Interpretable Debiasing of Vision-Language Models for Social Fairness

Supplementary Material

Overview of contents This supplementary material contains the following:

A. Training Details

Since our method can be applied to any encoders, we train the image/text encoder for Vision-Language Models (VLMs) and the image encoder for Large Vision-Language Models (LVLMs) (Table S1). For training the sparse autoencoder (SAE), we employ the Matryoshka [15] variant with top-k ($= 20$) sparsity and hierarchical grouping [17]. The original activations are extracted from the corresponding encoder layers for both training and validation, with a batch size of 4096 and expansion factors of 1, 2, 4, and 8 to control the dictionary size. The model is optimized for 110,000 epochs, and the weight for auxiliary loss is set to be 0.03, with the decay of learning rate starting at step 109,999. We divide the SAE neurons into four groups using the fractions [0.0625, 0.125, 0.25, 0.5625], meaning that 6.25%, 12.50%, 25.00%, and 56.25% of the neurons are assigned to each group, respectively; this grouping lets us evaluate hierarchical behavior at different levels of neuron specificity.

B. Data Details

Real Data To ensure the balanced selection of the social neurons, each of the training datasets of sparse autoencoder (SAE) [15] has a fair distribution of social attribute labels, as in the original dataset (statistics in Table S9). For instance, the gender ratio of the Bias in Bios dataset [6] is 54:44 for male and female labels. This case even more strictly applies to the evaluation datasets. As can be seen in Table S1, the gender ratio of the FairFace evaluation dataset[1] [11] is 50:50. However, for each group, the balance becomes uneven, notably in the 20-29 age range. This inherent skew in the data distribution helps explain the intersectional effect of why modulating the age neurons can effectively reduce gender bias. Additionally, we observe 40% (out of 25) of the age neurons are gender-skewed (Figure S2). Not modulating these gender-skewed age neurons increases both gender and age MaxSkews by 0.6% and 5.3% ($\alpha = 1$), suggesting these are indeed *age* neurons. Note that all the datasets are publicly available.

[1]From a total of 10,954 cropped images, we sample a subset with a gender balanced distribution, following [2].

Table S1. **Overview of multimodal models.** The table lists the image and text encoders used in VLMs and LVLMs considered in this work.

Model	Type	Image Encoder	Text Encoder
CLIP (ViT-B/32)	VLM	ViT-B/32	Transformer (512-d)
CLIP (ViT-L/14@336)	VLM	ViT-L/14@336	Transformer (768-d)
LLaVA-1.5-7B	LVLM	ViT-L/14@336	–
LLaVAOneVision	LVLM	SigLIP-so400m/14@384	–
InternVL2-8B	LVLM	InternViT-300M@448px	–

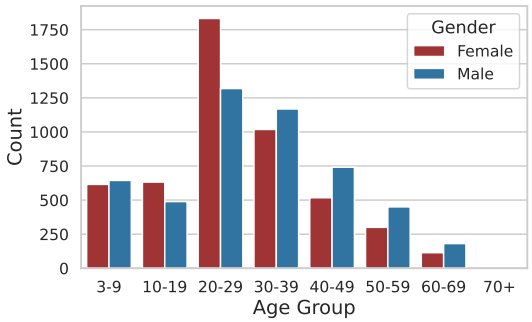

Figure S1. **Age distribution of FairFace evaluation dataset across genders.** Although the gender distribution is balanced, there is a skewed gender distribution per age group.

Figure S2. **Age neuron activating images.**

Synthetic Data To explore the possibility of extending the real-world data to synthetic data for SAE training and probing, we generate new synthetic datasets from the SBBench dataset [16]. We postprocess the images corresponding to the age and gender category to enable a direct comparison with realistic data, FairFace. Since the images from other categories (*e.g.*, religion, socio-economic status, disability) seem to have visual cues that are more closely related to the contextual factors like background or clothing, we mainly test age and gender categories in this work.

