De-stereotyping Text-to-image Models through Prompt Tuning

Eunji Kim* [1] **Siwon Kim*** [1] **Chaehun Shin** [1] **Sungroh Yoon** [1 2]

Abstract

Recent text-to-image (TTI) generation models have been reported to generate images demographically stereotyped in various sensitive attributes such as gender or race. This may seriously harm the fairness of the generative model to be deployed. We propose a novel and efficient framework to de-stereotype the existing TTI model through soft prompt tuning. Utilizing a newly designed de-stereotyping loss, we train a small number of parameters consisting of the soft prompt. We demonstrate that our framework effectively balances the generated images with respect to sensitive attributes, which can also generalize to unseen text prompts.

1. Introduction

Recently proposed zero-shot text-to-image (TTI) models, exemplified by Stable Diffusion (Rombach et al., 2022), have surprised people with their remarkable ability to generate images solely from text prompts, approaching human-level performance. However, it has been observed that TTI models generate stereotyped images for various attribute-neutral prompts, e.g., the images generated with “A photo of a pilot” are completely stereotyped to males (Bianchi et al., 2023). However, Stable Diffusion is capable of generating images with minor attributes such as female pilot given the gender-specified text prompt (Bansal et al., 2022). These findings have raised the suspicion that the stereotype may originate from text prompts, leading to the development of methods to debias text prompts (Chuang et al., 2023).

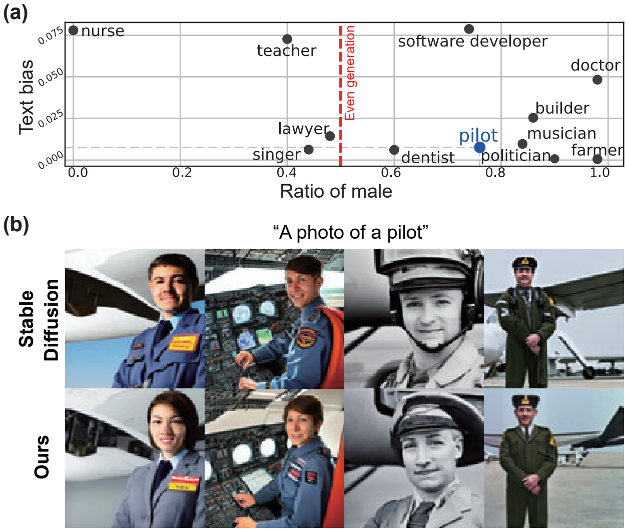

Figure 1. (a) Text bias vs. stereotypedness for 50 images from Stable Diffusion. (b) Example images generated by Stable Diffusion before and after applying our method, de-stereotyping gender bias.

Contrarily, our empirical observations indicate that stereotypes occur even when the text prompt is void of explicit biases. We assessed the bias of the text prompt by examining the difference between the cosine similarities of the text embedding with minor attribute embeddings and that with major attribute embeddings. If bias originates only from text, then the prompt with small difference should generate images with equal proportions of each attribute. However, as Figure 1(a) shows the normalized differences and the attribute ratios of generated images for various attribute-neutral texts, highly stereotyped images are generated from less biased text prompts as well (bottom right). The ratios have been measured with CLIP zero-shot classifier (Radford et al., 2021). This implies that the stereotypes are not solely derived from the text but also stem from image generation.

Motivated by this, we propose a novel framework to de-stereotype Stable Diffusion, one of the most widely used TTI models built on Latent Diffusion Models (LDMs) (Rombach et al., 2022). In our framework, the focus is not merely on attempting to eliminate bias in text, but rather on tuning the prompt to achieve a balanced ratio in generated images. Specifically, we append a de-stereotyping soft prompt to the original text prompt embeddings and train them with a newly proposed de-stereotyping loss. By incorporating image generation, our method alleviates the stereotype originating from both the text model and image generation model.

*Equal contribution [1]Department of Electrical and Computer Engineering, Seoul National University, Seoul, Korea [2]Interdisciplinary Program in Artificial Intelligence, Seoul National University, Seoul, Korea. Correspondence to: Sungroh Yoon <sryoon@snu.ac.kr>.

Workshop on Challenges in Deployable Generative AI at International Conference on Machine Learning (ICML), Honolulu, Hawaii, USA. 2023. Copyright 2023 by the author(s).

Since naive incorporation requires extensive memory, we design various memory-efficient engineering techniques.

