KAIST AI
Graduate School of AI

CVPR
JUNE 3-7, 2026
DENVER
COLORADO

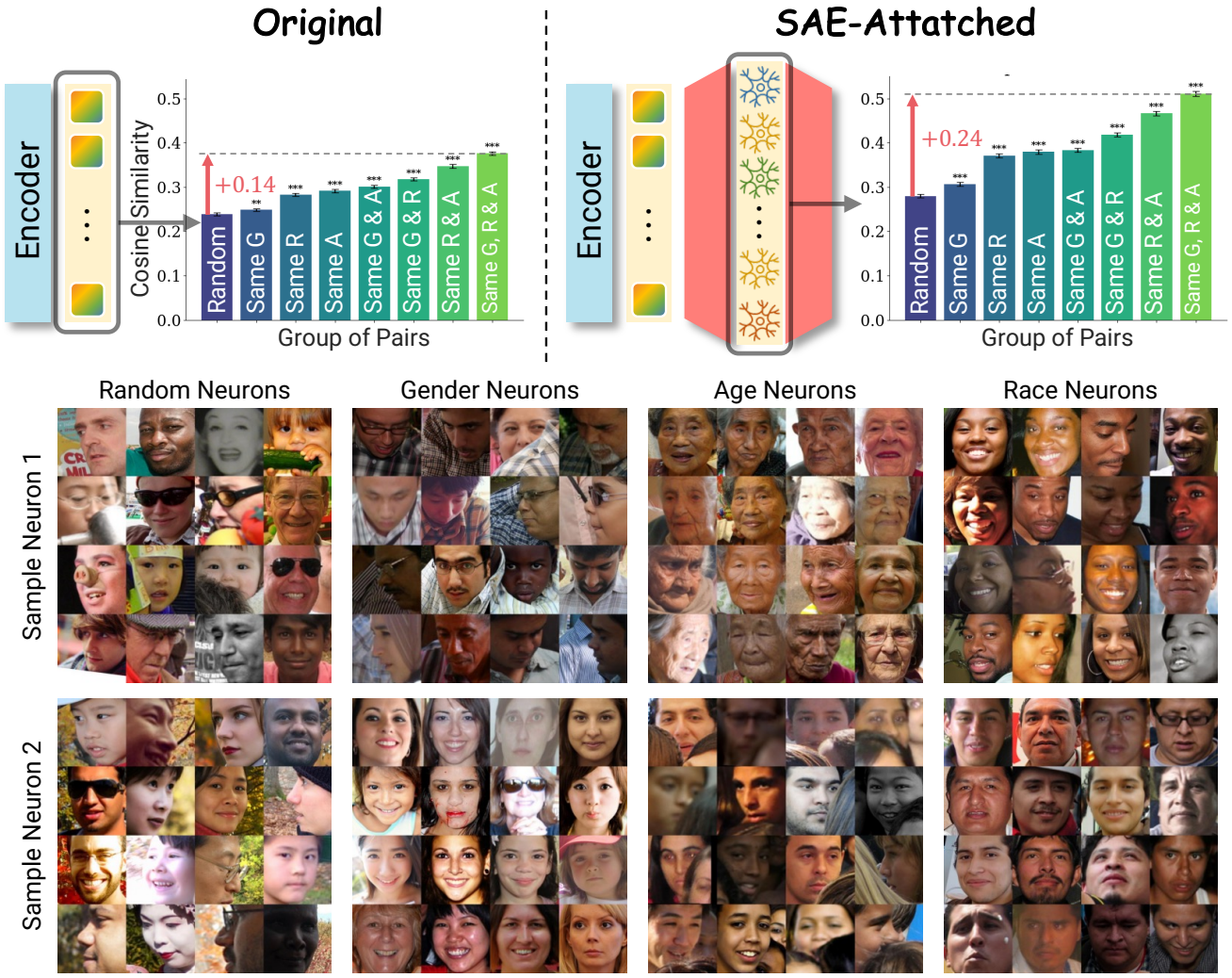
Interpretable Debiasing of Vision-Language Models for Social Fairness

Na Min An[1], Yoonna Jang[2], Yusuke Hirota[3], Ryo Hachiuma[3], Isabelle Augenstein[2], Hyunjung Shim[1]
[1]KAIST AI, [2]University of Copenhagen, [3]NVIDIA

Demographic Bias in Multimodal Models

While existing models retrieve image distribution of skewed demographics or answer definitively on ambiguous image-text pairs, our DEBIASLENS alleviates social biases across both image and text modalities.

