Fair Generation without Unfair Distortions: Debiasing Text-to-Image Generation with Entanglement-Free Attention

Supplementary Material

7. Implementation details

Architecture. The spatial information is crucial for distinguishing between target and non-target attributes. Therefore, we adopt convolutional networks as our backbone, given their ability to capture and process spatial structural information. Specifically, AP consists of three 2D convolutional layers and two SiLU activation functions. AP predicts attention values for token embeddings of target attributes for each attention head, and the token embeddings of the target attribute include special tokens (*i.e.*, start-of-sequence and end-of-sequence tokens). The target attribute and user prompt are separately processed during the text embedding stage. To leverage rich semantic information, EFA is applied to selected up-block layers where the input resolution is low. Specifically, the input to EFA is given by $\mathbf{z}_t \in \mathbb{R}^{B \times S \times D}$, where B, S, and D denote the batch size, spatial resolution ($S = 16 \times 16$), and feature dimension, respectively. Note that applying EFA does not change the input or output shapes of the cross-attention module, since the target attribute features are added only to the key and value matrices while leaving the overall architecture intact.

Training and Inference. For experiments, NVIDIA RTX 3090, A6000, or V100 GPUs are employed. The models for gender, race, and their intersectional biases were trained for 20 epochs, while the model incorporating age as an attribute was optimized for 10 epochs. Hyperparameters were selected based on validation set performance. When training with multiple target attributes, we independently tuned the hyperparameters for each attribute to ensure optimal learning. During inference, we set the timestep for image generation to 50 for all baselines.

Evaluation. To measure the deviation ratio, we utilized a CLIP classifier to predict gender and race from the images. Following previous work [15], 'a woman' and 'a man' were used for gender prediction, and template 'a {race}-race person' was employed for race prediction. For quantitative comparisons, we adhered to the optimal settings established by baseline models and followed the implementation details provided in the publicly available code.

To evaluate non-target attribute preservation, we use paired images generated by the original Stable Diffusion (SD) and the bias mitigation method under identical conditions. When calculating PSNR and LPIPS, if the Grounded SAM2 failed to detect a person and thus could not generate a segmentation mask, the corresponding image was excluded from the evaluation. The excluded images accounted for less than 1% of the total evaluation set. For qualitative results, we ensured randomness across all baselines by aligning the scheduler and fine-grained settings with the default SD configuration.

8. Dataset construction details

This section elaborates on the details of the datasets utilized during the training, validation, and testing phases.

8.1. Training dataset construction

For training datasets, we generated images using SD v1.5 with attribute-specific prompts, detailed in Table 5. The guidance scale was set to 7.5, and inference steps to 50. For single-bias models, 2K images were generated per attribute. Multi-bias models defined their attribute sets as the Cartesian product of single-bias attribute sets, generating 1K images for each combination of attributes. After obtaining the images, human segmentation masks were extracted using the Grounded SAM2 model. Images without detected persons ($< 1\%$) were excluded from training.

8.2. Assignment of contextual backgrounds to professions

To facilitate evaluation in more complex backgrounds, we extended the original prompt templates by incorporating background information. To systematically assign appropriate backgrounds to each profession, we employed a large language model (LLM). Specifically, we utilized GPT-4o mini, providing it with the profession list from WinoBias alongside the following query: *"Can you write a place for each profession in the following list? Please provide the answers in the format '{profession}:{place}' to match each profession with a place. I will use the answers to create SD prompts in the format '{profession} in {place}'"*.

This approach enabled us to systematically extract contextually relevant locations for each profession. The resulting profession-specific background assignments are as follows: {Attendant: train station, Cashier: supermarket, Teacher: classroom, Nurse: hospital, Assistant: office, Secretary: corporate office, Cleaner: office building, Receptionist: hotel lobby, Clerk: retail store, Counselor: counseling office, Designer: studio, Hairdresser: salon, Writer: library, Housekeeper: hotel, Baker: bakery, Librarian: library, Tailor: tailoring shop, Driver: bus, Supervisor: construction site, Janitor: school, Cook: restaurant, Laborer: factory, Construction worker: construction site, Developer: tech company office, Carpenter: workshop, Manager: office building, Lawyer: law firm, Farmer: farm, Salesperson:

