthetic data for building fair classification systems. We generate equal number of images per group using our approaches and train a classification model on them. To further enhance performance, we finetune only the softmax layer of the model with the original data.
- Through extensive experiments on multiple benchmarks, we show that all diffusion-based methods enhance the training of fair classifiers more effectively than FFR. Notably, they surpass debiasing methods like Group-DRO by a large margin, when the dataset bias ratios become severe.

2 Related Work

Bias Mitigation has been widely studied, with approaches falling into two categories: *known* and *unknown biases*. For *known* biases, the spurious attribute is known apriori (Kim et al. 2019; Li and Vasconcelos 2019; Sagawa et al. 2019; Arjovsky et al. 2019; Teney, Abbasnejad, and van den Hengel 2021; Tartaglione, Barbano, and Grangetto 2021; Wang et al. 2020, 2022; Basu, Addepalli, and Babu 2023). In Group-DRO (Sagawa et al. 2019), the worst-group training loss is optimized. Last Layer Retraining (Kirichenko, Izmailov, and Wilson 2022) shows that pretraining the model on the biased dataset and then only retraining the final classification layer with a group-balanced validation set can help debias the model. Semi-supervised approaches assume bias annotations only for a few samples (Nam et al. 2020; Jung, Chun, and Moon 2022). For *unknown* biases, where bias attributes and labels are not available (Creager, Jacobsen, and Zemel 2021; Lee et al. 2021; Li, Hoogs, and Xu 2022; Lahoti et al. 2020; Ahn, Kim, and Yun 2023; Liu et al. 2021; Huang et al. 2020; Hong and Yang 2021; et al. 2020; Basu, Mallick et al. 2024), some approaches use dual-branch networks—one amplifying bias, the other mitigating it (Nam et al. 2020; Lee et al. 2021; Liu et al. 2022). Contrastive-based methods (Zhang et al. 2022; Zhang and Ré 2022) refine feature representations by clustering same-class samples, identifying pseudo bias labels via model misclassifications. Such biases are also exhibited by diffusion models (Basu, Babu, and Pruthi 2023; Parihar et al. 2024).

Data Augmentation using Generative Models. Many recent works utilize generative models for data augmentation (Trabucco et al. 2023; Azizi et al. 2023; Du et al. 2024; Zheng, Wu, and Li 2023; Mariani et al. 2018). The trend began with GANs (Goodfellow et al. 2020). BAGAN (Mariani et al. 2018) was used to augment class-imbalanced datasets to enhance minority class performance. With diffusion models, DA-Fusion (Trabucco et al. 2023) employs Textual Inversion (Gal et al. 2022) to generate augmentations for images of every class, and then during training, in each batch, retain every original image with probability p and an augmented image with probability $(1-p)$. DiffuseMix (Islam et al. 2024) combines a partial natural image and its generated counterpart from the diffusion model, and thereafter combats adversarial attacks by blending a randomly selected structural pattern from a set of fractal images into the concatenated image to form the final augmented version for training.

Generative Models for Debiasing utilize generative models to debias classification systems (An et al. 2022; Ramaswamy, Kim, and Russakovsky 2021; Qraitem, Saenko, and Plummer 2023; Sharmanska et al. 2020). GAN-based approaches (An et al. 2022; Ramaswamy, Kim, and Russakovsky 2021) train generative models on the training images to synthesize bias-conflicting samples that can augment the original data. The diffusion-based methods do not train the models from scratch, rather manipulate existing pretrained models to generate group-balanced images. FFR (Qraitem, Saenko, and Plummer 2023) likewise generates group-balanced images from Stable Diffusion, trains the classification model on this synthetic data, before finetuning the latter with real data. Our work additionally explores the use of different finetuning mechanisms in text-to-image models to generate in-distribution images directly from the training image groups.

3 Problem Statement and Methodology

Preliminaries. The motivation of this work is to train fairer image classification models using synthetic data. Let $\mathcal{X}$ be the set of real training images, where each $x_i \in \mathcal{X}$ is associated with a class label $y_i \in \mathcal{Y}$, a bias label $a_i \in \mathcal{A}$, and a group label $g_i \in \mathcal{G}$ where $g_i = (y_i, a_i)$. A mapping function $f : \mathcal{X} \rightarrow \mathcal{Y}$ is optimized by training a model in order to classify the images. Such a model has two parts: a) Feature encoder e, which is typically pretrained on a large dataset, and b) Classifier c which is finetuned along with encoder e to learn the class labels from the images. This model is traditionally trained using the Cross-Entropy (CE) loss. The model becomes biased when group frequencies in the training data are imbalanced, resulting in superior performance on some groups in the test data and inferior performance on others.

3.1 Generating Synthetic Images

With advancements in generative modeling (Song, Meng, and Ermon 2020; Ramesh et al. 2022), we explore their potential in training fair classifiers by generating images that reflect the training distribution and improve generalization to minority groups. Below, we outline our image generation methods and strategies for leveraging synthetic data to enhance fairness.