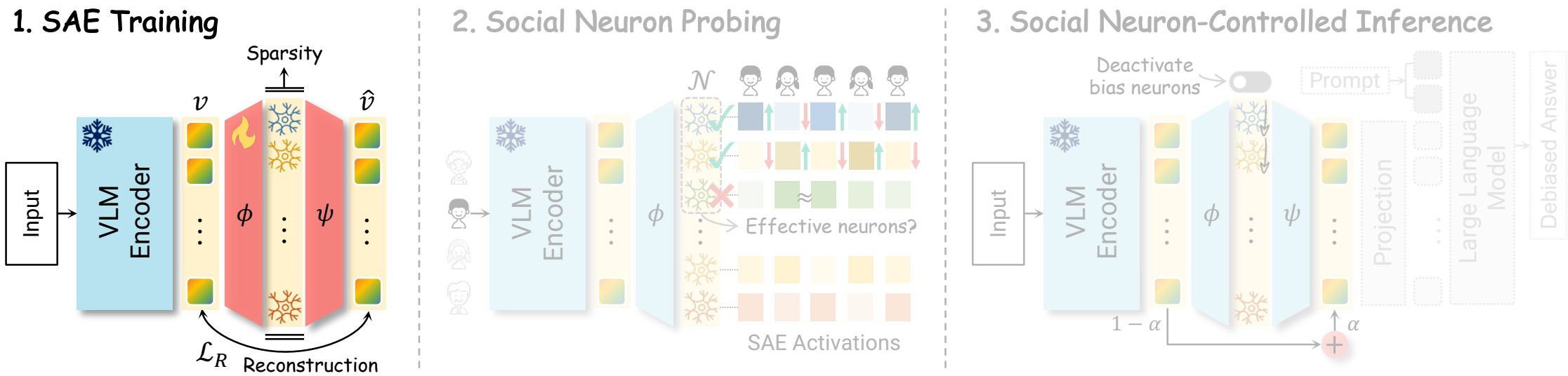
Motivation



SAE neurons can implicitly capture social attributes even without explicit supervision from demographic labels.

Hypothesis: Social neurons contributing to a model's bias exhibit differential SAE activation patterns across data groups representing different social attributes.