For experimental results, we first show that the learned de-stereotyping prompt enables Stable Diffusion to generate a balanced ratio of sensitive attributes as depicted in Figure 1(b). We also demonstrate the efficiency of our method with generalization and transferrability analysis. Specifically, once the soft prompt has been trained, it exhibits notable effectiveness when applied to both previously unseen text prompts and text classes. This suggests that a TTI model in a deployment, or to be deployed, can be de-stereotyped across various prompts using a single soft prompt embedding of only a few dimensions in a plug-and-play manner.

2. Related Works

2.1. Bias and De-Bias in Text-to-image Generation

Cho et al. (2022) and Bianchi et al. (2023) revealed that demographic stereotypes are amplified in TTI-generated images more than that in the training dataset. Based on the finding, Chuang et al. (2023) proposed DebiasVL to debias discriminative and generative vision-language models, adopting a text embedding projection. However, it necessitates retraining for every new prompt. Orgad et al. (2023) proposed fine-tuning the cross-attention layer of the TTI model, but it is not feasible for the non-expert users who typically lack the authorization to modify the model parameters. In contrast, our method can generalize to unseen prompts and does not require direct tuning of TTI model parameters. Friedrich et al. (2023) introduced Fair Diffusion, a method that randomly selects sensitive attributes and guides a model to generate images with those selected attributes. In contrast, our method determines the attribute of the generated image solely based on the randomness of noise used in traditional TTI models to mitigate biases. Kim et al. (2023) focused on detecting novel biases and applied an existing de-biasing technique (Fair Diffusion) to address the identified biases.

2.2. Prompt Tuning

Prompt tuning is one of the transfer learning methods for improving downstream task performance of large foundation models (Lester et al., 2021). By assuming that a prompt-based large generative model already has sufficient knowledge, well-designed prompts can extract knowledge for a target downstream task from the frozen model. Various prompt tuning approaches have demonstrated a significant improvement in performance, training only a small number of parameters while leaving the model frozen (Jia et al., 2022; Zhou et al., 2022). However, there have been no attempts to tune prompts to make large generative models trustworthy.

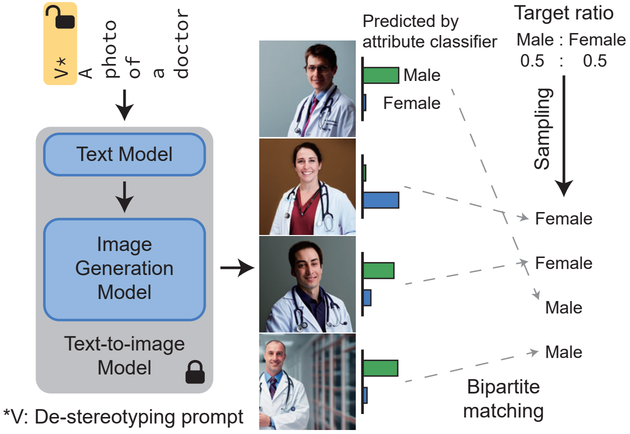

Figure 2. Key training scheme of our de-stereotyping method.

3. Problem Definition

3.1. Text-to-image Generation

TTI model generates images that faithfully depict the content specified in a given text prompt. It generally consists of two models: text model and image generation model. First, a text model $T_m(\cdot)$ takes a series of l tokens $\{x_1, x_2, ..., x_l\}$ as an input prompt X. The model then featurizes it as a token embedding X_e and encodes it to a text feature X_t. In the image generation model of Stable Diffusion (Rombach et al., 2022), latent z_0 is generated conditioned on X_t and an image y is decoded from the latent z_0 by an autoencoder.

3.2. Stereotyped generation

Let us refer to the object to be generated as a “class” and the elements that comprise a bias type as “attributes”. To determine if generated images are stereotyped with respect to a bias type, an auxiliary attribute classifier f that predicts the attributes of the images is employed. To define the stereotyped generation, let us denote a bias type with K attributes as $\mathcal{S} = \{s_1, ..., s_K\}$, ideal non-stereotyped ratio of attributes, or target ratio, as $\boldsymbol{p} = \{p_1, ..., p_K\}$, and N generated images depicting the class as $\mathcal{Y} = \{y_1, ..., y_N\}$. Then, the image generation of the class is stereotyped with the bias type if $\mathbb{E}_{y \sim \mathcal{Y}}[\mathbb{1}_{f(y)=s_k}] \not\approx p_k, \exists k$, where $s_k \in \mathcal{S}$. If the goal is a uniform generation of attributes, then $\boldsymbol{p} = \{1/K, ..., 1/K\}$.