Vanilla Stable Diffusion (SD). We generate images from each group $g = (y, a)$ by specifying only y and a in the prompts. Such generations are independent of the training data, leading to domain mismatches, and may result in inaccurate generations if the model fails to follow the text prompts precisely (see § 1).

LoRA-based Finetuned Stable Diffusion. We make the generative model aware of the training data by finetuning the SD model on each training group g separately. Each model is trained on $l = \min\{|g| : g \in \mathcal{G}\}$ samples, selected randomly from each training group. The images are generated using the model trained on g by specifying y and a in the prompt.

Dreambooth. To strengthen the resemblance between the training and generated images, we explore Dreambooth (Ruiz et al. 2022), a text-to-image personalization model that learns to imitate an object or a concept (e.g., a specific dog) from a small set of images depicting that object. It finetunes a pretrained text-to-image model by learning a unique identifier (e.g., "`[V]`") such that on inference time, if the model is queried by that identifier (e.g., "`photo of a [V] dog`"), it generates new images of the given object. Likewise, we sample 100 images from each training group, and train a separate Dreambooth model h on each group, where the prompt is of the form "`photo of a [V] y`".

Clustered Dreambooth. Dreambooth specializes in learning a concept from $3-5$ images. However, a training group like Blond Male consists of images of many individuals sharing a common trait, hair color. To prevent overwhelming a single Dreambooth model with multiple images of a training group, we explore a simple extension of first clustering the CLIP embeddings (Radford et al. 2021) of the images within each group. Let k_D^g denote the number of clusters, where D and g refer to the training dataset and a group in D respectively. We train a pool of Dreambooth models $\mathcal{H}^g = \{h_1^g, h_2^g, h_3^g, \cdots, h_{k_D^g}^g\}$ on the obtained clusters. We implement Clustered-Dreambooth (i.e., the Dreambooth pool $\mathcal{H}^g$) using LoRA-based finetuning (Hu et al. 2021b), which ensures lesser, feasible training time. Finally the trained models are utilized to generate images for each g. For simplicity, we assume equal k_D^g for each group g, and denote the number of clusters as k_D for the rest of the paper.

3.2 Stage 1: Training with the Generated Images

Once the generative models are trained with the individual data groups, we generate M images from each group g using Vanilla SD, LoRA-finetuned SD and Dreambooth. For Clustered Dreambooth, we generate M_D^{cl} images from each cluster in a group belonging to dataset D, such that the total number of images generated from the group is $M_D^{cl} \times k_D = M$. However, inspite of training models on the dataset groups, all the generated images may not always follow the prompt or may not be of high quality. Thus a filtering step is required for selecting images appropriate for our training. This is done using a CLIP-based scoring mechanism.

CLIP-based Filtering. To find the most relevant images, we apply a CLIP score in two ways for each image I.

1. **CLIP-Label**$(\mathbf{I}, \mathbf{p^c})$: We compute the image-text similarity of I with a prompt p^c, of the format "`Photo of a {c}`", where c is the class label.
2. **CLIP-Centroid**$(\mathbf{I}, \overline{\mathbf{z}}^{\mathbf{g}})$: To ensure that the chosen images