From the original SBBench synthetic dataset (760 and 672 samples from the age and gender categories), we remove duplicates to obtain 627 and 302 im-

ages, respectively. Since images often contain multiple humans, we use a recent text-to-image editing model `Qwen-Image-Edit` [19] to leave only one person with a certain social attribute (*i.e.*, old/young and male/female for the age and gender categories). Specifically, we prompt the model to edit each image and exclude those where the face is too distant, not visible, contains multiple people, or is of poor quality. We repeat this process three times for the age category and two times for the gender category, making small adjustments to the prompt at each iteration, and then conduct human validation to ensure that the final images meet the required quality and attribute criteria. The prompts used for the editing model are as follows:

Prompts for SBBench-Syn Data Construction

1st round generation Leave only one human who is {old/young/male/female}, and remove any other humans. Keep the background identical to the original. Make the image square.

2nd/3rd round generation (age) Leave only {old/young} {person/human}, and remove any other humans. Leave the background exact same as the original. Make the image square.

2nd round generation (gender) Leave only one {male/female}, and remove any other humans. Leave the background exact same as the original. Make the image square.

Prompts for SBBench-Syn-Crop Data Construction

1st round generation Leave only one human who is {old/young/male/female}, and remove any other humans. Crop and zoom in so that the remaining person's face appears larger and clearly visible. Make the image square.

2nd/3rd round generation (age) Leave only {old/young} {person/human}, and remove any other humans. Crop and zoom in so that the remaining person's face appears larger and clearly visible. Make the image square.

2nd round generation (gender) Leave only one {male/female}, and remove any other humans. Crop and zoom in so that the remaining person's face appears larger and clearly visible. Make the image square.

We use the above prompts to generate new synthetic datasets, SB-Syn and SB-Syn-Crop (sample synthesized images in Figure S6), to examine whether having less background context could help to find more effective social neurons. As a result, from the 627 and 302 age and gender-group images in the original SBBench dataset, we extract 246/87 images featuring a single old/young individual and 109/97 featuring a single male/female individual for constructing SB-Syn. Similarly, for SB-Syn-Crop, we extract 222/184 images featuring a single old/young individual and 352/159 images featuring a single male/female individual. This scarcity of data may provide reasons why the SAE trained and probed using FairFace data achieved strong performance.

Table S2. **Bias Score Results on Non-Overlapping Datasets.**

Datasets	PATA				Pairs				
Prompts	Adj	Occup	Act	Ster	Adj	Occup	Act	Ster	*Avg*
CLIP (ViT-B/16)	14.1	19.9	10.2	20.3	13.1	22.4	26.8	18.5	18.16
Prompt	7.3	16.9	10.0	16.0	11.9	23.9	15.2	15.0	14.53
DEBIASLENS	7.1	13.3	11.9	5.5	10.5	19.5	17.7	11.6	**12.14**

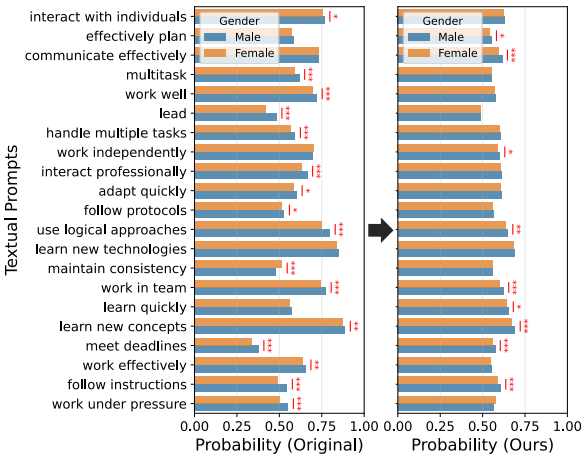

Figure S3. **Difference between male and female 'yes' probability across skill-related prompts for InternVL2-8b and DEBIASLENS-Intern.** Our method shows most of the skills having no statistically different probability across genders (*: $p < 0.1$, **: $p < 0.01$, ***: $p < 0.001$).

Table S3. **Computational Cost Results.** The trade-off score (↑ the better) is proportional to ΔBiasScore $-$ ΔVLAPerf.