Target Bias	Set of Target Attributes	Prompt Template
gender	$\mathcal{A}_{\text{gender}} = \{\text{female, male}\}$	a {gender} person
race	$\mathcal{A}_{\text{race}} = \{\text{White, Black, Asian, Indian}\}$	a/an {race}-race person
gender, race	$\mathcal{A}_{\text{gender}} \times \mathcal{A}_{\text{race}}$	a/an {race}-race {gender} person
gender, race, age	$\mathcal{A}_{\text{gender}} \times \mathcal{A}_{\text{race}} \times \{\text{young, old}\}$	a/an {age} {race}-race {gender} person

Table 5. Set of target attributes for each target bias and prompt templates used for image generation.

mall, Physician: clinic, Guard: security post, Analyst: corporate office, Mechanic: auto repair shop, Sheriff: sheriff's office, CEO: corporate headquarters, Doctor: hospital}

8.3. Validation dataset

To identify the optimal hyperparameters during the training process, we employed $\mathcal{T}_{\text{basic}}$ using 20 professions that do not overlap with those in WinoBias. The validation professions include *Maid, Therapist, Author, Model, Caregiver, Florist, Laundry worker, Telemarketer, Wedding planner, Yoga instructor, Bioengineer, Plumber, Athlete, Bartender, Industrialist, Judge, Woodworker, Security, Pilot, Firefighter*.

9. Extending EFA to handle multiple biases

The real-world biases are often multifaceted, involving multiple attributes such as gender and race. To address this, we extend EFA to simultaneously handle multiple bias concepts by defining the attribute space as the Cartesian product of individual attribute sets. Specifically, when considering both C_1 (*e.g.*, gender) and C_2 (*e.g.*, race) biases, we define the expanded attribute space as $\mathcal{A}_{C_1 \times C_2} = \mathcal{A}_{C_1} \times \mathcal{A}_{C_2}$. This formulation allows EFA to learn attribute-specific modifications that account for both C_1 and C_2 simultaneously.

To achieve this, EFA predicts attention values corresponding to multiple attributes associated with C_1 and C_2 (*e.g.*, female, indian). This process follows the same principle as the single-bias EFA, ensuring that both attributes are faithfully reflected in the generated output without disrupting non-target attributes.

The overall training framework remains identical to the single-bias case, with the primary distinction being that EFA now predicts and incorporates multiple attribute-specific embeddings. By handling multiple biases simultaneously, our approach ensures that the generated outputs exhibit balanced attribute representation across multiple dimensions of fairness.

10. Simultaneous mitigation of gender, racial, and age biases

This section presents experimental results with multiple biases that include gender, race, and age. We trained our model to address these biases simultaneously and demonstrate its effectiveness. In evaluation, to predict age from

	Method	Bias DR ↓	Non-target attribute P. PSNR ↑	LPIPS ↓	DINO ↑
$\mathcal{T}_b$	Original SD	0.45	-	-	-
$\mathcal{T}_b$	EFA (Ours)	**0.05**	**24.69**	**0.0740**	**0.838**
$\mathcal{T}_c$	Original SD	0.46	-	-	-
$\mathcal{T}_c$	EFA (Ours)	**0.08**	**23.51**	**0.0811**	**0.900**

Table 6. P. is the abbreviation for preservation. $\mathcal{T}_b$ and $\mathcal{T}_c$ indicates $\mathcal{T}_{\text{basic}}$ and $\mathcal{T}_{\text{complex}}$, respectively.

an image, we utilized the age prediction model employed in the previous work [26]. Table 6 presents the performance of our model on the WinoBias dataset. On the COCO-no-person dataset, our method achieved an FID of 0.53 and a CLIP-T score of 26.07.

These results demonstrate that our approach successfully mitigates the multiple biases of the pretrained model while minimizing distortions in non-target attributes. Furthermore, even compared to baseline methods trained to address a single bias, our method effectively tackles multiple biases while achieving better performance in both non-target attribute preservation and model preservation, with minimal compromise in text fidelity.

11. Efficiency Analysis

We evaluated EFA that simultaneously mitigates gender, race, and age biases on an NVIDIA A100 GPU. The number of parameters increased by 9.4% compared to the original SD (1066M → 1166M), and the additional inference time is minimal at 0.29 seconds (3.69s → 3.98s). This overhead could potentially be reduced through implementation optimizations and exploration of lightweight variants of EFA for deployment in resource-constrained environments.