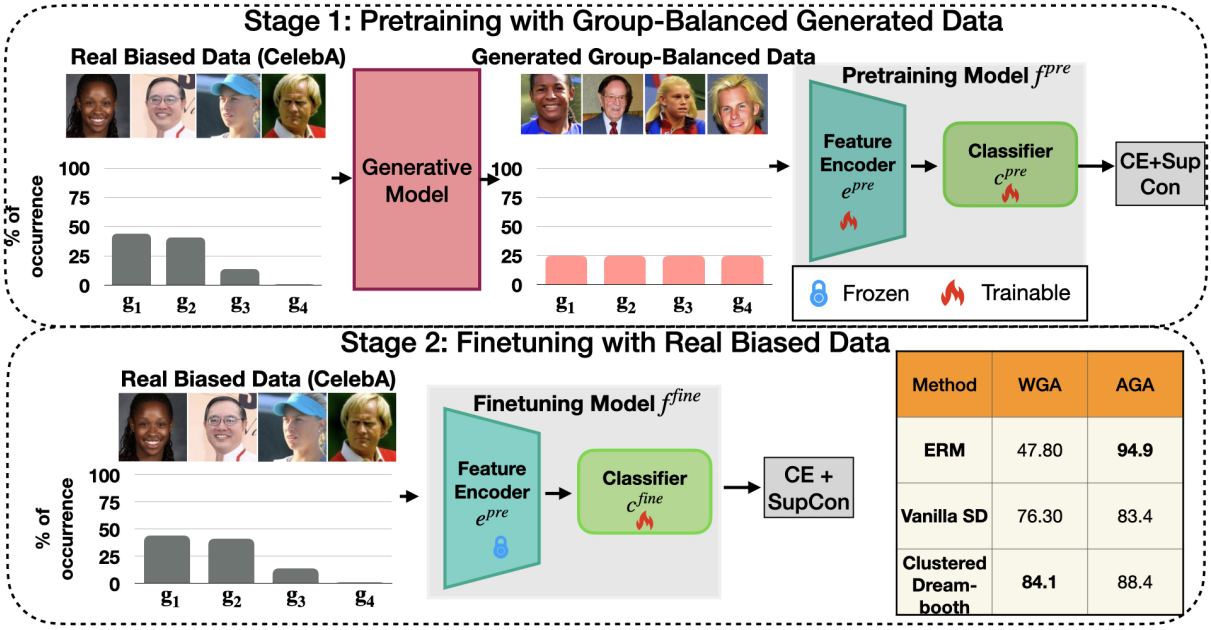

Figure 2: We describe the overview of the studied pipeline. In Stage 1, we generate images uniformly from each group (e.g., non-blond female (g_1), non-blond male (g_2), blond female (g_3), blond male (g_4)) using the generatice approaches, and train a classification model f^{pre} with CE and SupCon losses. In Stage 2, we finetune only the linear classifier on the original dataset.

most resemble the distribution in the given training group g, we compute the centroid of the CLIP embeddings of the images of the group, denoted by $\bar{z}^g = \frac{1}{M_g}\sum_{i=1}^{M_g} z_i^g$, where M_g is the size of the group in the training set, and z_i^g is the CLIP embedding of the i^{th} image of g. We calculate the CLIP similarity between each generated image and its corresponding group centroid.

The final scoring function becomes a combination of CLIP-Label(I, p^c) and CLIP-Centroid$(I, \bar{z}^g)$:

$$\begin{aligned}\text{CLIP-Score}(I, p^c, \bar{z}^g) = \alpha \cdot \text{CLIP-Label}(I, p^c) \\ + (1-\alpha) \cdot \text{CLIP-Centroid}(I, \bar{z}^g)\end{aligned} \tag{1}$$

where α is a hyperparameter. After selecting the top-ranked images from each group, a classification model f^{pre} is trained on them to learn fair representations. We empirically validate the effectiveness of the filtering step and find it essential for removing irrelevant images.

3.3 Stage 2: Finetuning with Original Data

After pretraining the classification model on group-balanced synthetic data, we adapt it to the real data by finetuning it with the real images using the method described below.

Last Layer Retraining with Real Data. We finetune the trained model f^{pre} on the entire real training dataset to help it adapt to real data. However, since the data is biased, finetuning the full network risks reintroducing the biases. To mitigate this, we freeze the feature encoder e^{pre} and only finetune the linear classification layer c^{pre}. Additionally, to address any class imbalance, each finetuning batch samples classes uniformly. We refer to this method as LLR_{all}, and the finetuned model as f^{fine}. We differ here from FFR, where the entire model is finetuned, increasing the dependency on thorough hyperparameter tuning. Our two-stage approach is illustrated in Figure 2 based on CelebA (Celeba).

We train both stages using a weighted sum of CE loss and Supervised Contrastive (SupCon) loss (Khosla et al. 2020) to enhance class separation in both stages: $L = \beta.L_{\text{CE}} + (1-\beta)L_{\text{sup-con}}$, where $\beta = 0.5$ in all experiments.

4 Experiments and Results

We present an overview of the datasets used for evaluation, followed by a comparative analysis of the diffusion model variants to determine their performances.

Datasets For Evaluation: Waterbirds (Sagawa et al. 2019) is a dataset of bird images, labeled as waterbird if the bird is commonly seen with waterbodies on background, and landbird otherwise. The dataset, with $4,795$ samples, suffers from background bias: only a few waterbirds in the dataset have land on the background, and few landbirds have water. We also consider two real-world datasets. The CelebA dataset (Celeba) consists of $202,599$ images of celebrities with annotations of 40 binary attributes. We choose Blond Hair as the target attribute, which is known to suffer from gender biases (Li, Hoogs, and Xu 2022; Seo, Lee, and Han 2022). UTKFace (Zhang, Song, and Qi 2017) is a dataset of human faces, having $20,000$ images with annotations of age, gender, and ethnicity. We use gender and age as the target and bias attributes respectively – where female adults dominate female children, and male children dominate male adults. For

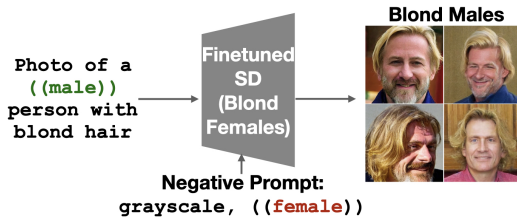

Figure 3: This is the Image Generation Pipeline for Bias Ratio= $\mathbf{0.999}$. Bias-conflicting samples (e.g., Blond Males in CelebA) are generated using models trained on bias-aligned images (e.g., Blond Females).

each dataset, we report the worst (WGA) and average group (AGA) accuracies, following previous works (Sagawa et al. 2019; LaBonte, Muthukumar, and Kumar 2023).

