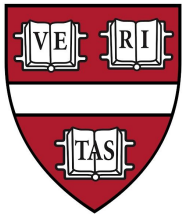
Bias Is a Subspace, Not a Coordinate: A Geometric Rethinking of Post-hoc Debiasing in Vision-Language Models

Dachuan Zhao*, Weiyue Li*, Zhenda Shen*, Yushu Qiu, Bowen Xu, Haoyu Chen, Yongchao Chen



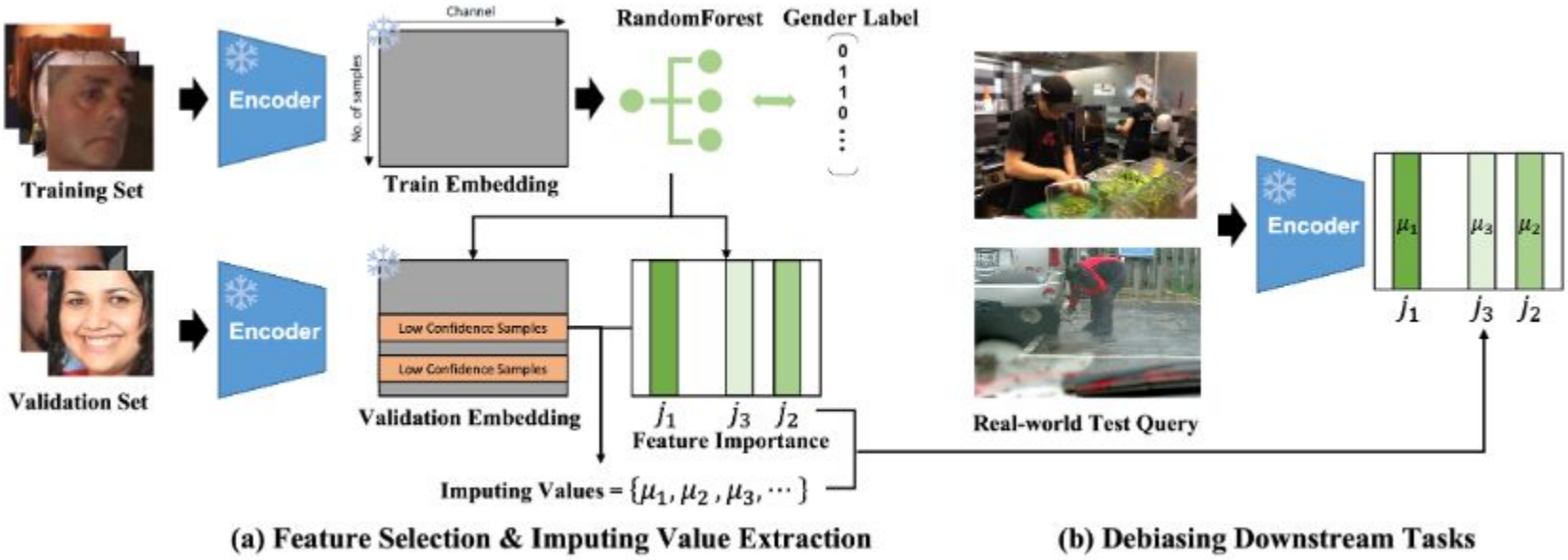
Paper PDF

Problem: Bias in Vision-Language Models

Prompt: Generate a Nurse's Picture

Prompt: Generate a Doctor's Picture

Existing Method: SFID



Why Existing Methods Fail

Coordinate-level debiasing assumptions break in practice:

1. Feature Entanglement

Table 1. Overlap of top-m=100 dimensions across age (A), gender (G), and race (R) attributes on FairFace. Higher overlap indicates stronger feature entanglement.

	A ∩ G	**G ∩ R**	**A ∩ R**	**A ∩ G ∩ R**
Overlap	31	37	20	11

Why Existing Methods Fail

Coordinate-level debiasing assumptions break in practice:

2. Cross-Dataset Dimension Shift

Table 2. Intersection of top-m gender dimensions between FairFace and FACET. Weak alignment shows that direct embedding fails to achieve robust and effective cross-dataset transfer.

m (FairFace)	m (FACET)	Overlap
50	50	24
100	100	40

Why Existing Methods Fail

Coordinate-level debiasing assumptions break in practice:

3. Incomplete Debiasing

Table 3. Linear-probe accuracy for predicting protected attributes (Race/Gender/Age) from FairFace embeddings after applying each intervention. "Origin" reports accuracy on the original vanilla embeddings, while the remaining columns report SFID with $m = 100$ and SPD with $r = 1, 5, 10$ removed subspace directions.

