Exploring Intrinsic Fairness in Stable Diffusion

Eunji Kim[1*] **Siwon Kim**[1*] **Robin Rombach**[§]
Rahim Entezari[4†] **Sungroh Yoon**[1,2†]

[1] Department of Electrical and Computer Engineering, Seoul National University
[2] Interdisciplinary Program in Artificial Intelligence, Seoul National University
[3] Stability AI

Abstract

Recent text-to-image models like Stable Diffusion produce photo-realistic images but often exhibit demographic biases. Previous debiasing efforts have predominantly focused on introducing training-based debiasing approaches, neglecting to investigate the root causes of these biases and overlooking Stable Diffusion's potential for generating unbiased images. In this paper, we demonstrate that Stable Diffusion inherently possesses fairness, which can be unlocked to achieve debiased outputs. We conduct carefully designed experiments to analyze the effect of initial noise sampling and text guidance on biased image generation. Our analysis reveals that an excessive correlation between text prompts and the diffusion process is a key source of bias.

1 Introduction

Recent text-to-image (T2I) generation models, such as Stable Diffusion (SD) [5, 16, 17], demonstrate photo-realistic image generation performance. Despite the ground-breaking image quality, these models often generate biased images, *i.e.*, an imbalanced ratio between major and minor sensitive attributes such as gender or race [2, 11, 15, 20]. Since T2I models are trained on real-world images that inherently contain bias, it is unsurprising that the generated images also reflect this bias. However, studies [15, 20] revealed that bias is often amplified in generated images compared to the training data, *i.e.*, the disparity in the ratio of major and minor attributes is exacerbated in generated images. While opinions regarding the definition of fairness may vary, there is consensus that such biases should not be exacerbated.

Several methods have been proposed to mitigate bias in SD [4, 6, 10, 21, 14], most of which involve additional training. This leads us to an important question: Are the generated images truly reflective of SD's inherent bias? If we can identify intrinsic fairness within SD, we could potentially reduce bias, lower costs, and maintain the essential image generation capabilities. To the best of our knowledge, this potential solution has not been explored.

In this paper, we investigate intrinsic fairness in SD and explore a potential direction to unleash it. We first propose a *mode test* in section 2 wherein we examine initial noise of SD. Our investigation particularly focuses on noise in the low-density regions of the probability distribution, which has been underexplored as images are typically generated from high-density regions. Mode test results suggest that a greater portion of noise than expected can generate a minor attribute. We then examine the effect of *weakening* of the text condition guidance that directs noise from high-density regions

[*]Co-first authors ({kce407, tuslkkk}@snu.ac.kr)
[§]Work done while at Stability AI
[†]Senior authorship with Rahim Entezari and Sungroh Yoon (corresponding author: sryoon@snu.ac.kr)

Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models
38th Conference on Neural Information Processing Systems (NeurIPS 2024).

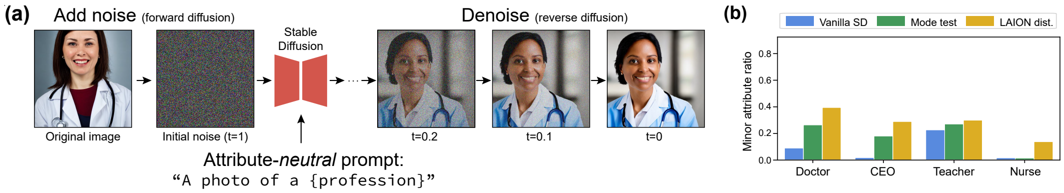

Figure 1: (a) Illustration of our mode test. Noise is added to minor attribute images, followed by a reverse diffusion process using an attribute-neutral prompt. (b) More minor attribute images are generated through the mode test (section 2).

to generate major attributes. As a means to achieve this, two approaches, 1) explicitly decreasing the strength of text condition and 2) perturbing it by adding noise, are examined in sections 3.1 and 3.2. The experimental results show that both approaches are effective in mitigating bias, but also undermining an image-text alignment, necessitating a more carefully designed perturbation scheme. As a final analysis, we demonstrate that perturbation accompanied by guidance toward the minor attribute during the early diffusion steps can be a potent alternative in section 3.3. Our analysis suggests that weakening the bond between the text guidance towards the major attribute and the diffusion process is a promising direction for debiasing.