Method	Par (M)	GPU hrs	Overhead (ms)	FLOPs	Trade-off
Full FT	6979.58	0.02	310.19	1.14e+13	1.29
LoRA FT	301.99	0.32	310.89	1.14e+13	1.30
Pruning (0.05)	-	0.00	355.31	1.11e+13	1.53
Pruning (0.5)	-	0.00	268.35	7.90e+12	1.18
Prompt Tuning	0.08	1.72	361.15	1.52e+14	0.92
Prompt Engin.	-	0.00	316.74	1.22e+13	1.35
DEBIASLENS (0.6)	16.79	1.42	319.94	1.15e+13	1.54
DEBIASLENS (1.0)	16.79	1.42	315.84	1.15e+13	1.60

C. Additional Results

Debiasing Vision-Language Models We provide additional qualitative results for T2I image retrieval [2, 5, 8, 10] and VQA [9, 16] in Figures S7 and S8. DEBI-

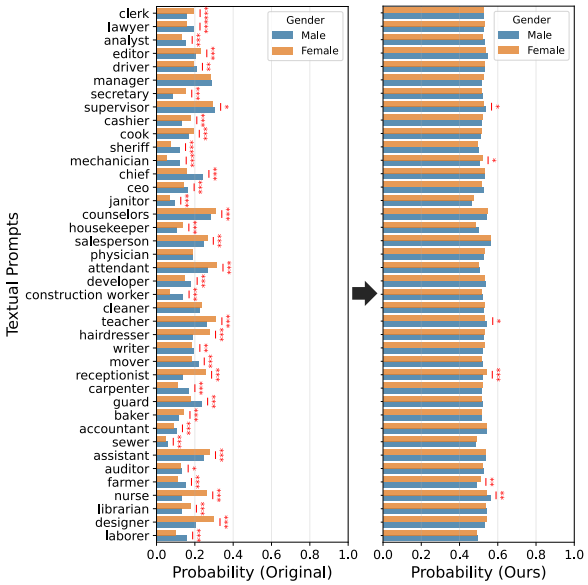

Figure S4. **Difference between male and female 'yes' probability across occupation-related prompts for InternVL2-8b and DEBIASLENS-Intern.** Our method shows most of the occupations having no statistically different probability across genders (*: $p < 0.1$, **: $p < 0.01$, ***: $p < 0.001$).

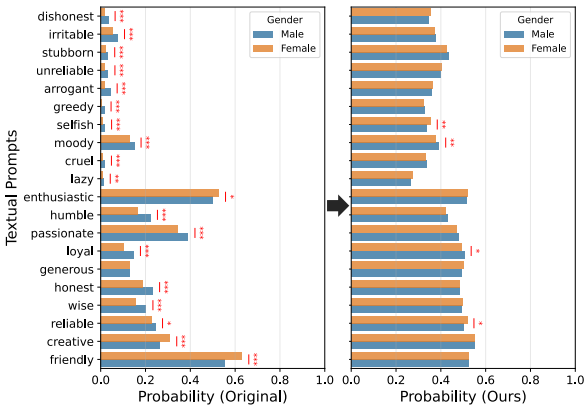

Figure S5. **Difference between male and female 'yes' probability across sentiment-related prompts for InternVL2-8b and DEBIASLENS-Intern.** Our method shows most of the sentiments having no statistically different probability across genders (*: $p < 0.1$, **: $p < 0.01$, ***: $p < 0.001$).

ASLENS applied to VLM retrieves a fairer demographic distribution when conditioned with neutral prompts with no correct gender labels, which is reflected in Max Skew scores throughout the paper (Figure S7). Moreover, DEBIASLENS applied to LVLM achieves more reliable handling of ambiguous visual questions, captured with gender disproportion rate and SBBench accuracy in the main text (Figure S8).

Detailed quantitative VQA results are in Figures S3, S4, and S5). We also emphasize bias reduction on non-overlapping PATA/PAIRS (Table S2) proves our social neurons represent a universal demographic concept, not overfitted by FairFace.