The stereotypedness is measured by discrepancy which is a widely-used metric based on statistical parity (Choi et al., 2020; Chuang et al., 2023). In line with existing works, since we target the uniform generation in this paper, the discrepancy is defined as follows:

$$\mathcal{D} = \max_{s \in \mathcal{S}} \mathbb{E}_{\mathcal{Y}}[\mathbb{1}_{f(y)=s}] - \min_{s \in \mathcal{S}} \mathbb{E}_{\mathcal{Y}}[\mathbb{1}_{f(y)=s}]. \quad (1)$$

4. Methods

Our goal is to train soft prompt embeddings that guide a TTI model to generate images with even ratio of sensitive attributes. Hereinafter, we refer to the soft prompt as "de-stereotyping prompt". The target of the de-stereotyping is Stable Diffusion[1]. We adopt CLIP (Radford et al., 2021) zero-shot classifier as the auxiliary attribute classifier, f. Our method does not rely on training data, but instead, it trains the de-stereotyping prompt using images generated by Stable Diffusion based on provided text prompts.

4.1. De-streotyping Prompt Tuning

De-stereotyping prompt embedding $P_e \in \mathbb{R}^{L_p \times D}$ is appended ahead of original prompt embedding X_e, where L_p is the number of de-stereotyping prompt tokens and D is the token embedding dimension of the text model. The output of the text model becomes $T_m([P_e, X_e])$. Note that the prompt embedding P_e is the only parameter to be tuned for de-stereotyping the image generation process of a TTI model. Figure 2 illustrates the key training scheme of our method.

4.2. De-stereotyping Loss

The newly proposed loss to train P_e is defined as sum of de-stereotyping loss $\mathcal{L}_{\text{DS}}$ and regularization loss $\mathcal{L}_{\text{reg}}$. The final loss is $\mathcal{L}_{\text{DS}} + \lambda \mathcal{L}_{\text{reg}}$, where λ is a balancing factor.

De-stereotyping loss $\mathcal{L}_{\text{DS}}$ We introduce a loss for training a de-stereotyping prompt using the cross entropy loss, which encourages the generated images to be classified as various attributes by a fixed attribute classifier:

$$\mathcal{L}_{\text{DS}} = \mathbb{E}_{\mathcal{Y}} \text{CrossEntropy}(\hat{t}_s, t_s), \tag{2}$$

where $\hat{t}_s$ is an attribute that a generated image contains and t_s is a pseudo attribute label for the generated image.

The pseudo labels are sampled with probability distribution $\boldsymbol{p}$, which is the target ratio of attributes. We sample N pseudo labels and assign them to the N generated images. As the training progresses, images with minor attributes begin to be generated. The random assignment of pseudo labels may incur unnecessary training signals to reverse the images that have been well-generated with a minor attribute to include the major attribute. To reduce such unnecessary signals, we introduce bipartite matching (Carion et al., 2020) with cost based on the Hungarian algorithm (Kuhn, 1955). The pseudo labels are assigned in such a way that they are maximally aligned with the attributes present in the generated images (Figure 2).

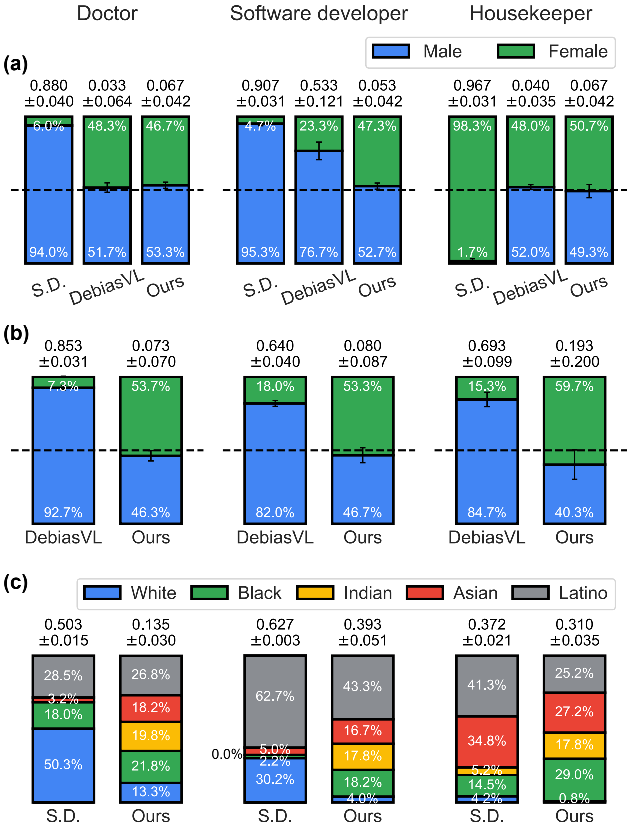

Figure 3. Attribute ratio within generated images with various methods. The discrepancy score is shown at the top of each bar. S.D. represents Stable Diffusion. Results for gender bias on (a) the seen text "A photo of a/an [class]" and (b) the unseen text "A/An [class] is laughing". (c) Results for racial bias on the seen text "A photo of a/an [class]".