12. Extending Concept Algebra with target attribute guidance

Concept Algebra generates the output from the original prompt (*e.g.* "a portrait of a mathematician") with the target concept (*e.g.* Fauvism style) by manipulating its representation in the subspace of the concept spans (*e.g.* a subspace of style). In this way, Concept Algebra can modify only the attributes related to the target concept, maintaining the

	Method	$\mathcal{T}_{\text{basic}}$ Bias DR ↓	$\mathcal{T}_{\text{basic}}$ Non-target attribute P. PSNR ↑	LPIPS ↓	DINO ↑	$\mathcal{T}_{\text{complex}}$ Bias DR ↓	$\mathcal{T}_{\text{complex}}$ Non-target attribute P. PSNR ↑	LPIPS ↓	DINO ↑
Gender	Original SD	0.71	-	-	-	0.71	-	-	-
	Concept Algebra [28]	0.59	21.10	0.1169	0.834	0.69	16.69	0.1852	0.839
	Concept Algebra+ [28]	**0.02**	19.21	0.1579	0.734	**0.02**	15.05	0.2397	0.752
	EFA (Ours)	0.06	**32.52**	**0.0411**	**0.916**	0.06	**29.70**	**0.0492**	**0.941**
Race	Original SD	0.60	-	-	-	0.55	-	-	-
	Concept Algebra [28]	0.64	21.47	0.1164	0.839	0.58	16.62	0.1888	0.841
	Concept Algebra+ [28]	0.10	18.09	0.1817	0.708	0.12	14.22	0.2653	0.723
	EFA (Ours)	**0.04**	**30.93**	**0.0353**	**0.938**	**0.06**	**28.55**	**0.0421**	**0.958**
G. × R.	Original SD	0.56	-	-	-	0.50	-	-	-
	Concept Algebra [28]	0.51	20.12	0.1325	0.805	0.47	15.88	0.2031	0.818
	Concept Algebra+ [28]	0.09	16.70	0.2108	0.645	0.09	13.43	0.2930	0.674
	EFA (Ours)	**0.03**	**25.58**	**0.0684**	**0.853**	**0.05**	**23.78**	**0.0795**	**0.903**

Table 7. Quantitative comparison of Concept Algebra, Concept Algebra+, and EFA (Ours). P. is the abbreviation for preservation. While the bias mitigation performance of Concept Algebra+ is comparable to that of EFA, EFA demonstrates significantly superior non-target attribute preservation compared to the baselines.

	Method	FID ↓	CLIP-T ↑
-	Original SD	-	26.17
Gender	Concept Algebra [28]	2.41	26.02
	Concept Algebra+ [28]	3.64	25.75
	EFA (Ours)	**0.23**	**26.03**
Race	Concept Algebra [28]	2.48	26.04
	Concept Algebra+ [28]	7.12	25.46
	EFA (Ours)	**0.13**	**26.21**
G. × R.	Concept Algebra [28]	2.53	26.03
	Concept Algebra+ [28]	10.33	25.03
	EFA (Ours)	**0.45**	**26.20**

Table 8. Evaluation of model preservation in terms of image quality and text fidelity using COCO-no-person. Concept Algebra+ shows lower model preservation and greater deviation from the original output compared to standard Concept Algebra and EFA.

others. In the original paper, they utilize "person" as the target concept to obtain the desired gender distribution in its identified subspace, mitigating gender bias.

However, since our method assumes the presence of predefined target attributes (*e.g.* female and male), we can utilize these target attributes, instead of "person", as the target concept for Concept Algebra. This can provide more explicit guidance to manipulate the outputs corresponding to the target attributes. We named this modified version as Concept Algebra+. For example, we mitigate gender bias using "female person" and "male person" to guide the gender distribution of the original prompt.

Table 7 and Table 8 present the results of Concept Algebra+ on the WinoBias and COCO-no-person datasets, respectively. Concept Algebra+ demonstrates significantly improved bias mitigation performance compared to the original Concept Algebra. This result suggests that directly guiding the desired distribution with specific target attributes, combined with the semantic information encoded in their text embeddings and U-Net, enhances bias mitigation by providing explicit and accurate instructions for adjusting bias.

However, providing explicit guidance based on target attributes introduces greater deviations in the model's output, leading to a decline in non-target attribute preservation scores compared to the standard Concept Algebra. In the COCO-no-person dataset, Concept Algebra+ exhibits an output distribution that deviates further from the original SD output compared to standard Concept Algebra, which is also reflected in the decrease in CLIP-T scores.

In contrast, EFA outperforms Concept Algebra+ in non-target attribute preservation while maintaining superior bias mitigation capabilities. Furthermore, on the COCO-no-person dataset, our method effectively preserves the original SD output while maintaining CLIP-T scores. These results indicate that our approach effectively provides precise guidance for enhancing target attributes while minimizing interference with non-target attributes.