		Replace Race				Replace Gender				Replace Age			
Class	**Origin**	**SFID**	**SPD(r=1)**	**SPD(r=5)**	**SPD(r=10)**	**SFID**	**SPD(r=1)**	**SPD(r=5)**	**SPD(r=10)**	**SFID**	**SPD(r=1)**	**SPD(r=5)**	**SPD(r=10)**
Race	0.7144	0.7086	0.7149	0.2745	0.1913	0.7127	0.7140	0.7150	0.7148	0.7118	0.7149	0.6964	0.3080
Gender	0.9466	0.9449	0.9466	0.9467	0.6832	0.9404	0.9286	0.6766	0.5925	0.9465	0.9466	0.9397	0.5549
Age	0.6023	0.6023	0.6026	0.5988	0.4745	0.5991	0.6024	0.6009	0.5976	0.5961	0.6023	0.3000	0.2958

Bias is a Subspace, Not a Coordinate

Our idea:

- Bias is distributed across correlated directions
- Learn a bias subspace using iterative linear probes
- Project embeddings away from the bias subspace

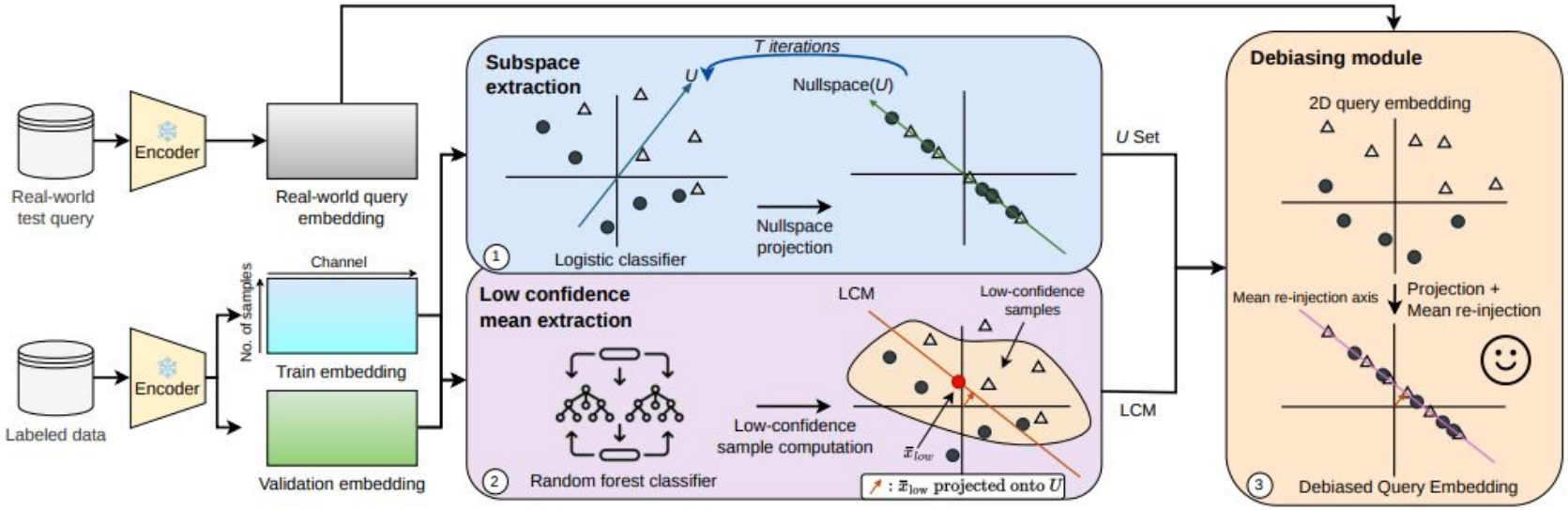
$$x' = x\,(I - U^\top U)$$

Project onto the orthogonal complement of the bias subspace

$$x'' = x' + U^\top\big(U\,\bar{x}_{\text{low}}\big)$$

Reinject a neutral mean to preserve semantics

Methodology



Addressing Three Failed Assumptions of SFID

- **Feature Entanglement**
  → Dropped accuracy of one category does not affect the other two
- **Incomplete Debiasing**
  → Protected-attribute probe accuracy approaches random chance
- **Cross-Dataset Dimension Shift**
  → Downstream cross-dataset performance shows stronger robustness

Table 3. Linear-probe accuracy for predicting protected attributes (Race/Gender/Age) from FairFace embeddings after applying each intervention. "Origin" reports accuracy on the original vanilla embeddings, while the remaining columns report SFID with $m = 100$ and SPD with $r = 1, 5, 10$ removed subspace directions.

