Bias Is a Subspace, Not a Coordinate: A Geometric Rethinking of Post-hoc Debiasing in Vision-Language Models

Supplementary Material

8. Qualitative Results for T2I Generation

Figure 2 presents qualitative examples of text-to-image generation results under the neutral prompt "*a person who works as a film director.*" The CoDi baseline tends to generate male-presenting images due to training bias, while CoDi + SPD produces more gender-neutral appearances, indicating effective debiasing.

male male male

CoDi

female female male

CoDi + SPD

Figure 2. Text-to-image generation results for the neutral prompt "a person who works as a film director." The first row shows CoDi outputs, and the second row shows CoDi with SPD debiasing. Gender labels ("male" / "female") are automatically assigned using BLIP-2 by asking "Does the person look like a male or a female?".

9. Impact of r on Downstream Tasks

In SPD, the number of removed bias-predictive directions r determines a trade-off between debiasing completeness and semantic retention. We first vary r to study its effect on attribute classification by training linear probes after projection, finding that a moderate projection depth effectively removes target-attribute information while largely preserving semantics. To check whether the optimal setting in downstream tasks aligns with classification, we evaluate text-to-image retrieval on Flickr30K under varying r. As shown in Tab. 8, $r = 5$ achieves the best balance between retrieval accuracy (R@K) and fairness (Skew@100), consistent with the choice used across downstream evaluations.

Table 8. Text-to-image retrieval (Flickr30K) results for CLIP (ResNet50) with varying r.

Model		Text-to-Image Retrieval R@1	R@5	R@10	Skew@100
SPD	$r = 1$	57.11±0.55	81.62±0.74	88.05±0.58	0.1613±0.1025
	$r = 3$	57.04±0.73	81.65±0.62	88.09±0.52	0.1527±0.0995
	$r = 5$	56.97±0.57	81.42±0.66	87.85±0.63	0.1177±0.0830
	$r = 7$	55.47±0.57	80.81±0.75	87.13±0.51	0.1165±0.0971
	$r = 10$	55.32±0.65	80.62±0.66	86.99±0.55	0.1168±0.0851

10. Computing Resource

Table 9 presents the hardware configuration and runtime of each module used in our experiments. The reported times include Random Forest training for the subsequent projection-direction extraction. Overall, SPD adds limited computational cost during inference while maintaining efficiency across all tested backbones.

Table 9. Compute resources used for experiments.

Component	**Details**
CPU	Intel Xeon Gold 6430
Numbers of CPU cores used	32
GPU	NVIDIA RTX A100
Numbers of GPU used	4
(CLIP ViTB-32 Image Encoder)	
Training RandomForest	310.47s
Direction Extraction	37.36s
(CLIP RN50 Image Encoder)	
RandomForest Training	466.52s
Projection Direction Extraction	35.37s
(CLIP ViTB-32 Text Encoder)	
Training RandomForest	1571.34s
Prodection Direction Extraction	93.77s
(CLIP RN50 Text Encoder)	
RandomForest Training	2567.06s
Projection Direction Extraction	146.31s
(CoDi Text Encoder)	
Training RandomForest	62.71s
Projection Direction Extraction	3.73s
(CoDi Image Dncoder)	
Training RandomForest	197.33s
Projection Direction Extraction	52.63s
Inference on CoDi with SPD	403.62s / 249 prompts