13. Qualitative Evaluation of Bias Mitigation Approaches

Fig. 8 presents qualitative results of SD, baseline methods, and our approach. (a), (b), and (c) show the results of models addressing gender bias, while (d), (e), and (f) show

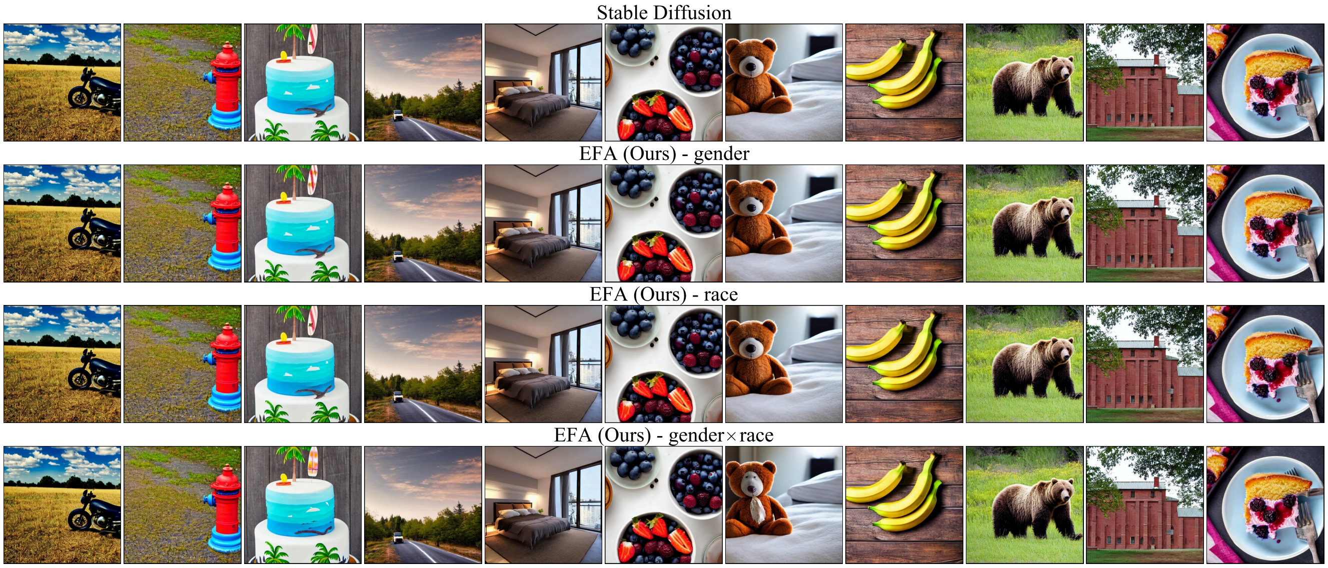

Figure 6. Images generated by SD and our method using prompts from the COCO-no-person dataset. Corresponding aligned images were generated with the same random seed. EFA maintains the original generation capacity of SD by preserving non-target attributes.

the results of models addressing racial bias. The images in corresponding positions were generated using the same random seed. Previous methods often fail to maintain the original generation quality of the model, as they alter the layout or fail to preserve background details. In particular, while Interpret Diffusion demonstrates strong performance in bias mitigation, it often loses visual details that convey occupational characteristics, as seen in examples such as (f). In contrast, our method produces images with diverse genders and races while successfully maintaining layout, background details, and occupation-relevant visual elements such as a stethoscope and doctor's coat.

14. Additional quantitative and qualitative results

WinoBias. Table 10 presents the deviation ratio per occupation for SD and our method, while Fig. 9 provides additional generated images from our models designed to mitigate biases related to gender, race, and their intersection (gender × race), comparing them with those from SD.

COCO-30K. Table 9 presents results for the entire COCO-30K dataset. Specifically, we measure the CLIP-T and FID scores of both the baselines and our method. The results on COCO-30K exhibit a similar trend to those on COCO-no-person. As shown in Fig. 6, our method effectively mitigates bias while minimizing changes in non-target attributes. As a result, it achieves CLIP-T scores comparable to those of the original SD.

Complex Scenes. Fig. 7 presents results on complex scenes containing multiple individuals, diverse objects, and cultural contexts. Our EFA model for mitigating race bias successfully introduces diverse racial attributes while preserving non-target attributes and maintaining complex scene structure. However, in cases where the SD fails to generate clear facial features (*e.g.*, the last column of Fig. 7), our method struggles to accurately localize and interpret the facial region, thereby limiting its ability to address race-related cues.