Varying Bias Severity. We extend our evaluation to scenarios where each dataset is severely biased, setting the bias ratio to 0.999 (i.e., 99.9% of training images belong to bias-aligned groups). The diffusion models must generate images for minority groups to counteract bias severity. For Vanilla SD and FFR, images are generated by prompting with the bias label a and class label y. As LoRA-finetuned SD, DreamBooth, and Clustered DreamBooth rely on the training group images, the bias-conflicting images are generated using models trained on bias-aligned groups (e.g., Blond Males are generated from the model trained on Blond Females). Interestingly, for the Dreambooth models, we find that during the bias-conflicting sample generation, removing the learnt '[V]' token from the prompt leads to more accurate depiction of the target group descriptions. On manual inspection, we find that this way of transferring the style of one group into another (described in Fig. 3) leads to images that visually follow the distribution of the input data, while imitating the target group. Generated images are filtered using the CLIP-Label score with $\alpha = 1$ (see eq. 1), as the minority groups lack sufficient samples for the CLIP-Centroid computation. The classifier is then pretrained on group-balanced synthetic images and finetuned on the severely biased dataset for each method.

Implementation Details: We use SD v1.4 for all experiments. For each dataset group, we generate $M = 5000$ images for each method, and score the images using the CLIP-Score defined in eq. 1. For simplicity, we set the weighting parameter α to be 0.5 for the original dataset ($\alpha = 1$ for the severely biased version). Accordingly, we select the top 75% images from each group. An ImageNet (Deng et al. 2009)-pretrained ResNet-50 (He et al. 2016) model is trained on the group-balanced data with CE and SupCon losses. The model is trained for 20 epochs using the SGD optimizer during both the training and finetuning stages. As the validation sets of the datasets are not group-balanced, we refrain from experimenting with hyperparameters, and uniformly set a learning rate of $1e-3$, weight decay of $1e-3$ and batch size of 128. We discuss the choice of clusters for the Clustered Dreambooth method in § 4.2.

4.1 Quantitative Results

Our goal is to obtain rich representations from synthetic images to enable classification models to be fair, even when finetuned on biased real data. Table 1 presents the worst and average group accuracies (WGA and AGA) across all test groups and dataset variants. To further motivate the importance of the generative models in training fair classifiers, we compare their performance with traditional debiasing methods like Group-DRO (GDRO) (Sagawa et al. 2019) and SELF (LaBonte, Muthukumar, and Kumar 2023). Key findings are summarized below:

Performance on the Original Dataset. Our classification pipeline, leveraging images from vanilla SD, outperforms FFR (Qraitem, Saenko, and Plummer 2023) by an average of 2.4%. This improvement can be attributed to our last layer retraining-based finetuning stage. DreamBooth achieves the highest worst-group accuracy for Waterbirds (89.3%), followed by Clustered DreamBooth (88.1%) and LoRA finetuning (86.5%). For CelebA and UTKFace, Clustered DreamBooth outperforms all other generative methods and is the only one to surpass ERM scores for UTKFace, improving upon LoRA by 7.4%. This highlights the advantage of training multiple DreamBooth models on clustered subsets, particularly for facial datasets. While Clustered DreamBooth has higher time complexity, k_D can be adjusted based on resource constraints, with $k_D = 1$ (a single DreamBooth model per group) as a special case. Notably, since DreamBooth models are themselves trained with LoRA, per-model training remains computationally efficient. Overall, fine-tuning consistently outperforms vanilla Stable Diffusion across all datasets, achieving accuracy comparable to traditional debiasing methods like Group-DRO.

Performance for Bias Ratio $= \mathbf{0.999}$. With severe bias ratio, traditional debiasing methods experience a drastic performance drop, whereas generative methods exhibit significantly lower degradation. For Waterbirds and CelebA, DreamBooth and Clustered DreamBooth outperform all other methods, with performance gaps between the original and biased versions remaining below 7%, compared to over 60% for traditional debiasing approaches. This emphasizes the importance of leveraging generative models for training fairer classifiers. Additionally, for these datasets, the accuracies of vanilla SD and LoRA finetuning remain similar. For UTKFace, all generative finetuning methods underperform compared to the vanilla SD pipeline, which achieves the highest accuracy. Manual inspection reveals many bias-conflicting samples are irrelevant or out-of-domain. We leave further investigation and mitigation of this problem for future work. Averaged across all datasets, fine-tuning approaches outperform vanilla Stable Diffusion, with Clustered Dreambooth achieving 75.5% WGA.

Generative-Based Pretraining + GDRO Finetuning ($\mathbf{0.999}$ bias ratio). Table 1 shows that the WGA of GDRO (Sagawa et al. 2019) suffers for high bias ratio. We next analyse the impact of finetuning the classification model pretrained by the Clustered Dreambooth images [1] for Water-

[1]We choose Clustered Dreambooth for different ablations and experiments only as an example.