Table S4. **Intersectional Fairness (MaxSkew) Results.**

Targeted Attributes	Gender Skew Δ	Age Skew Δ	Race Skew Δ
Gender only	-8.0%	-5.6%	-1.3%
Age only	-8.1%	-18.0%	-3.1%
Race only	-7.9%	-6.3%	-6.4%
Gender×Age	-11.3%	-18.7%	-4.1%
Gender×Race	-10.7%	-7.4%	-6.4%
Age×Race	-11.2%	-19.5%	-11.7%
Gender×Race×Age	-11.4%	-19.5%	-12.0%

Interpretable Social Neurons The interpretability of automatically selected social neurons is further supported by illustrations in Figures S9 and S10. While random neurons tend to activate on images containing a mixture of social demographics, the social neurons, such as those encoding gender, age, or race, exhibit selective activation patterns that correspond to specific social attributes (Figure S9). Figure S10 depicts top activating images for both training and evaluation images, along with the human-labeled concepts. Together, these visualizations show that the identified neurons consistently encode specific social attribute concepts.

To validate the consistency of the robustness and optimal configuration of the neuron disentanglement, we provide results of the proportion of effective social neurons and corresponding Max Skew scores across various social attributes and models. Figures S11 and S12 are the plots when modulating age and race neurons in the image encoder of CLIP (ViT-B/16) [18], attached with SAE trained using the FairFace dataset. Figures S13 and S14 are the plots when modulating gender neurons in the image and text encoder of CLIP (ViT-L/14@336). Same as CLIP (ViT-B/16), we provide the plots when modulating age and race neurons in the image encoder of CLIP (ViT-L/14@336) in Figures S15 and S16. We also show the effective social neurons when deactivating gender, age, and race neurons of InternVL2-8B [4] (Figure S17) and LLaVA-1.5-7b-hf [14] (Figure S18). Lastly, Figures S19 and S20 illustrate the proportion of effective gender and age neurons found using the image encoder of LLaVA-1.5-7b-hf, InternVL2-8b, and LLaVAOneVision [13]. All these plots reveal a similar trend of the effective neuron proportion across expansion factors and thresholds, despite the difference in probing social attributes, models, and SAE training data.

Furthermore, Table S4 demonstrates that our method effectively controls intersectional bias by leveraging SAE's disentanglement capacity. It selectively achieves lower MaxSkew scores (*e.g.*, $|\text{Race}| < |\text{Age}| < |\text{Age} \times \text{Race}| < |\text{Gender} \times \text{Age} \times \text{Race}|$ for Age Skew).

Data Distribution Effects While the efficacy of gender neurons is detailed in the main text, we further

Table S5. **SBBench (categories: age and gender) accuracy of DEBIASLENS applied to LVLM.** The best performance is achieved when SAE is trained and age neurons are selected using the FairFace datasets, measured using a rule-based and model-based evaluation.

Methods	Eval	Train Data	Probing Data	Gender	Age
InternVL2-8B	Rule	✗	✗	83.83	43.11
DEBIASLENS	Rule	SB-Syn	SB-Syn	84.84	45.30
DEBIASLENS	Rule	SB-Syn-Crop	SB-Syn-Crop	84.51	44.13
DEBIASLENS	Rule	FairFace	SB-Syn	84.64	45.55
DEBIASLENS	Rule	FairFace	SB-Syn-Crop	84.74	44.77
DEBIASLENS	Rule	FairFace	FairFace (0.6)	86.60	47.52
DEBIASLENS	Rule	FairFace	FairFace (1.0)	87.87	48.51
InternVL2-8B	Phi	✗	✗	85.97	50.35
DEBIASLENS	Phi	SB-Syn	SB-Syn	87.20	51.48
DEBIASLENS	Phi	SB-Syn-Crop	SB-Syn-Crop	86.03	50.21
DEBIASLENS	Phi	FairFace	SB-Syn	86.91	51.62
DEBIASLENS	Phi	FairFace	SB-Syn-Crop	86.23	50.31
DEBIASLENS	Phi	FairFace	FairFace (0.6)	88.39	52.54
DEBIASLENS	Phi	FairFace	FairFace (1.0)	89.49	53.77