2 Discovering Fairness in SD

Analysis setting. Before delving into our analysis, we outline the experimental setup used throughout the paper. The main analyses use SD-v1.5[1], with additional results for SD-v2 and SDXL in the Appendix to support the generalizability of our findings. We primarily focus on binary gender bias (male and female) in four different professions (doctor, CEO, nurse, and teacher). We use the CLIP zero-shot classifier [2] with the prompts ‘‘`A photo of a male/female`’’ to determine the gender in generated images. When testing with racial bias, text prompts ‘‘`A photo of a/an White person/Black person/Asian/Indian/Latino`’’ are utilized following [4]. The most frequent attribute in generated images is termed as major, while others are denoted as minor.

Analysis with noise. This paper addresses the issue of amplified bias that occurs even with attribute-neutral prompts. We examine the increased disparity between major and minor attributes in generated images compared to the training images. This suggests that initial noises, primarily sampled from high-density regions in the probability distribution, tend to strongly favor a major attribute when conditioned with an attribute-neutral prompt. However, it remains unclear whether noise in low-density regions is also prone to generate major attributes. Since the majority of generated images are from high-density regions, resolving this necessitates further investigation into the low-density regions.

To facilitate this investigation, we propose a *mode test*. Given that directly accessing low-probability noises is challenging due to their rare sampling, we opt to simulate them instead. Specifically, we intentionally generate minor attribute images with SD-v1.5 using minor attribute-specified prompts and then add noise to them, simulating a forward diffusion process. Inspired by SDEdit [12], we then apply reverse diffusion to the resulting noise while conditioning it with attribute-neutral prompts. Figure 1(a) depicts the overall flow of the mode test. If the images are regenerated with minor attributes despite using attribute-neutral prompts, it supports the presence of previously undetected noises in low-density regions that can be generated into minor attributes.

Figure 1(b) compares the minor attribute ratio in vanilla SD generated images and mode test generations. The ratio in LAION-5B [19], is also depicted for reference as reported in [20]. For all four professions, the mode test increases the ratio of minor attributes compared to the vanilla SD, aligning the results more closely with the LAION-5B distribution. This suggests that noises from low-density regions can generate minor attributes, even with neutral prompts. These noises were likely overlooked because initial noise sampling usually targets high-density regions. This observation indicates that SD has inherent fairness, and utilizing this fairness can help reduce bias.

[1] `https://huggingface.co/runwayml/stable-diffusion-v1-5`

[2] `https://huggingface.co/openai/clip-vit-base-patch32`

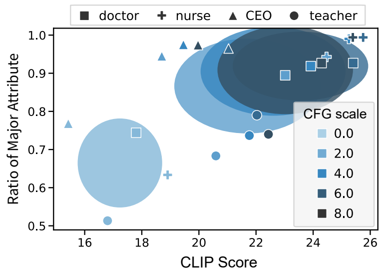

Figure 2: Impact of CFG: Increasing CFG scale increases both major attribute ratio and CLIP score (section 3.1).

Figure 3: Samples from vanilla SD-v1.5 and CADS ($\tau_1 = 0.6, \tau_2 = 0.9, s = 0.25$) applied, with "`a photo of a doctor`". CADS diversifies gender and race but sometimes compromises prompt alignment (section 3.2).

3 Key to Unlocking Fairness in SD

From our mode test analysis, we hypothesize that the text condition is the primary factor guiding initial noise from high-density regions to generate major attributes. If this is correct, reducing the influence of the text condition on the diffusion process should alleviate bias. To test this hypothesis, we conduct two experiments to intentionally weaken the effect of the text condition: 1) decreasing the classifier-free guidance scale (section 3.1) and 2) using noisy text conditions (section 3.2). We also examine the effect of directly guiding towards minor attributes (section 3.3).