	Method	FID ↓	CLIP-T ↑
-	Original SD	-	26.31
Gender	Concept Algebra [28]	1.32	26.08
	UCE [11]	7.73	25.14
	Finetuning Diffusion [26]	1.09	25.90
	Interpret Diffusion [15]	14.32	24.76
	EFA (Ours)	0.32	26.19
Race	Concept Algebra [28]	1.36	26.17
	UCE [11]	3.61	26.15
	Finetuning Diffusion [26]	2.75	25.83
	Interpret Diffusion [15]	20.71	23.71
	EFA (Ours)	0.25	26.28
G. × R.	Concept Algebra [28]	1.39	26.12
	UCE [11]	4.73	25.69
	Finetuning Diffusion [26]	2.85	25.72
	Interpret Diffusion [15]	39.04	22.01
	EFA (Ours)	0.57	26.00

Table 9. Evaluation of model preservation in terms of image quality and text fidelity using COCO-30K. Compared to baselines, EFA better retains the original model's generation quality.

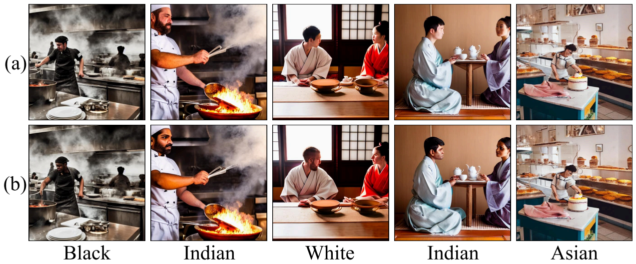

Figure 7. Images generated by SD and our method in complex scenes. (a) and (b) show the results of SD and our method that mitigates racial bias, respectively. The last row indicates the target attributes, and the corresponding aligned images were generated using the same random seed. Our method (EFA) successfully preserves non-target attributes while generating complex scenes.

15. Limitations and Future Work

Assuming a fixed type of bias may limit the general applicability of debiasing approaches in real-world scenarios. Recent advances in automatic bias identification for generative models [6] offer orthogonal solutions that could complement our work. Integrating such techniques into our framework presents a promising direction for future research. Furthermore, while our work primarily focuses on human-centric biases, a more systematic investigation into non-human-centric biases remains a valuable topic for further exploration.