Dataset	Method	Synthetic Data?	Original Dataset		Bias Ratio 0.999		Average Performance	
			Worst	Average	Worst	Average	Worst	Average
Waterbirds	ERM	✗	63.7	88.0	29.0	66.7	46.3	77.3
	FFR† (Qraitem, Saenko, and Plummer 2023)	✓	69.5	84.0	57.3	84.2	63.4	84.1
	Vanilla SD	✓	$74.6^{\pm 2.90}$	$80.5^{\pm 0.34}$	$69.9^{\pm 0.70}$	$80.1^{\pm 0.13}$	72.2	80.3
	LoRA-finetuning	✓	$86.5^{\pm 3.81}$	$89.9^{\pm 0.76}$	$61.5^{\pm 0.40}$	$84.0^{\pm 0.11}$	74.0	87.0
	Dreambooth (Ruiz et al. 2022)	✓	$\mathbf{89.3}^{\pm 0.75}$	$\underline{90.1}^{\pm 0.50}$	$82.4^{\pm 0.25}$	$\underline{88.3}^{\pm 0.22}$	$\underline{85.9}$	$\underline{89.2}$
	Clustered Dreambooth	✓	$88.1^{\pm 0.92}$	$\mathbf{90.2}^{\pm 0.11}$	$\mathbf{84.2}^{\pm 0.46}$	$\mathbf{88.5}^{\pm 0.14}$	**86.0**	**89.3**
	GDRO† (Sagawa et al. 2019)	✗	91.4	93.5	23.5	65.5	57.4	79.5
	SELF† (LaBonte, Muthukumar, and Kumar 2023)	✗	93.0	94.0	25.5	64.2	59.2	79.1
CelebA	ERM	✗	47.8	94.9	31.7	67.3	39.7	81.1
	FFR† (Qraitem, Saenko, and Plummer 2023)	✓	68.9	85.7	22.8	47.7	45.9	66.7
	Vanilla SD	✓	$76.4^{\pm 1.27}$	$84.2^{\pm 0.37}$	$77.1^{\pm 0.42}$	$\underline{84.7}^{\pm 0.67}$	76.7	84.4
	LoRA-finetuning	✓	$\underline{82.3}^{\pm 1.51}$	$87.2^{\pm 0.56}$	$73.5^{\pm 2.83}$	$83.2^{\pm 0.29}$	77.9	85.2
	Dreambooth (Ruiz et al. 2022)	✓	$82.1^{\pm 0.00}$	$\underline{87.9}^{\pm 0.32}$	$\underline{78.8}^{\pm 0.21}$	$84.6^{\pm 0.19}$	$\underline{80.4}$	$\underline{86.2}$
	Clustered Dreambooth	✓	$\mathbf{84.1}^{\pm 0.63}$	$\mathbf{88.4}^{\pm 0.19}$	$\mathbf{81.8}^{\pm 0.35}$	$\mathbf{85.9}^{\pm 0.28}$	**82.9**	**87.1**
	GDRO† (Sagawa et al. 2019)	✗	88.9	92.9	27.2	75.2	58.0	84.0
	SELF† (LaBonte, Muthukumar, and Kumar 2023)	✗	83.9	91.1	45.6	95.4	64.7	93.2
UTKFace	ERM	✗	74.3	84.5	31.0	48.9	52.6	66.7
	FFR† (Qraitem, Saenko, and Plummer 2023)	✓	67.4	81.4	55.0	68.0	61.2	74.7
	Vanilla SD	✓	$62.0^{\pm 3.89}$	$83.3^{\pm 0.92}$	$\mathbf{67.8}^{\pm 1.27}$	$\mathbf{82.7}^{\pm 0.61}$	64.9	$\underline{83.0}$
	LoRA-finetuning	✓	$68.6^{\pm 3.91}$	$\mathbf{85.6}^{\pm 0.76}$	$64.5^{\pm 2.59}$	$\underline{82.4}^{\pm 0.26}$	$\underline{66.5}$	**84.0**
	Dreambooth (Ruiz et al. 2022)	✓	$57.9^{\pm 3.91}$	$80.9^{\pm 0.76}$	$\underline{66.9}^{\pm 2.59}$	$77.1^{\pm 0.26}$	62.4	79.0
	Clustered Dreambooth	✓	$\mathbf{76.0}^{\pm 1.22}$	$\underline{83.5}^{\pm 0.35}$	$60.5^{\pm 1.22}$	$80.8^{\pm 0.35}$	**68.2**	82.1
	GDRO† (Sagawa et al. 2019)	✗	81.6	85.9	30.5	50.3	56.0	68.1
	SELF† (LaBonte, Muthukumar, and Kumar 2023)	✗	65.9	82.3	0.6	50.5	33.3	66.4

Table 1: Classification Performance for the Original Dataset and high-bias variant with respect to each finetuning variant on three datasets. While each finetuning method achieves worst group accuracy (WGA) comparable to debiasing methods like GDRO and SELF for the original datasets, they far outperform the latter for the high bias-ratio case. Clustered Dreambooth achieves highest WGA across all datasets (79.1%). Our results are averaged across 3 random seeds. † indicates implementation using the codebases of existing methods. The best and 2nd best scores are marked in bold and underline respectively.

