

Rethinking Training for De-biasing Text-to-Image Generation: Unlocking the Potential of Stable Diffusion

Supplementary Material

A. Experimental Details

A.1. Common Settings

For all experiments conducted with SD1.5, SD2, and SDXL, images are generated using 50 steps with the PNDM scheduler, while SD3 employs the FlowMatchEulerDiscrete scheduler with 28 steps. Image resolutions are set to 512×512 for SD1.5 and SD2³, and 1024×1024 for SDXL⁴ and SD3⁵. Unless stated otherwise, we use a CFG scale (α) of 6 for SD1.5 and SD2, and 4 for SDXL and SD3. The experiments are performed using NVIDIA RTX 8000, A40, and H100 GPUs.

A.2. Settings for Comparison Methods

All the baseline methods we evaluate were originally developed using SD1.4 and SD1.5, so our primary comparisons are conducted using SD1.5. Additionally, we assess our method against FairDiffusion, the only training-free baseline, using SD2 with the same hyperparameters as in SD1.5. Since SDXL features a distinct architecture from SD1.5, we report results only for vanilla SD and our method in this case.

For FairDiffusion, we adopt the default hyperparameters provided in its official implementation. For FTDiff, the pre-trained checkpoints for both the text encoder and LoRA are used. For UCE and SelfDisc, we strictly follow the training and inference procedures outlined in their official codebases. Detailed settings for each method are provided below.

FairDiffusion ⁶ The editing prompts are set to [“female person”, “male person”]. The direction of guidance—subtracting or adding—is randomly selected between the two concepts. The guidance scale for editing is set to 3. The threshold, the momentum scale, and momentum beta are set to 0.9, 0.5, and 0.6, respectively. The weights of the individual concepts are set equally to 1.

UCE ⁷ We trained UCE from scratch following the instructions provided in the official GitHub repo. UCE training

³<https://huggingface.co/stabilityai/stable-diffusion-2-base>

⁴<https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0>

⁵<https://huggingface.co/stabilityai/stable-diffusion-3-medium-diffusers>

⁶<https://github.com/ml-research/Fair-Diffusion>

⁷<https://github.com/rohitgandikota/unified-concept-editing>

requires a pre-defined set of professions and the model reported in the main paper is trained on 36 professions, but for fair comparison with other methods, we remove the professions overlapping with the eight professions used in our paper.

FTDiff ⁸ We use the pre-trained checkpoint provided by the authors. FTDiff provides various checkpoints where different components of SD are trained, such as prefix tuning or LoRA fine-tuning. We used ‘from-paper_finetune-text-encoder_09190215checkpoint-9800_exportedtext_encoder_lora.EMA.pth’.

SelfDisc ⁹ Since the official code does not provide a pre-trained checkpoint, we trained SelfDisc from scratch with default settings and hyperparameters provided by the authors. Then we generated images by uniformly sampling attributes.

A.3. Detailed Setting of Our Method

Our method requires no additional training, making it adaptable to any version of Stable Diffusion with minimal effort. The only parameter to configure is τ , which determines the initial time steps during which \hat{c} is used exclusively. As described in the main text, subsequent time steps alternate between using \hat{c} and c , ensuring a balanced integration of guidance throughout the process.

For SD1.5 and SDXL, we set τ to 0.9. In SD2 and SD3, where the guidance strength for adding attribute directions is weaker, τ is adjusted to 0 and 0.5, respectively. In the case of SD3, which incorporates three text encoders, we add the attribute direction across all positions for the T5 text encoder. For models like SDXL and SD3, where text encoding involves pooling as part of the conditioning, the attribute direction is applied to the pooled representation.

To maintain the scale of the original text embedding, we renormalize each position so that the norm of $c + \alpha a_k$ matches the norm of c . This ensures consistency in the embedding space while incorporating attribute-specific adjustments.