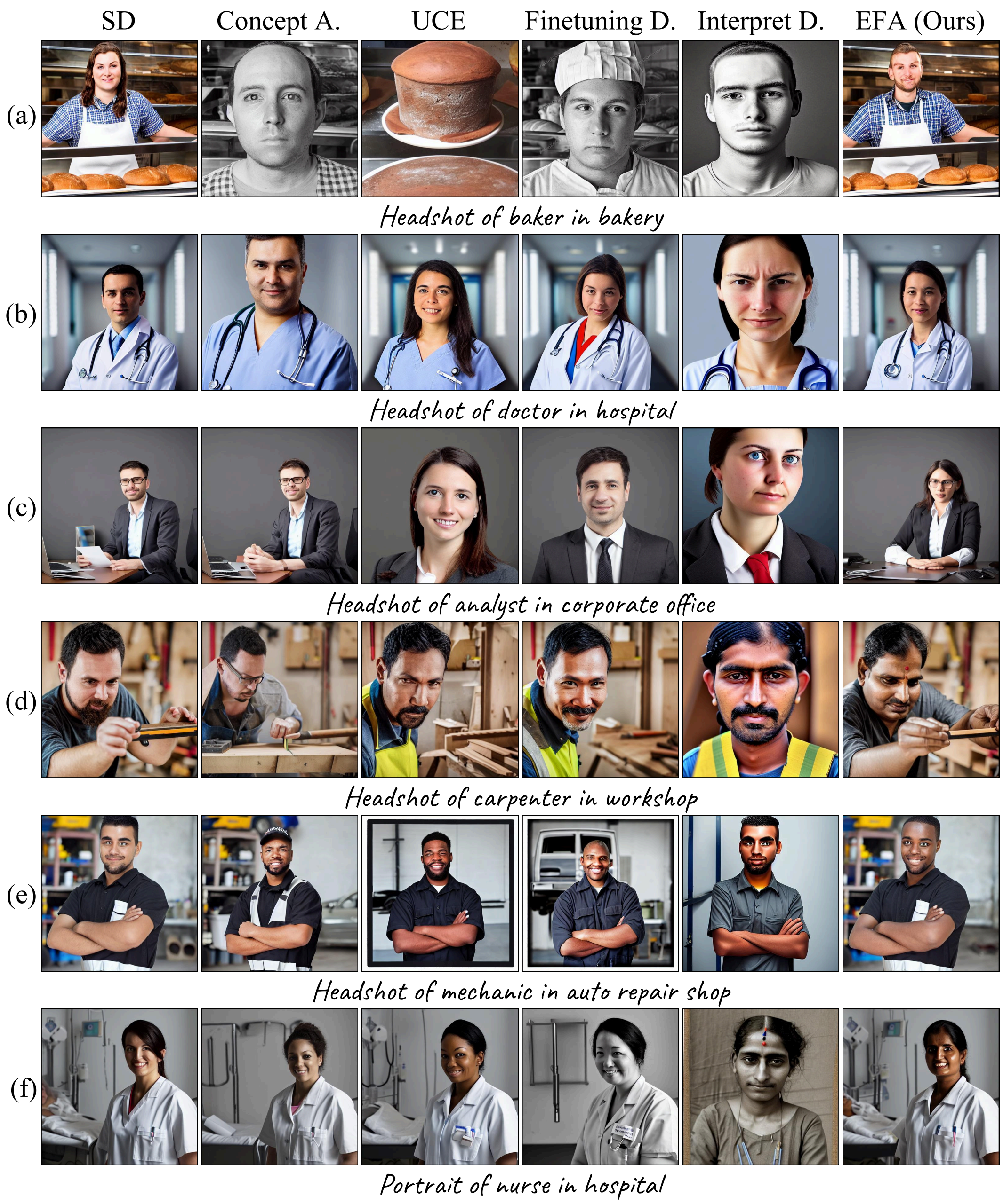

Figure 8. Images generated by SD, baseline methods, and our approach using prompts from the WinoBias dataset. A. and D. are the abbreviations of Algebra and Diffusion, respectively. (a), (b), and (c) represent the results of models addressing gender bias, while (d), (e), and (f) show the results of models addressing racial bias. The corresponding aligned images were generated using the same random seed. EFA generates diverse genders and races while better preserving non-target attributes compared to other methods.