3.1 Impact of Classifier Free Guidance

The Classifier-Free Guidance (CFG) [8] directs image generation to reflect the semantics of the text condition. Specifically, with CFG, the predicted noise $\tilde{\epsilon}_\theta$ can be written as $\tilde{\epsilon}_\theta(\boldsymbol{z}, \boldsymbol{c}) = (1+\alpha) \cdot \epsilon_\theta(\boldsymbol{z}, \boldsymbol{c}) - \alpha \cdot \epsilon_\theta(\boldsymbol{z})$, where $\boldsymbol{z}$ and $\boldsymbol{c}$ denote unconditional and conditional text prompt embedding, respectively, and α denotes the CFG scale. It is known that a larger α, *i.e.*, a stronger guidance, yields higher coherence of the image to the text condition at the cost of reduced sample diversity [8]. Conversely, this suggests that reduced CFG scale can diversify generated images.

Here we study how bias changes by varying the CFG scale from 0.0 to 8.0. Figure 2 shows the major attribute ratio (y-axis) and CLIP score (x-axis). Color intensity reflects the magnitude of the CFG scale. As the CFG scale decreases (indicated by lighter colors), the major attribute ratio decreases. These results support our hypothesis that weakening a text condition can alleviate bias. Consequently, it also compromises the alignment between the generated images and the text prompts.

3.2 Noisy Text Condition

We describe an alternative approach that weakens text conditions by perturbing them with injected noise. This approach is inspired by Condition-Annealed Sampling (CADS) [18], which proposes to add noise to a text condition to diversify compositions of generated images. The CADS operates as follows: a given text condition c is perturbed to $\hat{c}$ as

$$\hat{\boldsymbol{c}} = \sqrt{\gamma(t)}\boldsymbol{c} + s\sqrt{1-\gamma(t)}\boldsymbol{n}, \quad \gamma(t) = \begin{cases} 1 & 0 \le t \le \tau_1, \\ \frac{\tau_2 - t}{\tau_2 - \tau_1} & \tau_1 < t < \tau_2, \\ 0 & \tau_2 \le t \le 1, \end{cases} \tag{1}$$

where s controls the scale of noise, $\gamma(t)$ is the annealed coefficient determined by t, and $\boldsymbol{n} \sim \mathcal{N}(0, I)$. As diffusion models operate reverse from $t = 1$ to $t = 0$, perturbation with noise is applied to a text condition in earlier steps. $\hat{\boldsymbol{c}}$ is then normalized to have the same mean and standard deviation as $\boldsymbol{c}$.

To study the impact of the CADS-based approach on diversifying attributes, we conduct experiments addressing gender and racial bias. The results are shown in Figure 4 (a,b) where the ratio of the major attribute is depicted in y-axis. It is shown that the major attribute ratio decreases, indicating that bias is mitigated by CADS (all variations) compared to vanilla SD (blue) for both gender and racial bias. We also observe that as s increases from 0.15 (yellow) to 0.25 (red) or τ_1 decreases from 0.8 (green) to 0.6 (red), bias mitigation becomes more pronounced. An increase in s or a decrease in τ_1

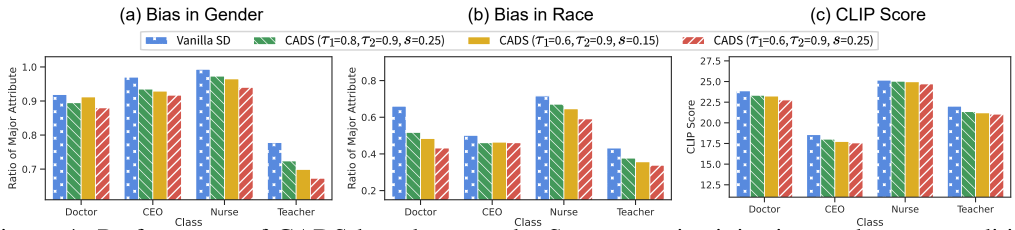

Figure 4: Performance of CADS-based approach. Stronger noise injection to the text condition (higher s and lower τ_1) mitigates bias (a, b) while increasing CLIP score (c) (section 3.2).

indicates stronger perturbation. These observations also validate our initial hypothesis that weakening text conditions helps mitigate bias. Figure 3 compares the images generated with vanilla SD and CADS. While CADS-generated images display diverse gender and race attributes, the alignment between prompt and generated images degrades as the intensity of perturbation increases. This is also evidenced by Figure 4(c) which shows decreased CLIP scores with CADS. These results indicate that while text prompt perturbation effectively reduces bias, it requires more careful design to maintain Stable Diffusion's image generation capabilities.