Regularization loss $\mathcal{L}_{\text{reg}}$ The de-stereotyping prompt should not change the original content of a given text. To ensure that the generated images still faithfully depict a given text, we additionally employ $\mathcal{L}_{\text{reg}}$. It minimizes the difference between two latents from the last diffusion timestep before and after appending the de-stereotyping prompt. To prevent the potential impact of regularization on the de-stereotyping, an anchor text is introduced as the text with the pseudo label t_s, which is previously determined during calculating $\mathcal{L}_{\text{DS}}$. For instance, if a given text is "A photo of a doctor" and the pseudo label is female, the anchor becomes "A photo of a female doctor". The regularization loss is defined with an anchor text embedding $X_e^{t_s}$:

$$\mathcal{L}_{\text{reg}} = ||z_0(T_m([P_e, X_e])) - z_0(T_m(X_e^{t_s}))||_2, \tag{3}$$

where $z_0(\cdot)$ indicates a latent conditioned on a given text.

4.3. Memory-efficient Prompt Tuning

As the batch size increases, sampled pseudo labels more accurately reflect the target ratio. However, propagating gradients with a large batch size requires a significant amount

[1] https://huggingface.co/CompVis/stable-diffusion-v1-4

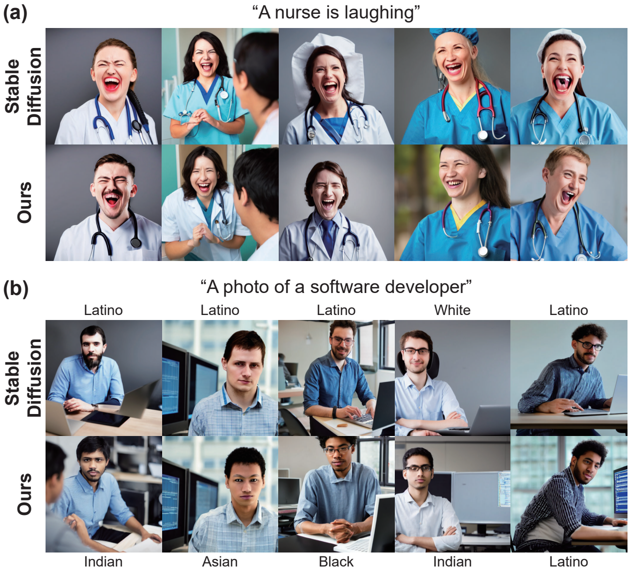

Figure 4. Examples of images generated by Stable Diffusion before and after applying our method. (a) De-stereotyping gender bias on unseen text prompt. (b) De-stereotyping racial bias.

of memory in Stable Diffusion. To utilize a larger batch size in limited memory, we propagate the gradient only through a subset of the generated images, which we refer to as an update batch. The remaining images are only dedicated to loss calculation, yet not used in back-propagation or parameter updates. We call this subset a non-update batch. We assign a higher weight to the cost in bipartite matching for the non-update batch images, so that if a misalignment is likely to occur, it should be encouraged within the update batch. Thus, we can obtain the informative gradients for updating the de-stereotyping prompt to generate images with minor attributes.

Stable Diffusion generates images through multiple diffusion steps, with text embeddings provided at each step. However, propagating gradients through all steps to update the de-stereotyping prompt requires a substantial amount of memory. To overcome this problem, we optimize the prompt using gradient skipping (Schwartz et al., 2023), which propagates the gradient solely through the last step.

5. Experimental Results

5.1. Experimental Settings

Following the previous works (Cho et al., 2022; Chuang et al., 2023), we evaluate our method on the two bias types: gender with male and female attributes, and race with White, Asian, Black, Indian, and Latino attributes. In our experiments, the input text prompt is "A photo of a/an [class]", and the attribute prompt used for CLIP zero-shot classification is "A photo of a/an [attribute]". The attribute ratio and discrepancy score are calculated with 100 and 200 generated images for gender and racial bias, respectively.[2] Implementation details are in Appendix A.