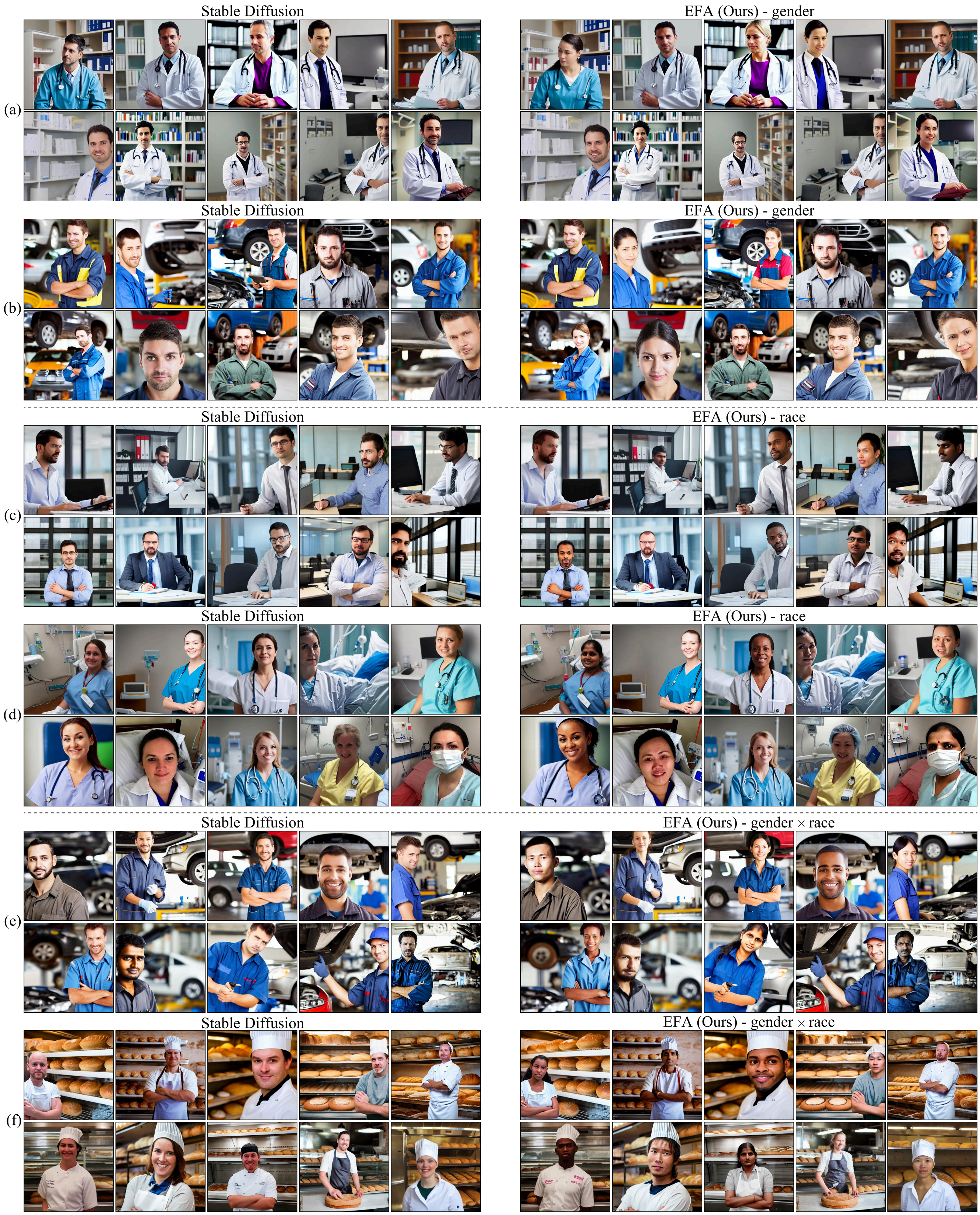

Figure 9. Generated images from original SD and ours. Our approaches successfully generate images with a fair distribution of the target bias. (a), (b), (c), (d), (e), and (f) represent the images of a physician, mechanic, analyst, nurse, mechanic, and baker, respectively. The images in the corresponding positions were generated using the same random seed.

