

Supplementary Materials: Mitigating Social Biases in Text-to-Image Diffusion Models via Linguistic-Aligned Attention Guidance

Anonymous Authors

In addition to the experiments and discussions presented in the paper, the appendix provides more experimental details and results, which further demonstrate the effectiveness of our method.

1 EXPERIMENT DETAILS

For scenarios of daily activities and style prompts in Fig.7 of the paper, we employ prompts “two persons standing next to a vase of flowers on a table” and “concept art of elite scientist by jama jurabae, emperor secret society, cinematic shot, trending on artstation, high quality, brush stroke”. Moreover, we provide additional qualitative comparisons about occupations and other prompt templates in the appendix.

2 ADDITIONAL QUALITATIVE COMPARISONS

Occupations. We provide qualitative comparisons involving multiple individuals within an image for additional occupations in Fig. 1 and Fig. 2. Different social groups are highlighted by the same color box as in the paper. As demonstrated in the results, SD-EI, FairDiffusion, and UCE tend to produce homogenized social groups within an image. Fine-tune presents similar facial characteristics across the same social groups, particularly for Indians, appearing as the same person. Besides, the generated results of Fine-tune are less consistent with the original results in terms of structural and semantic information. In contrast, our method generates more diverse results for different social groups while perfectly preserving the original structural and semantic consistency.

Other Senarios. To further illustrate the effectiveness of our method, we employ prompts for additional descriptors involving other objects, such as “two persons wearing hats” and “two persons covered by red headscarf” (Fig. 3), as well as descriptors for different occupations: “a photo of the faces of two artists smiling”, and “a photo of the faces of two teachers reading” (Fig. 4). All the results are debiased for intersections of gender and race. Across these scenarios, our method demonstrates more diversity in social groups within the images while maintaining the original structural and semantic consistency.

3 QUANTITATIVE COMPARISONS

To present an optional evaluation of diversity and fairness within images, we employ the Bias-P* metric, which considers individuals within images incapable of fitting all groups. When the individual counts N fall below the group counts n_a (denoted as $n_a(-)$), we evaluate the diversity by Simpson’s diversity index [1]. When the individual counts exceed the group counts (denoted as $n_a(+)$), we evaluate the fairness across different social groups. The formula is as follows:

	Method	Bias-P* (↓)			
		Gender (2+)	Race (4-)	Race (4+)	G. × R. (8-)
Gender	Stable Diffusion	.316 ± .08	.784 ± .09	.319 ± .16	.493 ± .13
	SD-EI	.269 ± .02	.738 ± .12	.331 ± .05	<u>.400 ± .07</u>
	FairDiffusion	.358 ± .04	.686 ± .11	.290 ± .15	.480 ± .10
	UCE	.452 ± .05	.740 ± .22	.357 ± .17	.665 ± .21
	Fine-tune	.251 ± .03	.910 ± .04	.336 ± .02	.473 ± .06
Race	Ours	<u>.255 ± .03</u>	<u>.719 ± .06</u>	<u>.320 ± .16</u>	.385 ± .07
	SD-EI	.288 ± .09	.778 ± .07	<u>.268 ± .14</u>	.444 ± .11
	FairDiffusion	.283 ± .07	.768 ± .08	.312 ± .16	.439 ± .12
	UCE	<u>.278 ± .07</u>	<u>.617 ± .28</u>	.303 ± .15	<u>.346 ± .18</u>
	Ours	.271 ± .07	.454 ± .05	.216 ± .11	.257 ± .04
G. × R.	SD-EI	.251 ± .06	.777 ± .06	.352 ± .02	.408 ± .07
	FairDiffusion	.315 ± .05	.741 ± .07	.280 ± .14	.472 ± .09
	UCE	.325 ± .11	.816 ± .07	.287 ± .17	.534 ± .16
	Fine-tune	<u>.251 ± .04</u>	<u>.499 ± .07</u>	<u>.204 ± .18</u>	<u>.240 ± .03</u>
	Ours	.236 ± .03	.389 ± .05	.191 ± .10	.167 ± .03

Table 1: Quantitative comparisons for the Bias-P* metric. The best results are highlighted in **bold** and the second to best is highlighted by underline.

$$Bias - P^* = \begin{cases} \frac{\sum_i N_i(N_i - 1)}{N(N - 1)}, & N < n_a, \\ \sqrt{\frac{1}{n_a} \sum_a (freq_a^p - \frac{1}{n_a})^2}, & N \geq n_a, \end{cases} \quad (1)$$

where i represents social groups and N denotes the total number of individuals within an image, N_i is the number of individuals in social group i . a represents all social groups, n_a denotes the number of them, and $freq_a^p$ indicates the frequency of attribute a within one image. Specifically, the number of n_a is 2 for gender, 4 for race, and 8 for their intersections.

For quantitative evaluation, we sample images containing individuals between two and eight ($2 \leq N < 8$). Each method is evaluated across five occupations, with 100 images per occupation. The results are shown in Table 1. As demonstrated in the results, our method outperforms other methods by a large margin considering different biases, indicating its effectiveness in mitigating biases with multiple individuals in the images.

REFERENCES

- [1] Edward H Simpson. 1949. Measurement of diversity. *nature* 163, 4148 (1949), 688–688.



Figure 1: Qualitative comparison of different methods for “Artists”. Different social groups are highlighted with colored boxes: “White Female”, “White Male”, “Black Female”, “Black Male”, “Indian Female”, “Indian Male”, “Asian Female”, “Asian Male”.



Figure 2: Qualitative comparison of different methods for “Scientists”.



Figure 3: Qualitative comparison for scenarios involving objects: “two persons wearing hats” (left) and “two persons covered by red headscarf” (right).