⁸<https://github.com/sail-sg/finetune-fair-diffusion>

⁹<https://github.com/hangligit/InterpretDiffusion>

B. Additional Results for Exploring Guidance to the Minority Region

B.1. Noisy Text Condition

Figure S1 demonstrates that introducing noise through CADS effectively reduces gender and racial biases in image generation for both SD2 and SDXL. As discussed in Section 3.2.2 regarding results for SD1.5 and SDXL, increasing the noise level in the text condition (achieved with larger s and smaller τ_1) decreases the prevalence of major attributes, leading to reduced bias in terms of both gender and race. For results using a teacher model, the impact is minimal in certain cases (*e.g.*, racial bias in SD2) or even counterproductive (*e.g.*, increased gender bias in SDXL), particularly when the initial bias in vanilla SD-generated images is less pronounced for specific professions.

Examples of images generated by vanilla SD and CADS, shown in Figure S3, highlight that CADS produces more diverse outputs and helps mitigate bias. However, CADS occasionally struggles to align the generated images with the provided text prompts, as reflected in a decrease in CLIP scores in Figure S1.

These findings indicate that perturbing text conditions with noise, as demonstrated by CADS, can be an effective strategy for reducing bias across different versions of SD. However, as noted in the main text, this approach may come at the cost of weaker alignment between generated images and their text prompts.

C. Additional Experimental Results with Debiasing

C.1. Image-Text Alignment

To evaluate the versatility of our method, we generate 100 images with text prompts where the gender is explicitly specified. The image generation process is repeated three times, and the mean and standard deviation of the attribute ratio are reported in Tables S4 and S5. Table S4 provides detailed values illustrated in Figure 8, showing the results for eight professions using SD1.5 with our method and other baselines. Table S5 presents results for SD2, SDXL, and SD3. The results show that our method successfully generates the explicitly specified attribute in the text prompts across all models with a ratio close to 1.0, ensuring versatility. Other methods, including FairDiffusion, fail to achieve a ratio of 1.0, indicating they often fail to generate attributes specified in the text prompts.

C.2. Debiasing Results

Table S6 shows the de-biasing results on eight different professions using SD2 and SDXL. The numbers indicate the proportion of the minor attribute. It is worth noting that SDXL produces a higher ratio of male teachers compared to

female teachers, which differs from SD1.5 and SD2. However, we maintain consistency by reporting the male ratio for teachers. The table demonstrates that our method effectively mitigates bias across different models.

A de-biasing technique should generalize to any prompt susceptible to a stereotype, *e.g.*, it should be able to successfully mitigate bias in images generated from various prompts such as “a photo of a lawyer in a workplace” and “a portrait of a fashion designer”. We test the generalizability with 35 different professions for 5 different prompt templates (“a painting of a/an {profession}”, “a/an {profession} working”, “a/an {profession} laughing”, “a/an {profession} in the workplace”, and “a/an {profession} digital art”), following [21]. For each template, we generate 24 images resulting in a total of 120 images per profession. In Table S7, we present the mean and standard deviation of the female ratio across five prompt templates.

C.3. Debiasing on Racial Bias

We additionally evaluate our method within racial bias. Here, we use a name list of attributes: [“White person”, “Black person”, “Asian”, “Indian”, “Latino”]. Other settings are set same as evaluating within gender bias. Table S8 shows that our method shows reduced ratio of major attribute compared to vanilla SD in all professions.

C.4. Debiasing Multiple Biases

We evaluate our method’s ability to debias both gender and race at the same time by sampling two separate weak text embeddings—one for gender attributes and another for race attributes—and then adding them to the original text embedding.

We measure performance by computing the discrepancy score, \mathcal{D} , the difference between the generated and target ratio. Formally, \mathcal{D} is defined as follows:

$$\mathcal{D} = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \left| \mathbb{E}_{\mathcal{Y}} [\mathbb{1}_{f(y)=s}] - \frac{1}{|\mathcal{S}|} \right|, \quad (2)$$

where \mathcal{S} is a set of attributes, $f(y)$ is the classified attribute of image y , and \mathcal{Y} is a set of generated images. A lower \mathcal{D} indicates more balanced generation and reduced bias.