3.3 Text Guidance with Minor Attribute

The results in the previous sections reveal that while perturbing text conditions can steer initial noise towards creating minor attributes—helping to reduce bias—uncontrolled perturbations can disrupt image-text alignment. To address this, it is beneficial to control the perturbation by providing guidance in the desired direction—in our case, the direction of a minor attribute.

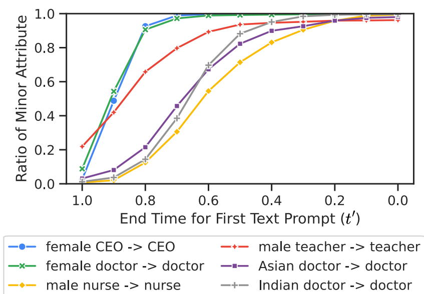

Figure 5: Ratio of the minor attribute. The x-axis indicates the variation in the initial steps that include a minor attribute in the text condition. As t' decreases in the intermediate steps, the minor attribute ratio increases (section 3.3).

Here we investigate whether conditioning the early diffusion steps with a minor attribute-specified prompt aids in bias mitigation by generating more images with minor attributes. Specifically, when generating images for a neutral prompt, we replace the text condition in the early diffusion steps from $t = 1$ to $t = t'$ with a minor attribute-specified prompt. We keep the neutral prompts for the remaining steps, from $t = t'$ to $t = 0$. Figure 5 shows the minor attribute ratio by varying the initial steps that include a minor attribute in the text condition (x-axis). When $t' = 1$, only the neutral prompt is used, leading to biased outputs. When $t' = 0$, using only the minor attribute prompt drives the minor attribute ratio close to 1. As t' decreases in the intermediate steps, the ratio of minor attributes steadily increases, indicating that guiding early diffusion with a minor attribute-focused prompt effectively mitigates bias.

4 Conclusion

In this paper, we tackle the bias in images generated by Stable Diffusion by systematically studying its root causes and exploring its intrinsic fairness. Our experiments reveal that excessive bonding between text prompts and the diffusion process is a key source of bias. Weakening this bond is crucial for debiasing; however, reducing text guidance with noise and lowering the classifier-free guidance scale can compromise image quality. We also found that the guidance towards the minor attributes in early diffusion steps can reduce bias. We believe our findings can inspire new debiasing strategies.

Broader Impacts. Our novel analysis of the low-density region in the initial noise space opens new avenues for exploring intrinsic fairness in Stable Diffusion, potentially leading to more equitable generative models.

Limitations. Our study primarily focuses on binary gender and five racial categories, which do not encompass all demographic groups. Future research should explore a wider range of biases.

Acknowledgments and Disclosure of Funding

This work was supported by Stability.ai, Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)], the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2024, the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022R1A3B1077720, 2022R1A5A708390811).

References

[1] 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.

[2] 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.

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

[4] 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.

[5] 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. *arXiv preprint arXiv:2403.03206*, 2024.

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

[7] 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.

[8] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[9] Eunji Kim, Siwon Kim, Chaehun Shin, and Sungroh Yoon. De-stereotyping text-to-image models through prompt tuning. *ICML 2023 Workshop on DeployableGenerativeAI*, 2023.

[10] 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 (CVPR)*, June 2024.

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

[12] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022.

[13] 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.

[14] Rishubh Parihar, Abhijnya Bhat, Saswat Mallick, Abhipsa Basu, Jogendra Nath Kundu, and R Venkatesh Babu. Balancing act: Distribution-guided debiasing in diffusion models. *arXiv preprint arXiv:2402.18206*, 2024.

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

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

[17] 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.

[18] Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M. Weber. CADS: Unleashing the diversity of diffusion models through condition-annealed sampling. In *The Twelfth International Conference on Learning Representations*, 2024.

[19] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[20] Preethi Seshadri, Sameer Singh, and Yanai Elazar. The bias amplification paradox in text-to-image generation. *arXiv preprint arXiv:2308.00755*, 2023.