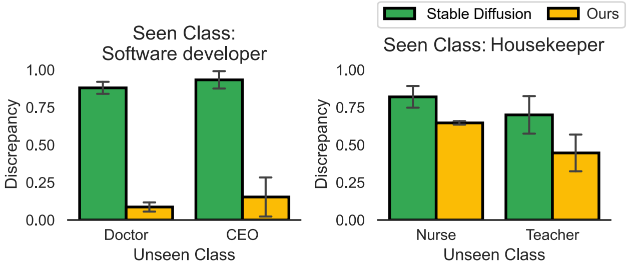

Figure 5. Transferability of de-stereotyping prompt to unseen classes.

5.2. De-stereotyping Results

We first evaluate the efficacy of the de-stereotyping prompt on the generation of the text prompt that was used for its training. Figure 3(a) shows the images generated with the learned soft prompt prepended to the original text prompt. It is shown that simply appending the trained prompt successfully enables an even generation of males and females, achieving a small discrepancy score. Figure 1(b) shows that some images of pilot generated by Stable Diffusion are changed to female after applying our method.

Figure 3(b) shows that our method can generalize well to unseen text prompt, "A doctor is laughing" with small discrepancy scores. On the other hand, DebiasVL fails to generalize to the unseen prompt showing a large discrepancy score of 0.853. This is compared to low discrepancy scores of DebiasVL for the seen prompt in Figure 3(a), implying that its optimization is strongly over-fitted on a training text prompt. Figure 4(a) displays examples images of generalization, where the laughing male nurses are generated after our de-stereotyping. More results are shown in Appendix B.1.

Our method also successfully reduces the discrepancy score on racial bias, as shown in Figure 3(c). The ratio of images with the dominant race decreases (White for doctor, Latino for software developer, and Latino for housekeeper). Please note that achieving a perfectly even balance is challenging due to the diversity of attributes involved in racial bias. Figure 4(b) shows the example images, and more images are in Appendix B.2.

5.3. Transferability analysis

We employ a de-stereotyping prompt that has been trained on a specific class to achieve a balanced generation of images of other classes. Figure 5 shows that appending the

[2] All reported results are presented along with the standard deviation calculated from three repetitions of the experiments.

de-stereotyping prompt trained on software developer to the prompt for doctor and CEO effectively reduces the discrepancy scores for each class. Also, the same result is observed when transferring the de-stereotyping prompt trained on housekeeper to nurse and teacher. These results demonstrate the transferability of our method to unseen classes, indicating the potential for TTI model providers or users to efficiently utilize the prompt for de-stereotyping their own prompts. More results are in Appendix B.1.

6. Conclusion

In this paper, we proposed a novel de-stereotyping framework by leveraging soft prompt tuning. Various experimental results demonstrated that Stable Diffusion becomes to generate images with an even ratio after the de-stereotyping. We also analyzed the noteworthy property of soft prompts, transferability, that enables further efficient de-stereotyping of TTI models in a plug-and-play manner. In future research, we aim to enhance the generalizability of the soft prompt for de-stereotyping text-to-image models. To achieve this, we will focus on optimizing memory usage with the expectation that a larger batch size will bring further improvements in de-stereotyping performance and the stability of training.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grants funded by the Korea government (Ministry of Science and ICT, MSIT) (2022R1A3B1077720 and 2022R1A5A708390811), Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (2022-0-00959 and 2021-0-01343: AI Graduate School Program, SNU), and the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2023.

References

Bansal, H., Yin, D., Monajatipoor, M., and Chang, K.-W. 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*, pp. 1358–1370, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

Bianchi, F., Kalluri, P., Durmus, E., Ladhak, F., Cheng, M., Nozza, D., Hashimoto, T., Jurafsky, D., Zou, J., and Caliskan, A. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1493–1504, 2023.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pp. 213–229. Springer, 2020.

Cho, J., Zala, A., and Bansal, M. Dall-eval: Probing the reasoning skills and social biases of text-to-image generative models. *arXiv preprint arXiv:2202.04053*, 2022.

Choi, K., Grover, A., Singh, T., Shu, R., and Ermon, S. Fair generative modeling via weak supervision. In *International Conference on Machine Learning*, pp. 1887–1898. PMLR, 2020.

Chuang, C.-Y., Jampani, V., Li, Y., Torralba, A., and Jegelka, S. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*, 2023.

Friedrich, F., Schramowski, P., Brack, M., Struppek, L., Hintersdorf, D., Luccioni, S., and Kersting, K. Fair diffusion: Instructing text-to-image generation models on fairness. *arXiv preprint arXiv:2302.10893*, 2023.

Jia, M., Tang, L., Chen, B.-C., Cardie, C., Belongie, S., Hariharan, B., and Lim, S.-N. Visual prompt tuning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, pp. 709–727. Springer, 2022.

