Balancing Act: Distribution-Guided Debiasing in Diffusion Models

Contents

A. Ablation Experiments

A.1. Guidance Strength γ

We ablate over guidance strength parameter γ for both the distribution guidance and sample guidance in Fig. 1. Increasing γ reduces the bias (better FD) at the cost of inferior image quality (increased FID). Guidance strength 1500 achieved a good tradeoff between image quality and balancing attributes. Further, as compared to sample guidance, the distribution guidance achieves a better tradeoff for all the guidance strengths.

A.2. Data efficiency of h-space classifiers

We ablate over the number of training examples used to train the *h-space* classifiers in Fig. 2. Specifically, we train a *gender*-attribute classifier in image space and h-space using the same number of images. We use ResNet50 [1] as the image classifier and a linear head on top of the *h-features* as the *h-space* classifier. We pass the estimate $\hat{x}_0$ of the clean image at timestep t to the image classifier to obtain the prediction. For *h-space* classification, we pass the h_t to the linear head. The *h-space* classifiers are *extremely data efficient* and can achieve $> 90\%$ accuracy for most of the time-steps, even when trained with just 500 training examples. This shows the efficacy of guidance in the h-space,

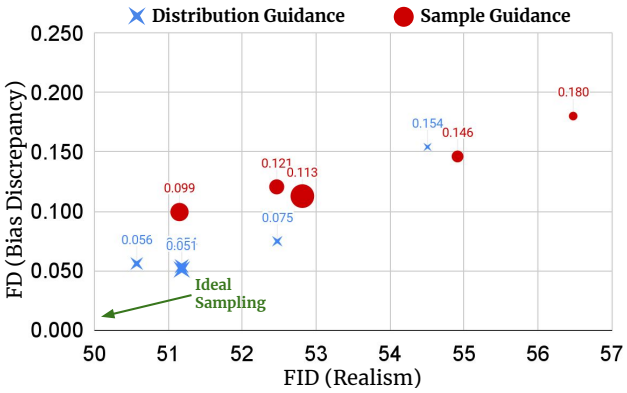

Figure 1. Ablation over guidance strength γ

which can be done with only a few hundred training examples.

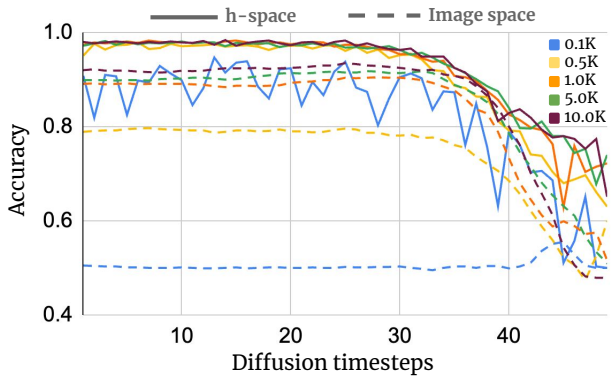

Figure 2. Ablation over the number of training examples for classifiers. h-space classifiers achieve excellent performance even with only 200 training examples. In contrast, ResNet50 classifiers require a large number of images to achieve similar performance.

A.3. H-space classifier architectures

We ablate over the different classifier architectures used for the classification in the h-space. We use a linear layer, an MLP and a small CNN architecture. The results are presented Tab. 1. We report the average test accuracy over 50 time steps, network parameters, and guidance time for generating a batch of images. With comparable accuracy, the

Table 1. Ablation over the classifier architecture for h-space classification

Classifier	Avg. accuracy	Guidance time ↓	# Parameters
Linear	0.921	1.0x	3.2M
CNN	0.908	2.5x	74M
MLP	0.952	2.0x	200M

linear layer has lesser parameters and performs guidance faster, and hence is used as the default classifier.