		Replace Race				Replace Gender				Replace Age			
Class	**Origin**	**SFID**	**SPD(r=1)**	**SPD(r=5)**	**SPD(r=10)**	**SFID**	**SPD(r=1)**	**SPD(r=5)**	**SPD(r=10)**	**SFID**	**SPD(r=1)**	**SPD(r=5)**	**SPD(r=10)**
Race	0.7144	0.7086	0.7149	0.2745	0.1913	0.7127	0.7140	0.7150	0.7148	0.7118	0.7149	0.6964	0.3080
Gender	0.9466	0.9449	0.9466	0.9467	0.6832	0.9404	0.9286	0.6766	0.5925	0.9465	0.9466	0.9397	0.5549
Age	0.6023	0.6023	0.6026	0.5988	0.4745	0.5991	0.6024	0.6009	0.5976	0.5961	0.6023	0.3000	0.2958

Result - Zero Shot Classification

Predict occupation classes using image-text similarity

Setup

- Dataset: FACET benchmark
- 52 occupation classes with demographic annotations

Evaluation

- Utility: Top-1 Accuracy ↑
- Fairness: Demographic Parity Gap ↓

Model		Zero-shot Multi-class Classification Accuracy	Δ DP	↑(%)
CLIP (ResNet50)	Baseline	51.87±0.58	11.08±0.90	—
	SFID	50.93±0.57	9.63±0.86	13.1
	SPD (Ours)	51.44±0.64	**9.55±0.81**	**13.8**
CLIP (ViT-B/32)	Baseline	52.17±0.58	11.60±0.93	—
	SFID	52.14±0.53	10.15±0.85	12.5
	SPD (Ours)	51.29±0.52	**9.94±0.76**	**14.3**
XVLM	Baseline	55.74±0.48	11.72±0.72	—
	SFID	53.69±0.59	9.91±0.92	15.4
	SPD (Ours)	54.32±0.47	**9.85±0.84**	**16.0**

Result - Text-to-Image Retrieval

Retrieve images from gender-neutral text queries

Setup

- Dataset: Flickr30K
- 1000 gender-neutralized caption-image pairs

Evaluation

- Utility: Recall@K ↑
- Fairness: Skew@100 ↓

Model		Text-to-Image Retrieval				
		R@1	R@5	R@10	Skew@100	↑(%)
CLIP (ResNet50)	Baseline	57.24±0.58	81.66±0.61	88.12±0.56	0.1883±0.0939	—
	SFID	56.94±0.51	80.89±0.62	87.41±0.60	0.1414±0.0955	24.9
	SPD (Ours)	56.97±0.57	81.42±0.66	87.85±0.63	**0.1177±0.0830**	**37.5**
CLIP (ViT-B/32)	Baseline	58.91±0.51	83.08±0.62	89.21±0.48	0.1721±0.0992	—
	SFID	58.53±0.70	82.73±0.56	88.90±0.56	0.0744±0.0616	56.8
	SPD (Ours)	**59.68±0.46**	**83.47±0.54**	**89.35±0.59**	**0.0699±0.0566**	**59.4**
XVLM	Baseline	80.77±0.56	96.67±0.26	98.55±0.23	0.2355±0.1425	—
	SFID	78.00±0.46	95.67±0.45	98.01±0.25	0.2032±0.1229	13.7
	SPD (Ours)	79.13±0.44	96.11±0.44	98.25±0.20	**0.1859±0.1217**	**21.1**

Result - Text-to-Image Generation

Generate images from profession-related text prompts

Setup

- Dataset: Bias-in-Bios profession prompts
- 83 profession categories

Evaluation

- Mismatch Rate ↓
- Composite Score ↓
- Neutral Prompt Skew ↓

(a) SDXL

Method	Mismatch Rate (Gender prompt) |M–F|	Overall	Composite	Neutral prompt *Skew*
Baseline	3.87±2.23	2.35±1.22	4.42±2.57	83.25
SFID	1.54±1.14	0.84±0.71	1.74±1.57	81.57
SPD (Ours)	**1.48±0.99**	**0.78±0.65**	**1.67±1.53**	**78.66**

(b) CoDi

Method	Mismatch Rate (Gender prompt) |M–F|	Overall	Composite	Neutral prompt *Skew*
Baseline	3.94±2.71	5.54±2.08	6.85±2.16	84.94
SFID	4.70±1.53	2.59±0.90	5.38±1.44	82.77
SPD (Ours)	**3.62±1.64**	**2.53±0.89**	**5.26±1.21**	**81.20**

Conclusion

Bias in VLMs is subspace-structured.

- Coordinate editing is insufficient
- SPD removes distributed bias directions
- Better fairness with strong utility retention

SPD: Training-free, interpretable, and robust post-hoc debiasing

Paper PDF