Interpretable Debiasing Framework

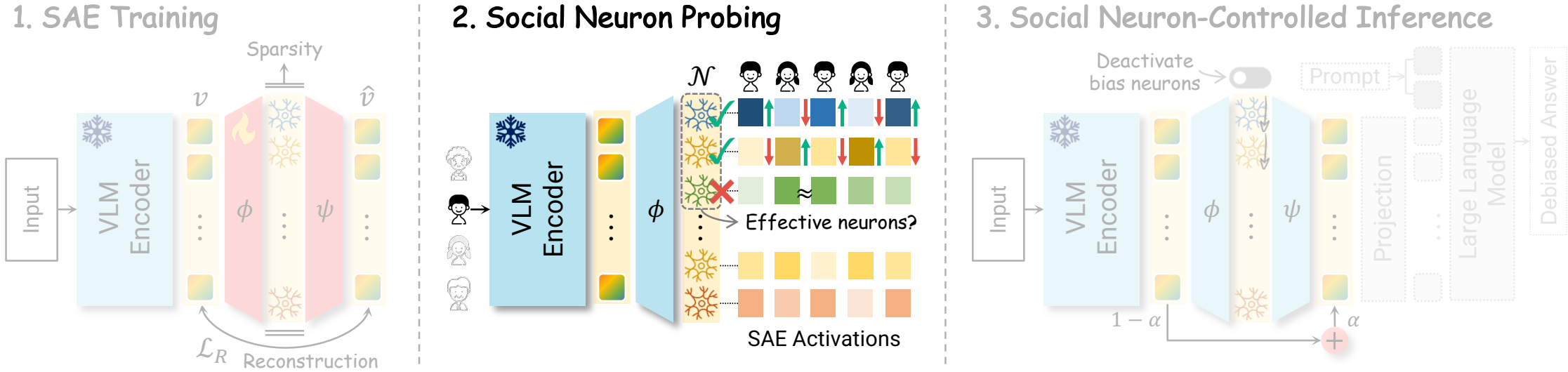


1. SAE is trained on top of the last layer of the VLM image/text encoder.

$$\mathcal{L}_1(\mathbf{v}) = \mathcal{L}_R(\mathbf{v}) + \lambda\|\phi(\mathbf{v})\|_1,$$

$$\mathcal{L}_R(\mathbf{v}) = \sum_{m\in\mathcal{M}} \left\|\mathbf{v} - \mathbf{W}_{\text{dec}}^{\top}\phi_{1:m}(\mathbf{v})\right\|_2^2.$$

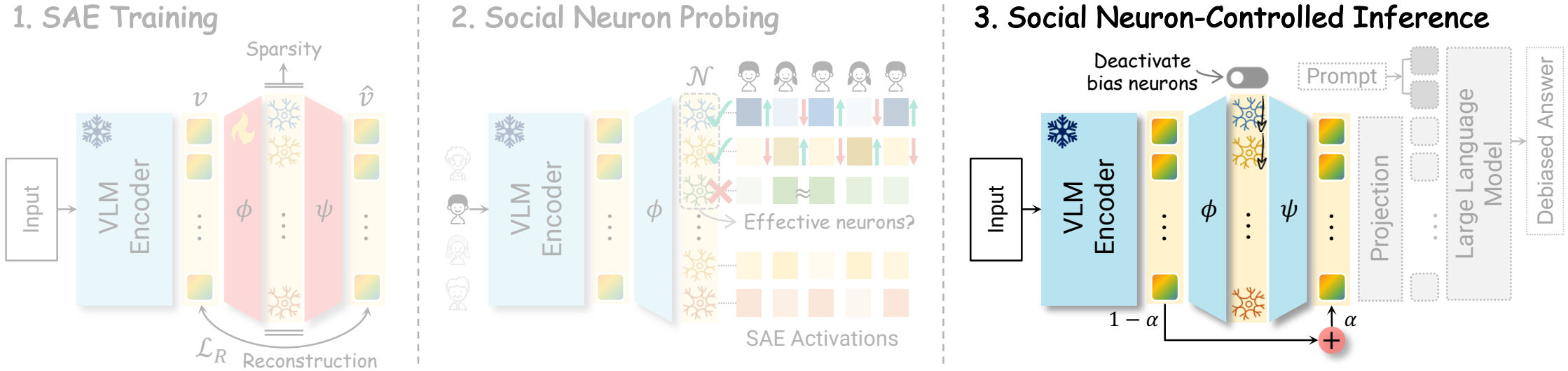
Interpretable Debiasing Framework



2. The social neurons are identified based on the consistency and specificity of SAE activations across data.

$$\mathcal{N}_g = \mathcal{E}_g \setminus \left(\bigcup_{h \in G, h \neq g} \mathcal{E}_h \right) \qquad \bar{\mathbf{s}}_j = \frac{1}{S_g} \sum_{i=1}^{S_g} \mathbf{x}_{i,j}^{(g)}, \text{for } j \in \mathcal{N}_g \qquad j_g^* = \arg\max_{j \in \mathcal{N}_g} (\bar{\mathbf{s}}_j).$$

Interpretable Debiasing Framework



3. The selected neurons are activated to generate debiased features, weighted summed with original features for further usage across downstream tasks.