Kim, Y., Mo, S., Kim, M., Lee, K., Lee, J., and Shin, J. Explaining visual biases as words by generating captions. *arXiv preprint arXiv:2301.11104*, 2023.

Kuhn, H. W. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Lester, B., Al-Rfou, R., and Constant, N. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059. Association for Computational Linguistics, November 2021.

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

Orgad, H., Kawar, B., and Belinkov, Y. Editing implicit assumptions in text-to-image diffusion models. *arXiv preprint arXiv:2303.08084*, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Schwartz, I., Snæbjarnarson, V., Benaim, S., Chefer, H., Cotterell, R., Wolf, L., and Belongie, S. Discriminative class tokens for text-to-image diffusion models. *arXiv preprint arXiv:2303.17155*, 2023.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.

A. Experimental Details

Text prompts for zero-shot attribute classification Text prompt for zero-shot attribute classification is determined depending on a given text prompt for image generation. For a given text prompt, [class] is replaced with attributes. For example, when images are generated with the text "A photo of a/an [class]", the attributes are classified with the text "A photo of a/an [attribute]". When images are generated with the text "A/An [class] is laughing", the attributes are classified with the text "A/An [attribute] is laughing". Among racial categories, since White and Black mean universal colors, we use "White person" and "Black person" to specify the meaning of race.

Bipartite matching The target attribute for each generated image is determined via bipartite matching based on the Hungarian algorithm. The match is determined as that minimizes the sum of negative likelihood for images in a batch by finding the optimal permutation ρ^* as follows:

$$\rho^* = \text{argmin}_\rho \sum_i -w_i \text{Likelihood}(\hat{t}_s^{(i)}, t_s^{(i)}(\rho)), \tag{4}$$

where w_i denotes matching weight for i-th generated image. We set w_i to 10000 for images in non-upate batch, otherwise 1.

De-stereotyping prompt We use two prompt tokens for de-stereotyping ($L_p = 2$). We initialize P_e with the text embeddings of sensitive attributes. We collect all embeddings of the sensitive attributes and initialize P_e with the average embeddings. For example, when we de-stereotype gender bias, the prompt embeddings are initialized with the average of text embeddings of male and female.

Implementation details Our experiments were conducted on RTX8000 GPUs using PyTorch and automatic mixed precision training. A balancing factor λ was set to 0.1. For gender bias, we used one GPU with an update batch size of 3 and a non-update batch size of 9. Since there are five distinct race categories, we used more computational resources and time to achieve a reliable target ratio, by employing three GPUs for racial bias with an update batch size of 9 and a non-update batch size of 27. We used AdamW optimizer (Loshchilov & Hutter, 2017) with a learning rate 1×10^{-4} and 5×10^{-4} for gender and race, respectively. Additionally, we accumulated the gradient for five iterations and trained the de-stereotyping prompt for gender and race for 70 and 100 iterations, respectively. We reduced the learning rate by 0.2 whenever there was a change in the dominant sensitive attribute among the generated images.

B. Experimental Results

B.1. Quantitative Results

Unseen text prompts Table 1 compares the discrepancy score of DebiasVL and our method. Both methods are optimized on the text prompt "A photo of a/an [class]" and evaluated on various unseen text prompts. Our method almost always achieves a lower discrepancy score than DebiasVL, indicating that our method is more generalizable to various text prompts.

Transferability to unseen classes Table 2 reports the discrepancy scores when de-stereotyping prompts are applied for generating images with unseen classes. It shows that the de-stereotyping prompt can be transferred to unseen classes with the same dominant attribute.

Table 1. Discrepancy of generated images with various methods on unseen text prompts.

	Doctor	CEO	Software developer	Nurse	Housekeeper
"A painting of a/an [class]"					
DebiasVL	0.773±0.050	0.840±0.053	0.933±0.031	0.813±0.050	0.073±0.081
Ours	0.433±0.117	0.327±0.127	0.133±0.031	0.160±0.131	0.127±0.023
"A/An [class] is laughing"					
DebiasVL	0.860±0.020	0.973±0.012	0.620±0.035	0.307±0.090	0.673±0.083
Ours	0.093±0.090	0.240±0.020	0.087±0.090	0.267±0.081	0.200±0.164
"A photo of a/an [class] in the workplace"					
DebiasVL	0.660±0.035	0.540±0.072	0.800±0.035	0.853±0.050	0.407±0.095
Ours	0.460±0.072	0.440±0.069	0.173±0.023	0.147±0.122	0.327±0.115