Method	GDRO		Clustered Dreambooth (Pretraining)+ GDRO	
	WGA	AGA	WGA	AGA
Waterbirds	23.5	65.5	81.8	89.8
CelebA	27.2	75.2	64.4	85.0

Table 2: Clustered Dreambooth + GDRO for Waterbirds and CelebA: i.e., pretrain the classification model using the synthetic images, and then finetune the same using GDRO.

birds and CelebA, using the GDRO loss, leading to a 58.3% and 37.2% performance increase for Waterbirds and CelebA respectively (Table 2). This analysis further highlights the positive effect of group-balanced synthetic image pretraining.

Time Complexity of Finetuning. We analyze the time complexity vs performance tradeoff for the finetuning techniques, averaged across all datasets and bias ratios. Clustered DreamBooth outperforms others, but it has higher time complexity of $\mathcal{O}(|\mathcal{G}_D|.k_D)$, where $|\mathcal{G}_D|$ represents the number of groups in the training set, and k_D : the number of clusters per group. In contrast, vanilla SD, which requires no finetuning, yields the lowest score across all datasets. These tradeoffs are presented in Table 3, providing insights to help practitioners select the most suitable method based on their application needs.

4.2 Design Choices

Choice of Clusters for Clustered Dreambooth. We select the number of clusters k_D for Clustered DreamBooth based on the size of the smallest group in the dataset, M_{g_s}. To ensure sufficient training data per cluster, we set $k_D \approx \frac{M_{g_s}}{20}$, keeping at least 20 samples per cluster. To limit complexity on larger datasets, we cap the number of clusters: $k_D = \min\left(\frac{M_{g_s}}{20}, 20\right)$. Thus, we use $k_D = 3$ for Waterbirds ($M_{g_s} = 56$), $k_D = 20$ for CelebA ($M_{g_s} = 1387$), and $k_D = 5$ for UTKFace ($M_{g_s} = 103$). For simplicity, k_D is fixed per dataset and shared across all groups, assuming similar intra-group variation. Further optimization of k_D is left for future work.

Group-Balanced Finetuning. After pretraining on generated data, we finetune the classification layer on real data (LLR_{all}). To assess the benefits of group-balancing the real data (sized to the smallest group), we explore: a) *Last Layer Retraining with Balanced Real Data* (LLR_{b}): Finetuning only the classification layer using the balanced real dataset instead

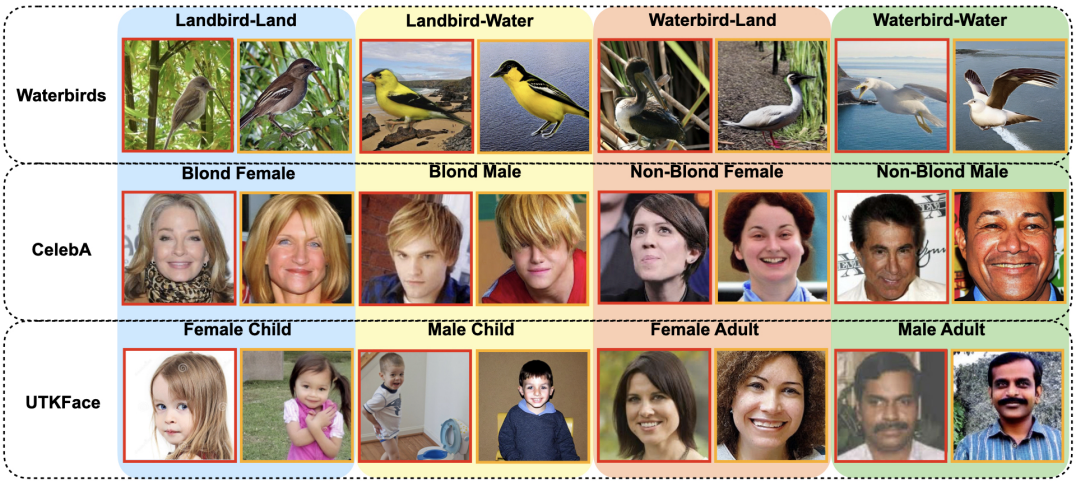

Figure 4: Training vs Generated Images from Clustered Dreambooth: In this figure, we present original images from each group in the studied datasets (images with red border) along with the generated ones (yellow border). We note how the generated images closely reflect the training distribution and the group-related attributes.

Figure 5: Dreambooth vs Clustered Dreambooth for UTKFace Female Children. Clustered Dreambooth samples are more diverse than the Dreambooth ones.

Method	Time Complexity	WGA	AGA
Vanilla SD	$\mathcal{O}(1)$	71.3	82.6
LoRA-finetuning	$\mathcal{O}(\lvert\mathcal{G}_D\rvert)$	72.8	85.4
Dreambooth	$\mathcal{O}(\lvert\mathcal{G}_D\rvert)$	76.2	84.8
Clustered Dreambooth	$\mathcal{O}(\lvert\mathcal{G}_D\rvert.k_D)$	**79.1**	**86.2**

Table 3: Classification Performance vs Time Complexity Tradeoff, averaged across all datasets and both bias ratios.

of the full biased set. b) *Full Fine-Tuning with Balanced Real Data* (FT_b): Finetuning the entire network with the balanced real dataset. Emperical results show that FT_b is beneficial only for CelebA, likely due to its larger size compared to Waterbirds and UTKFace. In contrast, LLR_b reduces worst group accuracy across all datasets, indicating that finetuning on the entire dataset yields better performance.