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

A Related Works

De-biasing text-to-image generation models. Most of existing methods grant fairness to SD by using additional resources. However, fully fine-tuning large T2I models is highly costly. Recent methods have relied on parameter-efficient fine-tuning techniques, such as prefix tuning [9], text embedding projection weight [4], or low-rank adaptation [21]. Additionally, there have been attempts to modify the cross-attention layer in the UNet of Stable Diffusion [7, 13]. Another line of work has proposed directly fine-tuning h-space vectors, which are vectors from the bottleneck layer of UNet known to contain rich semantics [10, 14]. However, there has been little examination of whether additional training is truly necessary.

Only a few de-biasing methods bypass additional training altogether, instead focusing on modifying text prompts by adding words or phrases. The most naive approach [1] involves adding ethically intervening words or phrases into the initial prompts. FairDiffusion [6] directly perturbs the diffusion direction by employing a concept editing method called SEGA [3].

B Experimental Details

B.1 Common Settings

For all experiments, we generate 1,000 images with 50 steps using the PNDM scheduler. Images are generated at 512×512 for SD-v1.5 and SD-v2 [3], and at 1024×1024 for SDXL [4]. Unless specified otherwise, we use a CFG scale α of 6 for SD-v1.5 and SD-v2, and a scale of 4 for SDXL. The experiments are done with NVIDIA RTX 8000 and A40.

B.2 Noisy Text Condition

We start with the default settings of CADS and set (τ_1, τ_2) to $(0.6, 0.9)$ and $s = 0.25$. To further explore the impact of the intensity and duration of noise injection on bias mitigation, we also extend our experiments with additional hyperparameters: $(\tau_1, \tau_2, s) = (0.8, 0.9, 0.25)$ and $(0.6, 0.9, 0.15)$.

[3] `https://huggingface.co/stabilityai/stable-diffusion-2-base`

[4] `https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0`

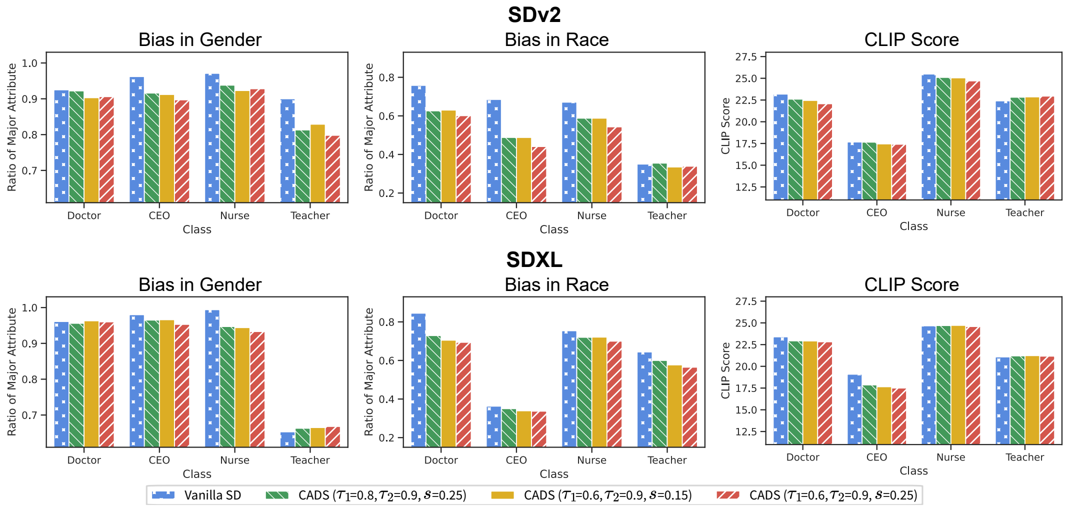

Figure 6: Change in ratio of major attribute and CLIP score when CADS is used with SD-v2 and SDXL.

C Additional Results for Exploring and Unlocking Fairness of Stable Diffusion

C.1 Noisy Text Condition

Figure 6 shows that adding noise via CADS reduces gender and racial bias within image generation of SD-v2 and SDXL. As explained in section 3.2 for SD-v1.5 results, increasing the amount of noise injected to the text condition (with larger s and smaller τ_1) decreases the ratio of major attributes, thereby reducing bias within both gender and race. For the result with teacher, the change is minimal (racial bias within SD-v2) or even increases the ratio of the major attribute (gender bias within SDXL), where bias in vanilla SD-generated images is not as severe as other professions.