B. Multi-attribute Debiasing

In this section, we generate balanced subgroups for a combination of the multiple attributes. We consider jointly balancing 2 and 3 sensitive attributes across all subgroups. Specifically, we explore the following settings: **1) Gender + Race**: 0.25 *black males*, 0.25 *black females*, 0.25 *white males*, 0.25 *white females* **2) Gender + Eyeglasses**: 0.25 *males with eyeglasses*, 0.25 *females with eyeglasses*, 0.25 *males without eyeglasses* and 0.25 *females without eyeglasses*. **3) Gender + Race + Eyeglasses**: 0.125 *for all the* 8 *subgroups formed*. The results are reported in Tab. 2 where we compute the FD score with a balanced reference set and FID with the original CelebA-HQ [2] dataset. The FID score quantifies the visual quality, whereas the FD score accounts for the bias in the generations, as explained in Sec.4.1 (main paper). We do not use a reference set to compute FID as balancing across sub-groups leads to considerably less number of samples in the reference set.

Table 2. Balancing attribute subgroups for fair generation

	Gender + Race		Gender + Eyeglasses		Gender + Race + Eyeglasses	
Method	FD	FID	FD	FID	FD	FID
Random Generation	0.684	49.45	0.636	49.45	0.768	49.45
Sample Guidance	0.436	45.49	0.3	47.41	0.496	47.83
Distribution Guidance	**0.224**	45.37	**0.2**	45.92	**0.408**	43.94

B.1. Generalization to imbalanced distributions

Here, we present results for generating an imbalanced distribution for the subgroups by providing a skewed reference distribution. Such a setting is helpful in data augmentation for under-represented subgroups as shown in Sec.4.8 in the main paper. We take the following two settings: **1) Gender + Eyeglasses**: 0.40 *males with eyeglasses,* 0.10 *males without eyeglasses,* 0.40 *females with eyeglasses, and* 0.10 *females without eyeglasses*. **2) Gender + Race**: 0.40 *black males,* 0.10 *white males,* 0.40 *black females and* 0.10 *white females*. These two configurations are contrary to the originally generated distribution as blacks and eyeglasses are minority groups. The results are provided in Tab. 3, where we report FD with a reference set and FID score with the CelebA-HQ dataset to evaluate bias and generation quality.

Table 3. Generating imbalance distribution across subgroups to generate more images for under-represented groups.

	Gender + Eyeglasses (0.40,0.10,0.40,0.10)		Gender + Race (0.40,0.10,0.40,0.10)	
Method	FD	FID	FD	FID
Random Generation	1.1	49.45	1.444	49.45
Sample Guidance	0.472	48.66	0.756	62.48
Distribution Guidance	**0.38**	47.68	**0.464**	45.51

C. Debiasing text-to-image Diffusion Model

C.1. Social biases in facial attributes

We extend our experiments with Stable diffusion (SD) [5], previously outlined in Sec. 4.6 (main paper) by implementing our distribution guidance technique for debiasing across other (*race* and *age*) attributes. We observe that SD generations with a neutral prompt, such as *'a photo of a firefighter'* or *'a photo of a doctor'*, are images of people who are predominantly of white origin. Similarly, when prompted with *'a photo a person'*, SD majorly generates images of young people. We attempt to mitigate these biases with the proposed distribution guidance method following Sec.4.6. The qualitative results for the same are provided in Fig. 3. As evident from the results, our method achieves fairness across *race* and *age* attributes using the same neutral prompts. We present quantitative results in Tab. 4, where our method achieves superior FD scores computed using CLIP attribute classifier as explained in Sec.D.1.

C.2. Background bias in bird generation

Our method is applicable for mitigating biases that are not social as well (i.e. spurious correlations). WaterBirds [6] is a widely used synthetic dataset for demonstrating spurious correlations, which consist of images of birds across different backgrounds. The images of birds (landbirds and waterbirds) are collated from the CUB dataset [8], and the backgrounds from the Places dataset [9]. The majority of the waterbirds (sea birds) are affixed with a water background, and landbirds with land backgrounds. We generate a dataset of $10K$ images using SD with prompts *'photo of a land background'* and *'photo of a water background'*. Additional negative prompts like - *ocean, water, sea, shore, river* are added to generate pure *land* images. Similarly negative prompts *land, mountain, sand, forest, rocks* are added to generate pure *water* images. Next, we train a background *h-space* classifier for guidance. When queried with neutral prompts for landbird and waterbird, SD follows the spurious correlation in its generation. Specifically, when prompted for a landbird - *'a photo of a crow'*, SD predominantly generates a crow with land as the background, whereas when prompted for a waterbird - *'a photo of a duck'*, SD generates images with water in the background. When we apply distributional guidance, an equal number of images across both backgrounds are generated irrespective of the bird type. See Fig. 4 for qualitative results.