Table S1 reports the average discrepancy score over eight professions, which we mainly use to evaluate our method, demonstrating that our method effectively addresses multiple biases at once.

C.5. Human Evaluation

We additionally conduct a human evaluation to assess two key aspects: 1) fairness and 2) image diversity and quality. A total of 40 participants took part in the survey. The following

paragraphs provide a detailed overview of the survey process and the results.

Fairness As shown in Figure S2(a), participants were shown 10 images per profession and asked to count the number of images exhibiting minor attributes (gender). The images were generated using one of three methods: vanilla SD, the main competitor FairDiff, and our method. After viewing the images, each participant was asked to count how many images displayed the minor attributes. Table S2 shows the average discrepancy scores over for this task. These human-generated results are consistent with the findings from the attribute classifier, further supporting the effectiveness of our method in reducing bias from a human perspective.

Image Quality and Diversity For evaluating image quality and diversity, we used captions from the COCO-30k dataset. Participants were shown two groups of 10 images each, one generated using our method and the other using vanilla SD. As shown in Figure S2(b), they were then asked to select the better group based on quality, including text-image alignment, and diversity, with the option to select a tie. As shown in Table S3, the most common response was “Tie” across all the results, indicating that our method successfully maintains the image quality and diversity comparable to vanilla SD, as perceived by humans.

C.6. Qualitative Results

Figure S5 shows examples of images generated with vanilla SD and our method using SD1.5, SD2, SDXL, and SD3. The examples demonstrate our method successfully generates images with minor attributes, regardless of the model version used. Images with major attributes are also produced without compromising quality.

Qualitative Comparison with FairDiffusion. Figure S4 provide additional qualitative comparisons between FairDiffusion and our method. The individuals in the images generated with FairDiffusion often exhibit physical traits of both women and men simultaneously, while those in the images generated with our method do not.

$\mathcal{D}(\downarrow)$	Vanilla SD	Ours-R	Ours-R	Ours-S
Gender	0.404	0.163	-	0.157
Race	0.177	-	0.090	0.078

Table S1. Average of difference of ratio with target within 1,000 generated images with SD1.5. “Ours-**R**” and “Ours-**S**” indicate debiasing gender and race **R**espectively and **S**imultaneously.

Fairness	SD	Fairdiff	Ours
$\mathcal{D}(\downarrow)$	0.341	0.153	0.155

Table S2. Average discrepancy score across 8 professions from human evaluation, based on 1,000 generated images using SD1.5.

Ratio	SD	Ours	Tie
Quality	0.30	0.25	0.45
Diversity	0.20	0.17	0.63

Table S3. Winning ratio, including ties, for better diversity and quality from human evaluation, based on 1,000 generated images using SD1.5.