4.3 Qualitative Analysis

Clustered Dreambooth vs Original. Figure 4 shows original images alongside synthetic ones for each data group, showing strong alignment with group characteristics and domain preservation, leading to high WGA even during pretraining.
Dreambooth for UTKFace. Table 1 shows that the performance of models trained on vanilla Dreambooth drops in case of UTKFace (original). To investigate, we present *female children* faces in Figure 5. The Dreambooth images follow a narrow demographic and age range, whereas the Clustered Dreambooth images are more diverse in demographics, age groups, and backgrounds. This shows Clustered Dreambooth's strength in capturing variety within training groups, unlike a single model trained on all group images.

5 Conclusion

In this work, we investigated Stable Diffusion and various finetuning approaches, including Dreambooth and LoRA, to enhance fairness in image classification by generating representative images for each training group. We also explored Clustered Dreambooth, which addresses intra-group diversity by clustering images within each group and training separate Dreambooth models per cluster, preventing a single model from being overwhelmed by excessive variation. Using these approaches, we generated group-balanced images, pretrained a classifier on them, and finetuned it on real data. Experiments on three fairness benchmarks demonstrated that diffusion-based finetuning, particularly Clustered Dreambooth, consistently outperforms vanilla SD and FFR (Qraitem, Saenko, and Plummer 2023), achieving comparable or superior results to SOTA debiasing methods like GroupDRO—especially as dataset biases become more severe. We refer the readers to the extended version for more ablations, details on the loss functions used, qualitative examples and failure cases.

Limitations. Our work has some limitations. Unlike vanilla SD, finetuned variants require dedicated training, with image quality sensitive to hyperparameters. For Clustered Dreambooth, optimizing the cluster count k_D through further experiments could improve performance. Additionally, generated images alone cannot fully mitigate biases, necessitating a two-stage approach. Despite these challenges, our study highlights the potential of diffusion models to improve fairness in classification tasks.

Acknowledgments

The work of Abhipsa Basu is partially supported by the Qualcomm Innovation Fellowship and the Ministry of Education Fellowship of the Government of India. We thank Soumya Dutta, PhD Student, LEAP Lab, Indian Institute of Science Bangalore, for their invaluable feedback, as well as the anonymous reviewers for their constructive suggestions.

References

Ahn, S.; Kim, S.; and Yun, S.-Y. 2023. Mitigating Dataset Bias by Using Per-Sample Gradient. In *The Eleventh International Conference on Learning Representations*.

An, J.; Kim, T.; Ko, D.; Lee, S.; and Woo, S. S. 2022. Aˆ 2: Adaptive Augmentation for Effectively Mitigating Dataset Bias. In *Proceedings of the Asian Conference on Computer Vision*, 4077–4092.

Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Azizi, S.; Kornblith, S.; Saharia, C.; Norouzi, M.; and Fleet, D. J. 2023. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*.

Basu, A.; Addepalli, S.; and Babu, R. V. 2023. Rmlvqa: A margin loss approach for visual question answering with language biases. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11671–11680.

Basu, A.; Babu, R. V.; and Pruthi, D. 2023. Inspecting the geographical representativeness of images from text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5136–5147.

Basu, A.; Mallick, S. S.; et al. 2024. Mitigating biases in blackbox feature extractors for image classification tasks. *Advances in Neural Information Processing Systems*, 37: 106411–106439.

Celeba. 2015. CelebA dataset. http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html.

Creager, E.; Jacobsen, J.-H.; and Zemel, R. 2021. Environment inference for invariant learning. In *International Conference on Machine Learning*, 2189–2200. PMLR.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Du, X.; Sun, Y.; Zhu, J.; and Li, Y. 2024. Dream the impossible: Outlier imagination with diffusion models. *Advances in Neural Information Processing Systems*, 36.

et al., D. 2020. Latent adversarial debiasing: Mitigating collider bias in deep neural networks. *arXiv preprint arXiv:2011.11486*.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hong, Y.; and Yang, E. 2021. Unbiased classification through bias-contrastive and bias-balanced learning. *Advances in Neural Information Processing Systems*, 34: 26449–26461.

Hu, E.; et al. 2021a. LoRA: Efficient Fine-Tuning of Large Models. Hugging Face Blog. Accessed: 2025-03-06.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021b. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Huang, Z.; Wang, H.; Xing, E. P.; and Huang, D. 2020. Self-challenging improves cross-domain generalization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, 124–140. Springer.

Islam, K.; Zaheer, M. Z.; Mahmood, A.; and Nandakumar, K. 2024. DiffuseMix: Label-Preserving Data Augmentation with Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 27621–27630.

Jung, S.; Chun, S.; and Moon, T. 2022. Learning fair classifiers with partially annotated group labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10348–10357.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33: 18661–18673.

Kim, B.; Kim, H.; Kim, K.; Kim, S.; and Kim, J. 2019. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9012–9020.