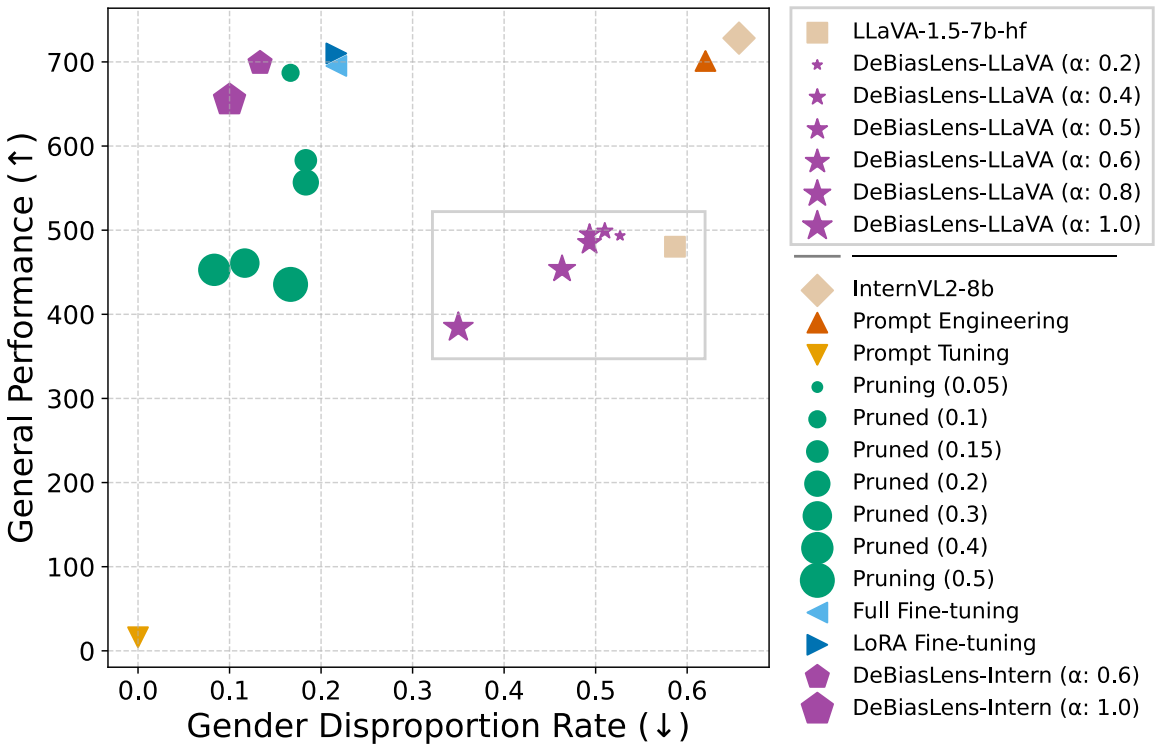
$$\mathbf{z}'[j] = \begin{cases} \gamma & \text{if } j \in \mathcal{Z}_{\mathrm{B}} \\ \mathbf{z}[j] & \text{otherwise} \end{cases} \qquad \hat{\mathbf{v}} = \psi(\mathbf{z}'). \qquad \mathbf{v}' = \alpha\hat{\mathbf{v}} + (1-\alpha)\mathbf{v}$$