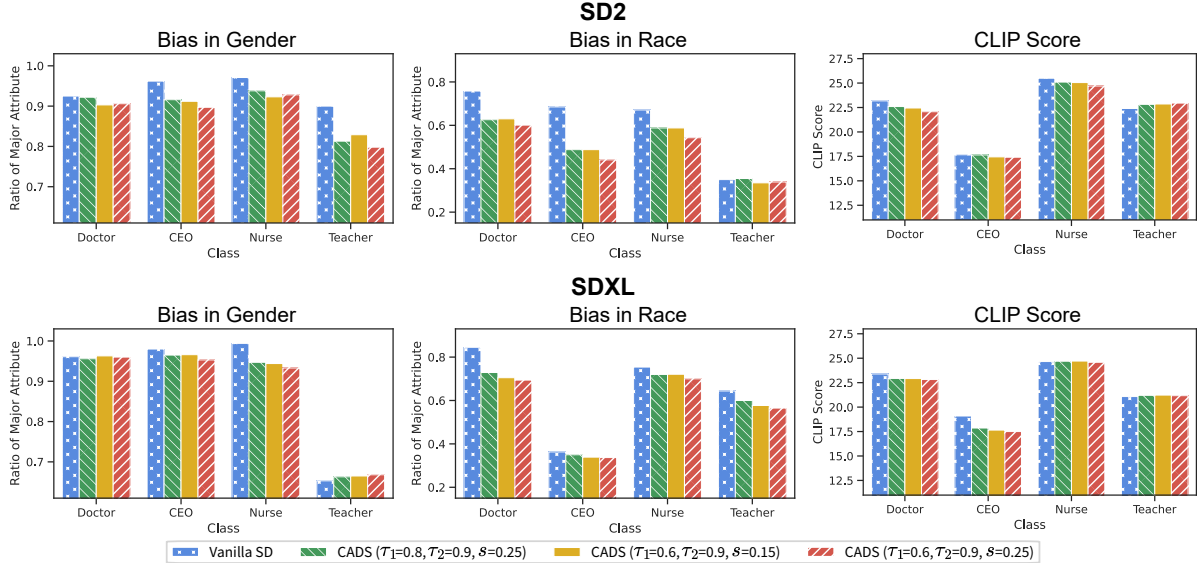


Figure S1. Change in the ratio of major attributes and CLIP score when applying CADs to SD2 and SDXL.

Profession	SD	FairDiffusion	FTDiff	SelfDisc	UCE	Ours
A photo of a female {profession}						
CEO	1.00±0.01	0.95±0.02	0.97±0.00	1.00±0.00	0.99±0.00	0.99±0.01
Doctor	1.00±0.01	0.89±0.01	1.00±0.00	1.00±0.00	0.99±0.01	0.99±0.01
Pilot	0.99±0.01	0.87±0.03	0.99±0.01	1.00±0.00	0.97±0.01	0.99±0.00
Technician	1.00±0.00	0.85±0.01	1.00±0.00	0.99±0.00	0.99±0.00	1.00±0.00
Fashion designer	1.00±0.00	0.91±0.01	0.92±0.01	1.00±0.00	0.99±0.00	0.98±0.01
Nurse	1.00±0.00	0.90±0.01	0.99±0.00	1.00±0.00	0.99±0.01	0.99±0.01
Librarian	1.00±0.00	0.93±0.01	1.00±0.00	0.99±0.00	1.00±0.00	0.99±0.01
Teacher	1.00±0.01	0.97±0.02	0.99±0.01	1.00±0.00	1.00±0.00	1.00±0.00
A photo of a male {profession}						
CEO	1.00±0.00	0.64±0.04	1.00±0.00	0.64±0.03	0.89±0.01	0.95±0.01
Doctor	1.00±0.00	0.46±0.06	0.98±0.01	0.72±0.03	0.99±0.01	0.99±0.01
Pilot	0.99±0.01	0.29±0.04	0.89±0.01	0.51±0.04	0.94±0.01	0.91±0.01
Technician	1.00±0.00	0.39±0.02	0.98±0.02	0.67±0.05	0.99±0.00	1.00±0.00
Fashion designer	0.99±0.00	0.54±0.10	0.94±0.02	0.66±0.02	0.74±0.04	0.98±0.01
Nurse	1.00±0.01	0.40±0.03	0.93±0.01	0.58±0.03	0.54±0.02	0.90±0.01
Librarian	1.00±0.00	0.55±0.04	1.00±0.00	0.74±0.02	0.91±0.04	1.00±0.00
Teacher	0.96±0.01	0.50±0.06	0.86±0.03	0.64±0.02	0.98±0.01	0.99±0.01

Table S4. Ratio of the attributes within images generated by SD1.5 using attribute-specified text prompts. The numerical values represent the attributes specified by the text prompts. Mean and standard deviation of three runs (ratio of each run is obtained using 100 images) are reported.