Figure 3. **Debiasing results on stable diffusion** for *race* and *age* attribute. We present random samples generated by the original stable diffusion model and with distribution guidance. **a)** Balancing across race involving different professions *e.g.* firefighter, doctor. **b)** Balancing across age attribute for neutral prompts.

Table 4. Balancing *race* for multiple professions and *age*

Method	Race-Doctor	Race-Firefighter	Age
Random Generation	0.356	0.423	0.488
Distribution Guidance	**0.191**	**0.186**	**0.194**

D. Implementation details

D.1. h-space classifiers

Training data. We created a paired training of *h-space* features and attribute labels $\mathcal{D}^h_{clf}$ for training *h-space* attribute classifiers. We start with CelebA-HQ [3] dataset $\mathcal{D}_{unf}$. Next, we used an off-the-shelf image space attribute classifier to obtain attribute labels for $\mathcal{D}$. We used CLIP as a classifier for both *gender* and *race* attributes as *race* labels are not present in CelebA-HQ dataset. Specifically, we pass $\mathcal{D}_{unf}$ to the CLIP[4] image encoder and obtain its similarity with text prompts - *'a male'* and *'a black person'*. The highest and lowest similarity images are then filtered to create a labeled attribute dataset $\mathcal{D}$. We used $|\mathcal{D}| = 2000$, where 1000 images are for the positive class, and others are for the negative class unless mentioned otherwise. For *eyeglass* attribute, we used ground truth labels from the CelebA-HQ dataset, as the predictions from CLIP were inaccurate. We then embed $\mathcal{D}$ into the *h-space* representation using DDIM[7] inversion to obtain labeled dataset $\mathcal{D}^h_{clf}$ to train the classifiers.

Model architecture. The h-space classifiers are implemented as a single linear layer for each diffusion timestep $\mathbf{t}$. We used DDIM inversion with $T = 49$ timesteps and obtain $\mathcal{H}^i = \{h^i_t\}^{t=49}_{t=0}$ as a set of 49 *h-vectors* for each image $\mathbf{i}$. As each classifier is linear with two output neurons (positive/ negative class), they can be jointly represented as a single *fully connected* layer with $2T$ output neurons.

Optimization. We train the *h-space* classifiers with the following hyper-parameters - batch size 64, learning rate 0.001, and for 5 epochs on a single NVIDIA $A5000$ gpu. The overall training time for a single attribute classifier is 484.37s.

D.2. Evaluation Metrics

We created a reference set $\mathcal{D}_{ref}$ of $5K$ images that follow the reference attribute distribution $\mathbf{p^a_{ref}}$, using the attribute labels from CelebA-HQ [2]. The ground truth labels were used for gender and eyeglasses from CelebA-HQ. For race, however, as the labels are not available, they were obtained from CLIP as discussed in Sec. D.1. To compute FID, we use $\mathcal{D}_{ref}$ to compute the reference statistics. For FD, we compute the discrepancy of the predicted attribute distribution from the reference distribution following Sec.4.1. We use the resnet-18 [1] architecture to implement the attribute classifiers trained on CelebA-HQ [2]. This set of classifiers *needs* to be different from the one used to obtain $\mathcal{D}^h_{clf}$ for fair evaluation.

Figure 4. **Debiasing results on stable diffusion for *backgrounds* while generating birds**. The proposed **Distribution Guidance** can balance the number of birds in various backgrounds.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 3

[2] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 2, 3

[3] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 3

[4] 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. 3

[5] 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

[6] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019. 2

[7] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 3

[8] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. *The Caltech-UCSD Birds-200-2011 Dataset*. 2011. 2

[9] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Antonio Torralba, and Aude Oliva. Places: An image database for deep scene understanding. *arXiv preprint arXiv:1610.02055*, 2016. 2