Table 2. Transferability of de-stereotyping prompt to other classes

Seen class	Dominant	Class				
	attribute	Doctor	CEO	Software developer	Nurse	Housekeeper
Stable Diffusion		0.880±0.040	0.933±0.058	0.907±0.031	0.820±0.072	0.967±0.031
Ours						
Doctor	Male	0.067±0.042	0.233±0.101	0.180±0.035	0.973±0.031	0.853±0.042
CEO	Male	0.040±0.040	0.060±0.035	0.393±0.117	0.960±0.000	0.927±0.031
Software developer	Male	0.087±0.031	0.153±0.130	0.053±0.042	0.967±0.023	0.893±0.042
Nurse	Female	1.000±0.000	0.993±0.012	0.993±0.012	0.120±0.106	0.620±0.035
Housekeeper	Female	0.973±0.012	0.960±0.020	0.907±0.050	0.647±0.012	0.067±0.042

B.2. Qualitative Results

Figures 6 and 7 show examples of images generated by Stable Diffusion before and after applying our method. With our method, Stable Diffusion become to generate images with non-dominant attributes.

C. Ablation Study

We analyze the effect of de-stereotyping loss $\mathcal{L}_{\text{DS}}$, which encourages the soft prompt to enable de-stereotyping. Table 3 shows that the de-stereotyping prompt fails to effectively reduce bias, leading to a high discrepancy score, when training without $\mathcal{L}_{\text{DS}}$. Even when $\mathcal{L}_{\text{DS}}$ is used, making the TTI model to generate images with a target attribute ratio is challenging without bipartite matching.

Table 3. Ablation study for $\mathcal{L}_{\text{DS}}$ on gender bias with text prompt “A photo of a software developer”.

$\mathcal{L}_{\text{DS}}$	Bipartite matching	Discrepancy
✓	✓	0.053±0.042
✓		0.447±0.090
		0.933±0.031