Profession	A photo of a female {profession}			A photo of a male {profession}		
	Vanilla SD	FairDiffusion	Ours	Vanilla SD	FairDiffusion	Ours
SD2						
CEO	1.00±0.01	0.93±0.02	0.99±0.00	1.00±0.00	0.87±0.05	1.00±0.00
Doctor	0.99±0.00	0.57±0.03	1.00±0.00	1.00±0.00	0.87±0.01	1.00±0.00
Pilot	0.99±0.01	0.79±0.05	0.99±0.00	0.99±0.01	0.56±0.01	0.99±0.00
Technician	0.97±0.01	0.58±0.06	0.98±0.00	1.00±0.00	0.78±0.03	1.00±0.00
Fashion designer	1.00±0.00	0.73±0.01	1.00±0.00	1.00±0.00	0.74±0.01	1.00±0.00
Nurse	0.99±0.01	0.64±0.03	1.00±0.00	1.00±0.00	0.74±0.04	1.00±0.00
Librarian	0.99±0.01	0.79±0.03	0.99±0.00	1.00±0.00	0.78±0.02	1.00±0.00
Teacher	0.99±0.01	0.79±0.03	1.00±0.00	1.00±0.00	0.78±0.02	1.00±0.00
SDXL						
CEO	1.00±0.00	-	1.00±0.00	1.00±0.00	-	1.00±0.00
Doctor	0.99±0.01	-	1.00±0.00	1.00±0.00	-	1.00±0.00
Pilot	0.99±0.01	-	1.00±0.00	0.99±0.01	-	0.99±0.01
Technician	1.00±0.00	-	1.00±0.00	1.00±0.00	-	1.00±0.00
Fashion designer	1.00±0.00	-	1.00±0.00	1.00±0.00	-	1.00±0.00
Nurse	1.00±0.00	-	1.00±0.00	1.00±0.00	-	1.00±0.00
Librarian	1.00±0.01	-	1.00±0.00	1.00±0.00	-	1.00±0.00
Teacher	1.00±0.00	-	1.00±0.00	1.00±0.00	-	1.00±0.00
SD3						
CEO	1.00±0.00	-	1.00±0.00	1.00±0.00	-	1.00±0.00
Doctor	1.00±0.00	-	1.00±0.00	1.00±0.00	-	1.00±0.00
Pilot	1.00±0.00	-	1.00±0.00	1.00±0.00	-	1.00±0.00
Technician	1.00±0.00	-	1.00±0.00	1.00±0.00	-	1.00±0.00
Fashion designer	1.00±0.00	-	1.00±0.00	1.00±0.00	-	1.00±0.00
Nurse	1.00±0.00	-	1.00±0.00	1.00±0.00	-	1.00±0.00
Librarian	1.00±0.00	-	1.00±0.00	1.00±0.00	-	1.00±0.00
Teacher	1.00±0.00	-	1.00±0.00	1.00±0.00	-	1.00±0.00

Table S5. Ratio of the attributes within images generated by SD2, SDXL, and SD3 using attribute-specified text prompts. The numerical values represent the attributes specified by the text prompts.

Profession	SD2			SDXL		SD3	
	Vanilla SD	FairDiffusion	Ours	Vanilla SD	Ours	Vanilla SD	Ours
CEO	0.038	0.176	0.147	0.020	0.201	0.000	0.262
Doctor	0.075	0.299	0.258	0.039	0.250	0.126	0.286
Pilot	0.065	0.504	0.159	0.064	0.286	0.017	0.295
Technician	0.007	0.333	0.101	0.002	0.011	0.000	0.029
Fashion designer	0.062	0.088	0.152	0.270	0.384	0.002	0.008
Librarian	0.071	0.387	0.142	0.345	0.448	0.000	0.097
Teacher	0.100	0.486	0.128	0.653	0.541	0.008	0.228
Nurse	0.029	0.338	0.021	0.006	0.025	0.000	0.001

Table S6. Ratio of minor attribute within 1000 generated images with SD2, SDXL, and SD3.