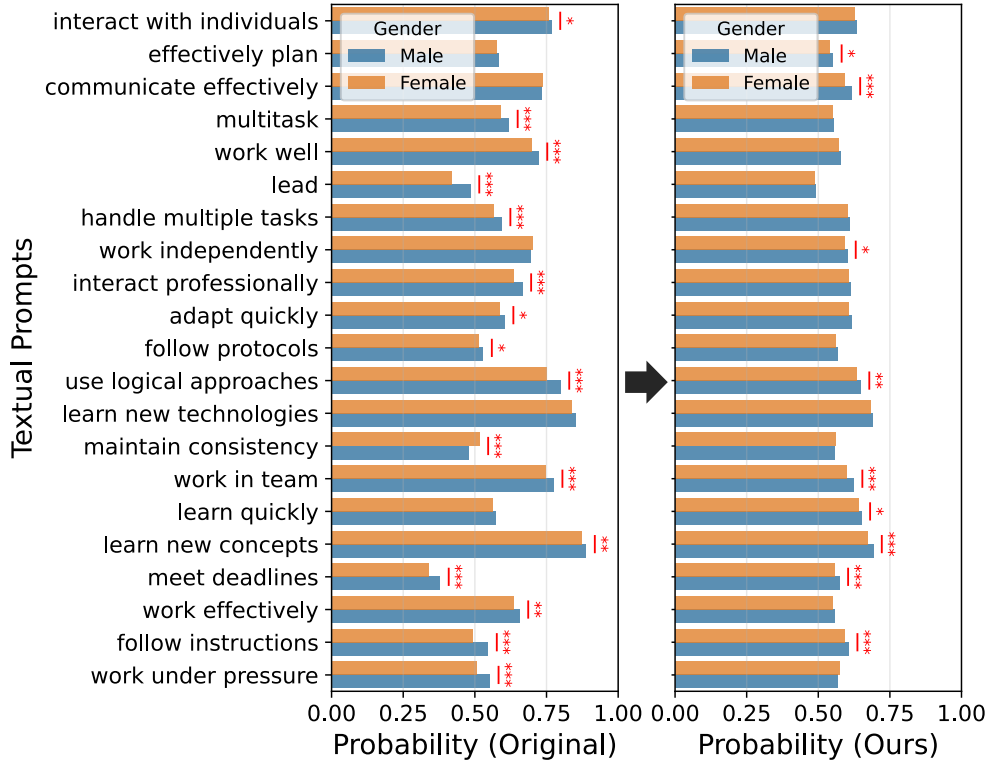
VLM Debiasing

Methods	Interpretable?	Max Skew (↓)			
		Adj	Occup	Act	Ster
CLIP (ViT-B/16) [67]	–	22.9	33.7	19.5	33.8
CLIP (ViT-B/16)†	–	21.9	33.5	19.8	32.5
Prompt [8]	✗	12.3	29.9	20.0	-
Prompt†	✗	11.9	29.8	19.3	28.7
Projection [14]	✗	15.4	37.4	15.0	52.0
Bend-VLM† [19]	✗	10.8	**10.2**	9.8	9.1
SANER [28]	✗	8.9	14.5	**7.7**	-
DEBIASLENS (I)	✓	14.2	21.5	20.0	18.3
DEBIASLENS (T)	✓	**7.1**	16.2	14.2	**8.1**
DEBIASLENS (I+T)	✓	11.1	19.4	18.0	10.3
CLIP (ViT-L/14@336)†	–	19.9	31.5	23.2	30.0
MMNeuron [12]	✓	17.3	23.5	26.0	20.3
DEBIASLENS (I)	✓	**12.0**	**20.4**	**17.2**	**11.2**
DEBIASLENS (T)	✓	16.3	27.6	26.9	21.2
DEBIASLENS (I+T)	✓	16.2	25.1	19.9	24.2

Applying interpretable DEBIASLENS to the image and/or text encoder of two widely used CLIP variants, we notice a significant decrease in Max Skew scores.