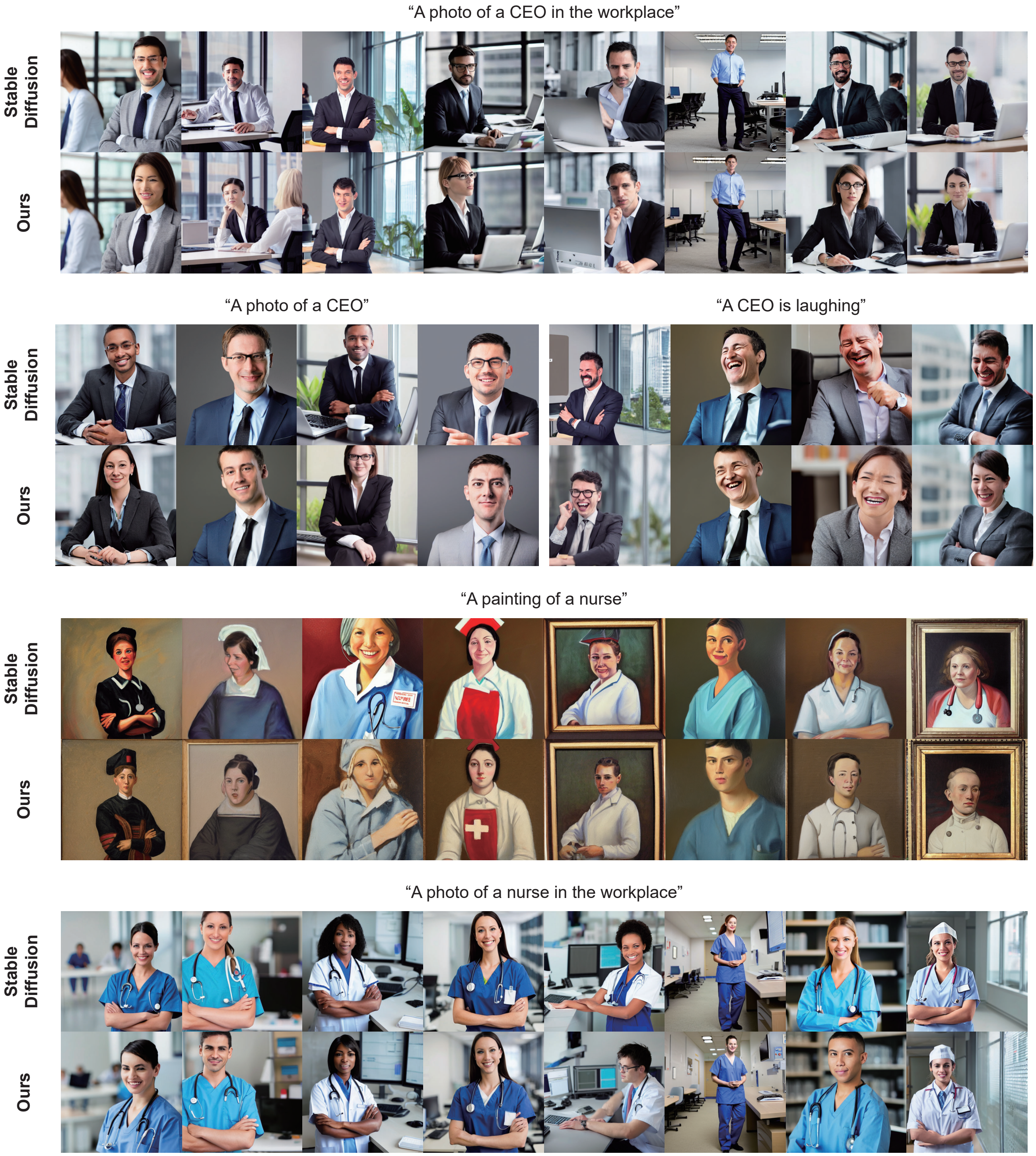

Figure 6. Examples of images generated by Stable Diffusion before and after applying our method on gender bias.

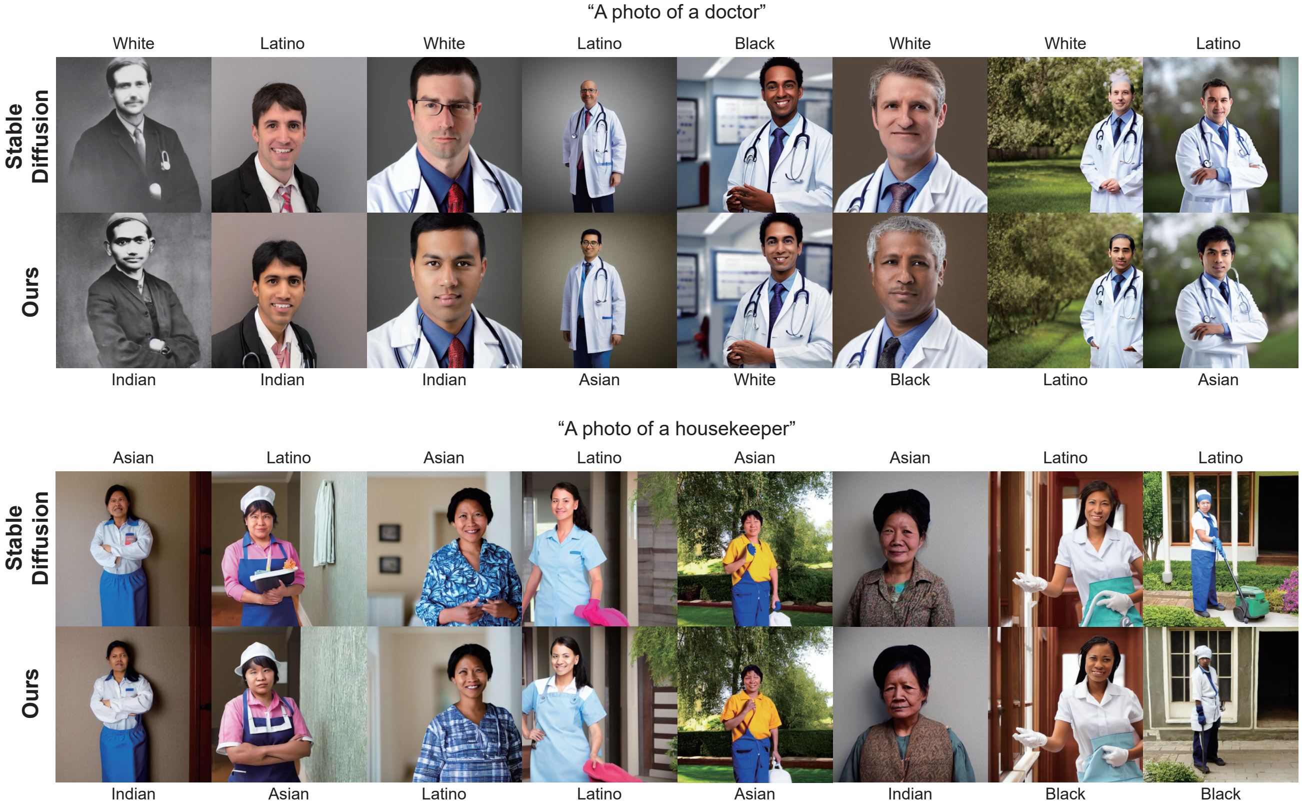

Figure 7. Examples of images generated by Stable Diffusion before and after applying our method on racial bias.