Profession	Vanilla SD	FairDiffusion	UCE	FTDiff	SelfDisc	Ours
Attendant	0.10±0.04	0.48±0.09	0.49±0.05	0.25±0.12	0.39±0.19	0.13±0.09
Cashier	0.61±0.21	0.65±0.07	0.49±0.27	0.50±0.07	0.16±0.15	0.66±0.26
Teacher	0.67±0.25	0.69±0.14	0.49±0.32	0.67±0.12	0.10±0.09	0.63±0.32
Nurse	0.98±0.02	0.84±0.07	0.03±0.05	0.81±0.10	0.82±0.18	0.98±0.02
Assistant	0.45±0.18	0.63±0.03	0.49±0.36	0.42±0.15	0.04±0.07	0.46±0.14
Secretary	0.93±0.15	0.86±0.06	0.08±0.07	0.60±0.17	0.16±0.11	0.91±0.07
Cleaner	0.34±0.18	0.45±0.13	0.68±0.18	0.47±0.22	0.46±0.11	0.25±0.16
Receptionist	0.93±0.07	0.86±0.09	0.33±0.23	0.73±0.18	0.02±0.04	0.97±0.05
Clerk	0.40±0.25	0.61±0.12	0.31±0.16	0.43±0.09	0.63±0.06	0.38±0.19
Counselor	0.48±0.14	0.54±0.05	0.34±0.21	0.46±0.17	0.37±0.19	0.50±0.15
Designer	0.43±0.32	0.65±0.13	0.42±0.32	0.53±0.22	0.03±0.04	0.52±0.37
Hairdresser	0.64±0.21	0.55±0.17	0.36±0.17	0.63±0.26	0.34±0.08	0.63±0.21
Writer	0.39±0.12	0.58±0.07	0.55±0.28	0.58±0.23	0.30±0.15	0.44±0.19
Housekeeper	0.98±0.04	0.73±0.17	0.07±0.05	0.95±0.03	0.33±0.19	0.96±0.05
Baker	0.45±0.19	0.61±0.11	0.36±0.14	0.65±0.14	0.40±0.08	0.37±0.32
Librarian	0.83±0.22	0.80±0.19	0.08±0.04	0.62±0.19	0.75±0.09	0.85±0.15
Tailor	0.24±0.25	0.53±0.22	0.61±0.26	0.33±0.27	0.18±0.10	0.23±0.24
Driver	0.06±0.09	0.43±0.09	0.79±0.11	0.30±0.17	0.20±0.07	0.06±0.07
Supervisor	0.18±0.14	0.49±0.12	0.61±0.18	0.15±0.16	0.63±0.15	0.13±0.10
Janitor	0.03±0.04	0.19±0.09	0.54±0.22	0.07±0.09	0.02±0.02	0.03±0.03
Cook	0.32±0.18	0.48±0.19	0.33±0.32	0.52±0.08	0.31±0.16	0.38±0.20
Laborer	0.03±0.04	0.34±0.10	0.82±0.23	0.15±0.17	0.48±0.14	0.03±0.05
Construction worker	0.00±0.00	0.21±0.16	0.76±0.28	0.05±0.07	0.06±0.06	0.00±0.00
Developer	0.14±0.15	0.38±0.05	0.65±0.18	0.25±0.25	0.39±0.22	0.07±0.08
Carpenter	0.02±0.04	0.39±0.17	0.70±0.07	0.07±0.13	0.44±0.11	0.03±0.05
Manager	0.16±0.16	0.45±0.13	0.72±0.22	0.27±0.17	0.53±0.10	0.24±0.26
Lawyer	0.13±0.14	0.43±0.08	0.81±0.14	0.24±0.08	0.64±0.22	0.22±0.18
Farmer	0.02±0.02	0.32±0.09	0.97±0.03	0.14±0.07	0.41±0.15	0.02±0.02
Salesperson	0.09±0.07	0.44±0.15	0.68±0.13	0.16±0.06	0.44±0.14	0.13±0.13
Physician	0.10±0.11	0.48±0.18	0.76±0.20	0.19±0.07	0.70±0.10	0.15±0.12
Guard	0.03±0.02	0.30±0.11	0.84±0.10	0.09±0.05	0.64±0.09	0.03±0.03
Analyst	0.08±0.08	0.49±0.05	0.71±0.22	0.17±0.11	0.42±0.14	0.09±0.08
Mechanic	0.03±0.03	0.43±0.12	0.67±0.15	0.08±0.10	0.71±0.11	0.03±0.03
Sheriff	0.03±0.02	0.49±0.09	0.64±0.18	0.08±0.10	0.92±0.04	0.04±0.07
CEO	0.03±0.05	0.31±0.13	0.89±0.06	0.13±0.08	0.63±0.17	0.06±0.09

Table S7. Ratio of female attribute across five text prompts. Images are generated using SD1.5.