Figures 8, 9, and 10 illustrate some examples of generated images with a vanilla SD and CADS, using SD-v1.5, SD-v2, and SDXL, respectively. CADS generates more diverse images, reducing bias. However, it occasionally fails to generate images that match with the given text prompt. This is also reflected in the decrease in the CLIP score shown in Figure6.

The findings suggest that injecting noise to perturb the text condition, as demonstrated by CADS, aids in mitigating bias across various versions of SD. Nonetheless, as discussed in the main text, it may potentially compromise the alignment between images and text.

C.2 Minor Attribute Guidance

Figure 7 illustrates the experimental results regarding minor attribute guidance with SD-v2 and SDXL, as elaborated in section 3.3. As the end time (t') for the minor attribute-specified prompt decreases, the ratio of minor attribute increases. With SDXL, employing a minor attribute-specified prompt from $t = 1$ to $t = 0.5$ ($t' = 0.6$) results in over approximately 90% of the images being generated with minor attributes across most professions. With SD-v2, a longer duration of employing a text prompt specifying a minor attribute was required to achieve a similar minor attribute ratio. This observation demonstrates that guiding the diffusion process with a prompt specifying a minor attribute during the initial diffusion steps is effective across various versions of SD.

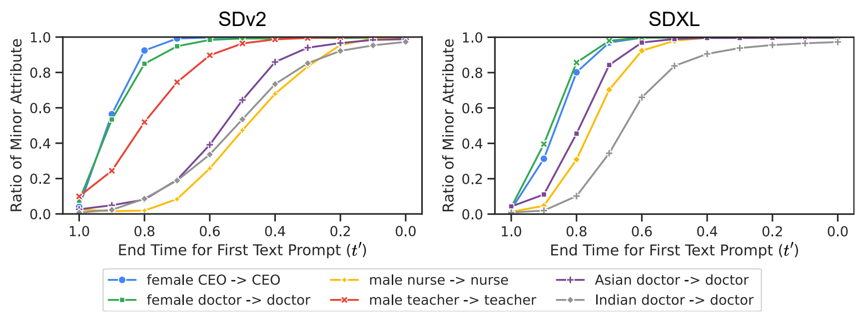

Figure 7: Ratio of minor attributes within the generated images using both minor attribute-specified text prompt and attribute-neutral text prompt.