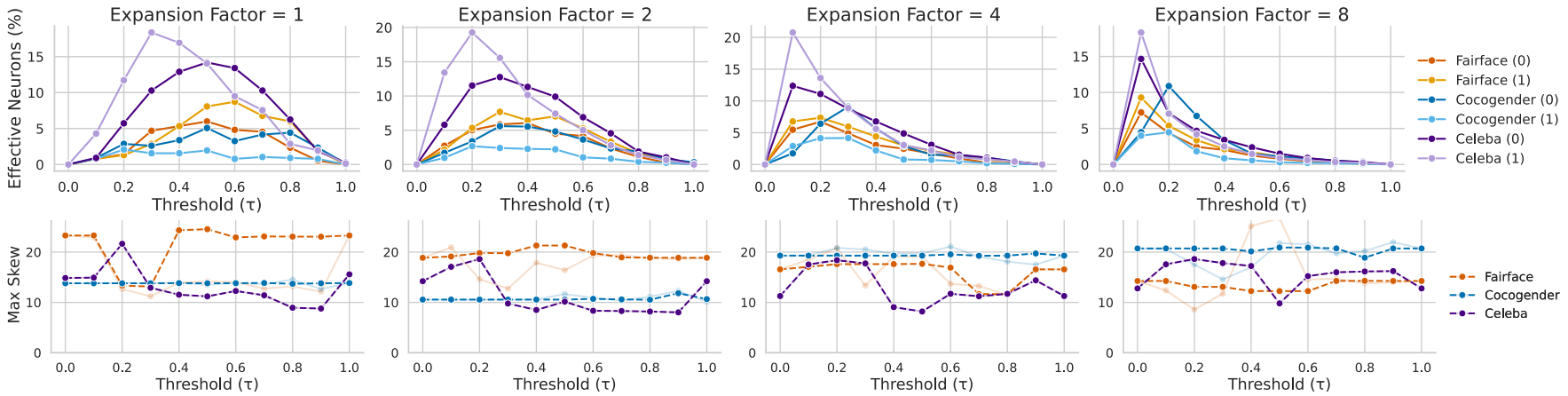
LVLM Debiasing

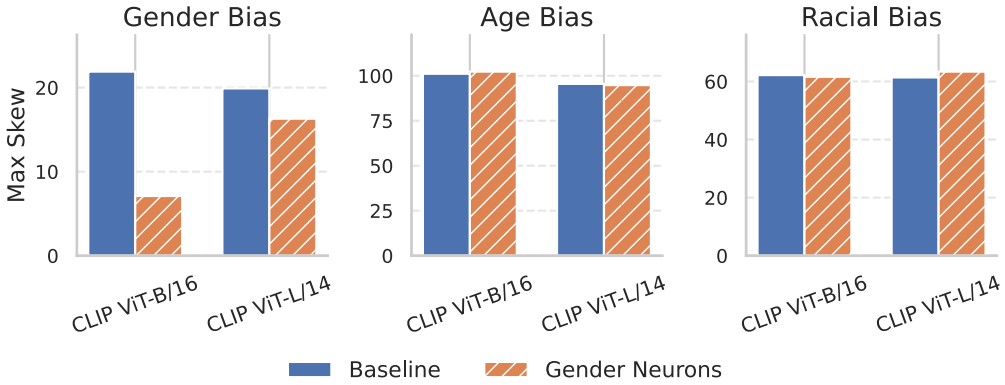


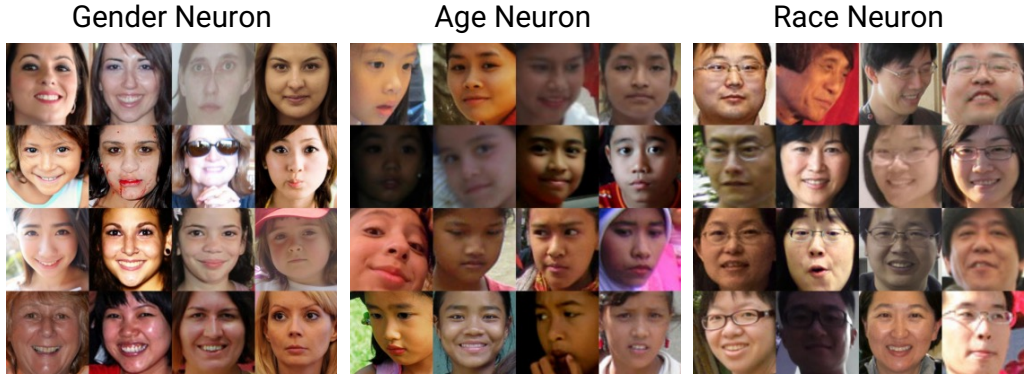


Our method achieves the best trade-off among other existing debiasing methods.
The probability difference between genders becomes non-significant for most textual prompts.