Profession	Vanilla SD					Ours				
	White	Black	Indian	Asian	Latino	White	Black	Indian	Asian	Latino
CEO	0.091	0.165	0.015	0.228	0.501	0.025	0.313	0.087	0.330	0.245
Doctor	0.660	0.049	0.014	0.031	0.246	0.194	0.268	0.126	0.186	0.226
Pilot	0.389	0.028	0.111	0.259	0.213	0.110	0.203	0.234	0.301	0.152
Technician	0.488	0.060	0.004	0.016	0.432	0.172	0.271	0.052	0.177	0.328
Fashion designer	0.222	0.075	0.030	0.549	0.124	0.095	0.236	0.105	0.473	0.091
Nurse	0.716	0.132	0.010	0.099	0.043	0.269	0.282	0.103	0.289	0.057
Librarian	0.886	0.010	0.000	0.045	0.059	0.516	0.173	0.014	0.203	0.094
Teacher	0.432	0.125	0.005	0.120	0.318	0.128	0.245	0.064	0.333	0.230

Table S8. Ratio of each attribute within 1000 generated images with SD1.5. The bold indicates the highest ratio among attributes.

(a)


Part 1 (1/8)

Part 1 is about counting images of certain gender based on the 10 images generated by each model. Please note that the gender to be counted may differ for each question.



Q. How many images do you think feature **female** among the images above?

☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10



Q. How many images do you think feature **female** among the images above?

☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10



Q. How many images do you think feature **female** among the images above?

☐ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ 10

Next

(b)

Part 2 (1/10)

Look at the 10 images generated by two different models using the prompt below and answer each question.

A desk with various electronics and a computer with a screensaver


Q1. Which group of images do you think is more **diverse**?
(By 'diverse,' we mean that the images have different styles, colors, compositions, etc.)

☐ A ☐ B ☐ Tie


Q2. Which group of images do you think has better **quality**?
(By 'quality,' we mean that the images are naturally generated corresponding to the prompt.)

☐ A ☐ B ☐ Tie

A



B



Next

Figure S2. Example of a survey page for human evaluation focusing on (a) fairness and (b) image diversity and quality.



Figure S3. Examples of images generated using vanilla SD and CADs with SD2 and SDXL.

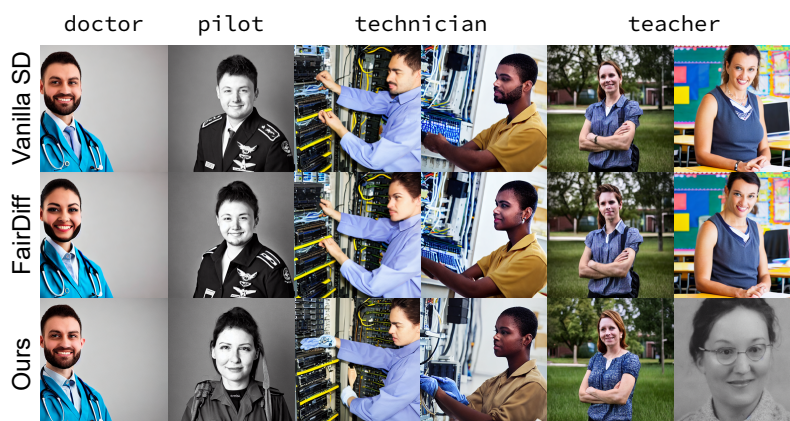


Figure S4. Examples of generated images with vanilla SD, FairDiff, and our method, using SD1.5.



Figure S5. Examples of generated images with vanilla SD and our method.