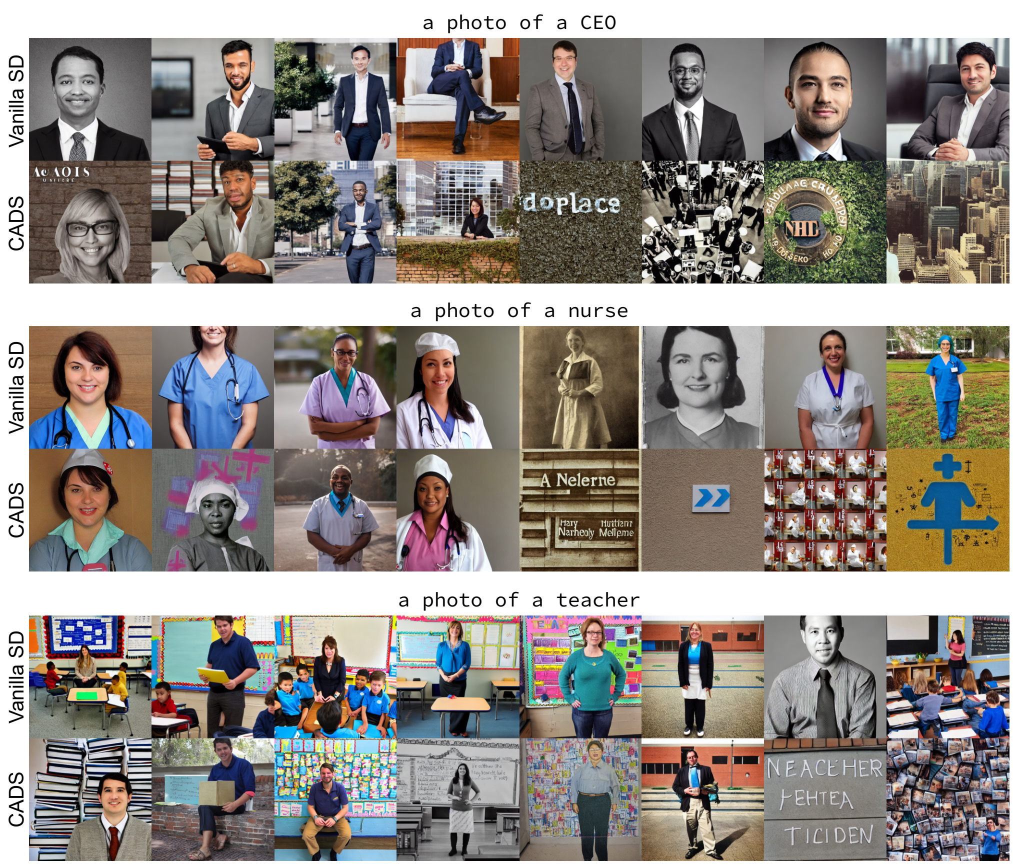

Figure 8: Examples of generated images with vanilla SD and CADS, using SD-v1.5.

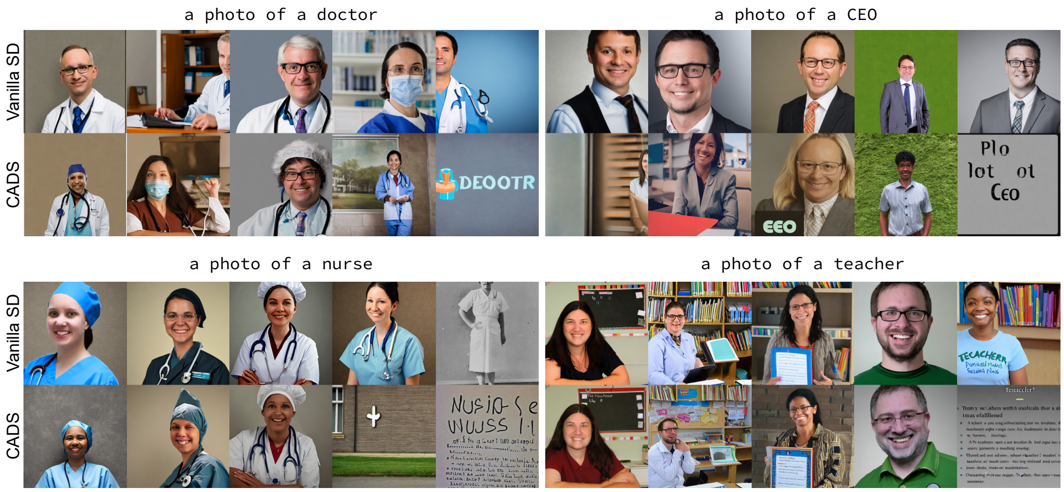

Figure 9: Examples of generated images with vanilla SD and CADS, using SD-v2.

Figure 10: Examples of generated images with vanilla SD and CADS, using SDXL.

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: We reviewed the abstract and introduction and checked that they accurately reflect the paper's contributions and scope.

   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: We discussed the limitations of our work in Conclusion.

   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: Our paper does not include theoretical results.

Guidelines: The paper does not contain any theoretical result.

- 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: We included all the information needed to reproduce the experimental results in the main text and appendix.

   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: We provide open access to the code in the abstract. Additionally, experimental details are demonstrated in the main text and the appendix.

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, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We specified all details in the Experimental Results section and the appendix.

   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: [Yes]

   Justification: We mentioned the number of experiments we conducted.

   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: We included the information on the computer resources in the appendix.

   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: We have read through Code of Ethics and checked that our research conform with them.

   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: We discussed broader impacts and limitations in the conclusion section.

    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: We did not release any data or models.

    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: We included citations for the cited papers.

    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: We did not introduce any new assets in the paper.

    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: Our paper does not include such experiments or researches.

    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: Our paper does not include such potential risk.

    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.