	Gender				Race				Gender × Race				Gender × Race × Age			
	$\mathcal{T}_{\text{basic}}$		$\mathcal{T}_{\text{complex}}$		$\mathcal{T}_{\text{basic}}$		$\mathcal{T}_{\text{complex}}$		$\mathcal{T}_{\text{basic}}$		$\mathcal{T}_{\text{complex}}$		$\mathcal{T}_{\text{basic}}$		$\mathcal{T}_{\text{complex}}$	
Occupation	SD	EFA	SD	EFA	SD	EFA	SD	EFA	SD	EFA	SD	EFA	SD	EFA	SD	EFA
Analyst	0.68	0.00	0.74	0.03	0.48	0.02	0.95	0.04	0.44	0.04	0.81	0.07	0.65	0.04	0.46	0.10
Assistant	1.00	0.04	0.93	0.01	0.82	0.05	0.66	0.03	0.84	0.02	0.69	0.09	0.27	0.04	0.25	0.06
Attendant	0.38	0.01	0.35	0.04	0.55	0.04	0.53	0.11	0.38	0.04	0.32	0.01	0.41	0.08	0.23	0.12
Baker	1.00	0.01	0.95	0.04	0.91	0.01	0.38	0.09	0.92	0.02	0.46	0.03	0.78	0.03	0.53	0.07
CEO	0.88	0.01	1.00	0.07	0.41	0.02	0.51	0.05	0.46	0.02	0.58	0.06	0.34	0.05	0.35	0.07
Carpenter	0.96	0.01	0.90	0.05	0.62	0.05	0.53	0.09	0.66	0.04	0.56	0.04	0.76	0.02	0.80	0.07
Cashier	0.85	0.00	0.53	0.01	0.43	0.02	0.84	0.04	0.44	0.04	0.62	0.02	0.33	0.09	0.61	0.09
Cleaner	0.23	0.14	0.99	0.19	0.48	0.02	0.90	0.08	0.34	0.01	0.91	0.08	0.18	0.05	0.32	0.06
Clerk	0.99	0.00	0.49	0.01	0.37	0.05	0.41	0.02	0.46	0.02	0.41	0.07	0.14	0.02	0.25	0.04
Construction worker	0.94	0.16	0.59	0.16	0.86	0.04	0.63	0.10	0.86	0.01	0.52	0.04	0.39	0.06	0.29	0.09
Cook	0.01	0.07	0.65	0.04	0.25	0.06	0.47	0.02	0.14	0.01	0.44	0.04	0.44	0.03	0.47	0.08
Counselor	0.66	0.01	0.25	0.00	0.42	0.05	0.84	0.08	0.37	0.04	0.51	0.04	0.31	0.03	0.54	0.09
Designer	0.98	0.10	0.85	0.00	0.35	0.02	0.29	0.03	0.43	0.03	0.38	0.04	0.22	0.05	0.25	0.05
Developer	0.66	0.03	0.25	0.13	0.74	0.02	0.40	0.08	0.64	0.01	0.26	0.06	0.53	0.05	0.29	0.17
Doctor	0.88	0.01	0.35	0.05	0.85	0.02	0.38	0.04	0.81	0.03	0.24	0.08	0.41	0.04	0.44	0.03
Driver	0.29	0.01	1.00	0.00	0.73	0.02	0.68	0.02	0.51	0.01	0.73	0.03	0.47	0.03	0.48	0.05
Farmer	0.90	0.10	0.15	0.19	0.58	0.02	0.27	0.06	0.61	0.04	0.16	0.02	0.85	0.04	0.81	0.05
Guard	0.16	0.41	0.89	0.38	0.55	0.05	0.64	0.07	0.33	0.01	0.66	0.07	0.32	0.06	0.25	0.09
Hairdresser	0.45	0.05	1.00	0.07	0.74	0.08	0.27	0.08	0.59	0.03	0.33	0.01	0.51	0.19	0.48	0.20
Housekeeper	0.64	0.13	0.94	0.06	0.85	0.04	0.66	0.05	0.69	0.04	0.69	0.07	0.34	0.03	0.65	0.06
Janitor	0.88	0.19	0.13	0.13	0.76	0.02	0.83	0.06	0.75	0.04	0.41	0.02	0.32	0.07	0.29	0.06
Laborer	0.28	0.13	0.93	0.11	0.21	0.07	0.27	0.15	0.16	0.04	0.34	0.03	0.27	0.05	0.61	0.09
Lawyer	0.95	0.05	0.88	0.01	0.98	0.02	0.89	0.06	0.95	0.02	0.84	0.02	0.48	0.04	0.42	0.06
Librarian	0.93	0.05	0.99	0.01	0.38	0.04	0.33	0.02	0.44	0.04	0.39	0.02	0.69	0.05	0.75	0.05
Manager	0.60	0.00	0.09	0.03	0.66	0.03	0.33	0.02	0.60	0.01	0.16	0.04	0.38	0.05	0.38	0.09
Mechanic	0.99	0.04	0.28	0.07	0.78	0.05	0.43	0.16	0.80	0.01	0.25	0.03	0.41	0.05	0.42	0.08
Nurse	0.99	0.16	0.95	0.01	0.32	0.03	0.66	0.02	0.41	0.01	0.69	0.06	0.79	0.05	0.71	0.07
Physician	0.74	0.05	0.99	0.01	0.71	0.06	0.33	0.04	0.66	0.02	0.36	0.04	0.39	0.03	0.32	0.03
Receptionist	1.00	0.03	1.00	0.06	0.38	0.09	0.30	0.08	0.46	0.08	0.40	0.05	0.81	0.11	0.75	0.13
Salesperson	0.28	0.01	0.49	0.01	0.83	0.01	0.42	0.02	0.51	0.04	0.36	0.06	0.63	0.05	0.27	0.12
Secretary	0.98	0.03	1.00	0.00	0.98	0.08	0.94	0.05	0.96	0.01	0.95	0.07	0.43	0.07	0.20	0.05
Sheriff	0.32	0.06	0.94	0.04	0.30	0.02	0.59	0.04	0.21	0.02	0.63	0.04	0.80	0.02	0.89	0.07
Supervisor	0.95	0.01	1.00	0.15	0.21	0.02	0.95	0.06	0.28	0.02	0.96	0.05	0.21	0.05	0.27	0.07
Tailor	0.41	0.01	0.99	0.00	0.61	0.02	0.17	0.13	0.46	0.01	0.29	0.03	0.24	0.04	0.40	0.06
Teacher	0.91	0.03	0.80	0.04	0.99	0.03	0.71	0.09	0.94	0.02	0.64	0.04	0.30	0.05	0.40	0.07
Writer	0.69	0.01	0.32	0.00	0.67	0.03	0.33	0.02	0.58	0.01	0.14	0.06	0.44	0.04	0.57	0.05
Average	0.71	0.06	0.71	0.06	0.60	0.04	0.55	0.06	0.56	0.03	0.50	0.05	0.45	0.05	0.46	0.08

Table 10. Deviation ratio per occupation of the original SD and our model.