Interpretable Social Neurons







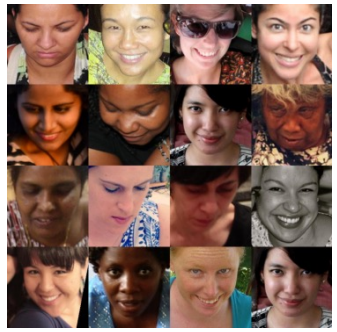
Modulating few gender neurons mitigates only gender bias, indicating high neuron specificity.
Each social neuron corresponds to a human-interpretable concept of a social attribute.

Data Distribution Effects

CelebA

COCOgender

FairFace



Bias in Bios

```
"He is 72 years old and has been practicing for 46
years. Dr. Oneacre is affiliated with Baylor
Medical Center at Carrollton and Texas Health
Presbyterian Hospital Flower Mound."

"He has been practicing for 46 years and received
his medical degree from Jawaharlal Institute of
Postgraduate Medical Education and Research. No
Reviews Favorite"

"The practice was established in 1980, and has
included projects in regional planning, housing,
schools, commercial buildings, agriculture,
rehabilitation, etc."
```

COCOgender captions

```
"A man standing behind a robot behind an
orange ball."

"A boy throws a baseball into the air."

"A man looking upward at a tennis ball. "

"A couple of people standing on a court
with a ball."

"A man holding a soccer ball in a field"

"a man catching a ball with a young boy"
```

Original

SB-Syn

SB-Syn-Crop

Methods	Eval	Train Data	Probing Data	Gender	Age
InternVL2-8B	Rule	✗	✗	83.83	43.11
DeBiasLens	Rule	SB-Syn	SB-Syn	84.32	44.59
DeBiasLens	Rule	SB-Syn-Crop	SB-Syn-Crop	84.71	45.55
DeBiasLens	Rule	FairFace	SB-Syn	86.32	47.17
DeBiasLens	Rule	FairFace	SB-Syn-Crop	86.49	47.21
DeBiasLens	Rule	FairFace	FairFace	86.68	47.52
DeBiasLens	Rule	FairFace	FairFace (α=1.0)	87.87	48.51
InternVL2-8B	Phi	✗	✗	85.97	50.35
DeBiasLens	Phi	SB-Syn	SB-Syn	85.68	50.42
DeBiasLens	Phi	SB-Syn-Crop	SB-Syn-Crop	87.07	51.51
DeBiasLens	Phi	FairFace	SB-Syn	87.68	53.46
DeBiasLens	Phi	FairFace	SB-Syn-Crop	87.81	53.46
DeBiasLens	Phi	FairFace	FairFace	88.39	52.54
DeBiasLens	Phi	FairFace	FairFace (α=1.0)	89.49	53.77

Deactivating SAE neurons probed with the Fairface dataset show higher performance than in-distribution datasets.

Ablation Study

α	CLIP ViT-B/16				LLaVA-1.5-7b-hf	
	ImgNette ↑ [29]		**FairFace** ↓ [41]		**MME** ↑ [12]	**VLA** ↓ [21]
	Image	Text	Image	Text	Image	Text
0.0	99.5	99.1	18.8	16.7	1440.10	0.62
0.2	99.3	99.0	17.8	12.8	1479.28	0.57
0.4	99.0	99.1	16.2	8.7	1496.10	0.53
0.5	98.5	98.9	15.2	7.4	1483.89	0.50
0.6	97.5	98.5	14.2	7.1	1454.26	0.50
0.8	88.6	96.3	11.7	9.2	1360.22	0.48
1.0	59.1	87.8	10.6	13.2	1152.33	0.41

While we select the α to be 0.6 throughout the main experiments to ensure a balance between general and bias mitigation performance, practitioners could select α based on their trade-off preference.

Towards Unbiased Multimodal Models

- ✓ We propose DEBIASLENS, an interpretable debiasing framework for actively identifying and modulating the social neurons.
- ✓ Beyond demonstrating strong empirical performance, across multiple VLM architectures and domains, DEBIASLENS transforms bias mitigation into a black-box correction into an interpretable intervention.
- ✓ We envision the framework as a foundational step toward trustworthy and responsible AI, inspiring future research into developing unbiased multimodal systems.

Thank you! ☺