Kirichenko, P.; Izmailov, P.; and Wilson, A. G. 2022. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*.

LaBonte, T.; Muthukumar, V.; and Kumar, A. 2023. Towards last-layer retraining for group robustness with fewer annotations. *arXiv preprint arXiv:2309.08534*.

Lahoti, P.; Beutel, A.; Chen, J.; Lee, K.; Prost, F.; Thain, N.; Wang, X.; and Chi, E. 2020. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33: 728–740.

Lee, J.; Kim, E.; Lee, J.; Lee, J.; and Choo, J. 2021. Learning debiased representation via disentangled feature augmentation. *Advances in Neural Information Processing Systems*, 34: 25123–25133.

Li, Y.; and Vasconcelos, N. 2019. Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9572–9581.

Li, Z.; Hoogs, A.; and Xu, C. 2022. Discover and Mitigate Unknown Biases with Debiasing Alternate Networks. In *The European Conference on Computer Vision (ECCV)*.

Liu, E. Z.; Haghgoo, B.; Chen, A. S.; Raghunathan, A.; Koh, P. W.; Sagawa, S.; Liang, P.; and Finn, C. 2021. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, 6781–6792. PMLR.

Liu, S.; Zhang, X.; Sekhar, N.; Wu, Y.; Singhal, P.; and Fernandez-Granda, C. 2022. Avoiding spurious correlations via logit correction. *arXiv preprint arXiv:2212.01433*.

Mariani, G.; Scheidegger, F.; Istrate, R.; Bekas, C.; and Malossi, C. 2018. Bagan: Data augmentation with balancing gan. *arXiv preprint arXiv:1803.09655*.

Metaxa, D.; Gan, M. A.; Goh, S.; Hancock, J.; and Landay, J. A. 2021. An image of society: Gender and racial representation and impact in image search results for occupations. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–23.

Nam, J.; Cha, H.; Ahn, S.; Lee, J.; and Shin, J. 2020. Learning from Failure: Training Debiased Classifier from Biased Classifier. In *Advances in Neural Information Processing Systems*.

Parihar, R.; Bhat, A.; Basu, A.; Mallick, S.; Kundu, J. N.; and Babu, R. V. 2024. Balancing act: distribution-guided debiasing in diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6668–6678.

Qraitem, M.; Saenko, K.; and Plummer, B. A. 2023. From Fake to Real (FFR): A two-stage training pipeline for mitigating spurious correlations with synthetic data. *arXiv preprint arXiv:2308.04553*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Ramaswamy, V. V.; Kim, S. S.; and Russakovsky, O. 2021. Fair attribute classification through latent space de-biasing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9301–9310.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2022. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*.

Sagawa, S.; Koh, P. W.; Hashimoto, T. B.; and Liang, P. 2019. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*.

Seo, S.; Lee, J.-Y.; and Han, B. 2022. Unsupervised Learning of Debiased Representations with Pseudo-Attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16742–16751.

Sharmanska, V.; Hendricks, L. A.; Darrell, T.; and Quadrianto, N. 2020. Contrastive examples for addressing the tyranny of the majority. *arXiv preprint arXiv:2004.06524*.

Shin, J.; Kang, M.; and Park, J. 2023. Fill-up: Balancing long-tailed data with generative models. *arXiv preprint arXiv:2306.07200*.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Tartaglione, E.; Barbano, C. A.; and Grangetto, M. 2021. End: Entangling and disentangling deep representations for bias correction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13508–13517.

Teney, D.; Abbasnejad, E.; and van den Hengel, A. 2021. Unshuffling data for improved generalization in visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1417–1427.

Trabucco, B.; Doherty, K.; Gurinas, M.; and Salakhutdinov, R. 2023. Effective data augmentation with diffusion models. *arXiv preprint arXiv:2302.07944*.

Wang, J.; Liu, Y.; and Wang, X. E. 2021. Are gender-neutral queries really gender-neutral? mitigating gender bias in image search. *arXiv preprint arXiv:2109.05433*.

Wang, Z.; Dong, X.; Xue, H.; Zhang, Z.; Chiu, W.; Wei, T.; and Ren, K. 2022. Fairness-Aware Adversarial Perturbation Towards Bias Mitigation for Deployed Deep Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10379–10388.

Wang, Z.; Qinami, K.; Karakozis, I. C.; Genova, K.; Nair, P.; Hata, K.; and Russakovsky, O. 2020. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8919–8928.

Zhang, M.; and Ré, C. 2022. Contrastive adapters for foundation model group robustness. *Advances in Neural Information Processing Systems*, 35: 21682–21697.

Zhang, M.; Sohoni, N. S.; Zhang, H. R.; Finn, C.; and Ré, C. 2022. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. *arXiv preprint arXiv:2203.01517*.

Zhang, Z.; Song, Y.; and Qi, H. 2017. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5810–5818.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.

Zheng, C.; Wu, G.; and Li, C. 2023. Toward understanding generative data augmentation. *Advances in neural information processing systems*, 36: 54046–54060.