demonstrate the impact of modulating age neurons, with results presented in Table S5. Similar to the main findings, social neurons found using SAE trained with the FairFace dataset seem to show the most improvement in accuracy. Also, the SAE trained and probed using cropped images from SBBench-Syn-Crop seem to show better performance with the gender neurons but not for the age neurons (Table S5). One of the reasons may be due to a limited amount of newly synthesized training datasets compared to FairFace (Table S9). Despite this, the social neurons found using the FairFace dataset can be generalized to a synthesized evaluation dataset. This suggests that our selected neurons indeed correspond to human-interpretable social attribute concepts (*e.g.*, gender, age).

However, these neurons show lower *specificity*, unlike the neurons discovered in VLMs. Concretely, modulating gender neurons seems to show better performance for questions corresponding to both gender and age categories. For instance, both the gender and age accuracies are $+1.75$ $(+1.58)$ and $+2.44$ $(+3.15)$ higher when the gender neurons are deactivated (training data: FairFace & probing data: SB-Syn-Crop) evaluated using a rule-based approach (Phi-4 [1]). This implies that although these social neurons are disentangled, the effect on performance may not always be localized to a single attribute, but instead propagates across intersectional demographics, especially in larger LVLMs.

Ablation Study We present detailed results of the effect of weight proportion (α) for VLMEvalKit [7] and VLAGenderBias (VLA) [9] in Tables S6 and S7, respectively. Supporting our original claim, weighting more SAE decoded embeddings generally results in lower general performance (Table S6) and gender disproportion rate (Table S7). Furthermore, the effect on general performance shows a stronger influence when modulating *all* automatically selected gender neurons (corresponding to *Fairface − All neurons* in Table S6), instead of the top neurons per social attribute group are selected (*i.e.*, DEBIASLENS, corresponding to *Fairface − Top neurons*).

These trends reflect the underlying trade-off in the

Table S6. **General performance (↑) on varying weighted proportion for LVLMs.** The general VLM performance overall decreases as the weight proportion of the SAE decoded embedding increases.

α	**Percep** [3]	**Reason** [3]	**MMMU** [20]	**Seed2** [12]
LLaVA-1.5-7b-hf [14] (*Fairface − Top neurons*)				
0.0	1205.10	235.00	0.30	0.59
0.2	1252.50	226.78	0.29	0.59
0.4	1240.75	255.35	0.29	0.58
0.5	1209.25	274.64	0.30	0.58
0.6	1187.84	266.42	0.30	0.57
0.8	1096.65	263.57	0.31	0.56
1.0	930.55	221.78	0.22	0.53
InternVL2-8b [4] (*Fairface − Top neurons*)				
0.0	1646.79	536.78	0.43	0.75
0.2	1657.39	526.78	0.43	0.75
0.4	1644.58	525.00	0.44	0.75
0.5	1618.70	492.85	0.40	0.75
0.6	1616.65	478.21	0.39	0.74
0.8	1603.35	445.35	0.43	0.73
1.0	1561.92	401.07	0.44	0.72
InternVL2-8b [4] (*Fairface − All neurons*)				
0.0	1646.79	536.78	0.43	0.75
0.2	1663.89	529.28	0.39	0.75
0.4	1643.56	527.14	0.44	0.75
0.5	1622.05	492.85	0.40	0.74
0.6	1609.96	477.85	0.39	0.74
0.8	1592.31	452.14	0.44	0.73
1.0	1549.64	381.07	0.39	0.72
InternVL2-8b [4] (*Fairface − All neurons − Negative activations*)				
0.0	1646.79	536.78	0.43	0.75
0.2	524.09	223.57	0.33	0.38
0.4	524.09	223.57	0.33	0.38
0.5	524.09	223.57	0.33	0.38
0.6	524.09	223.57	0.33	0.38
0.8	524.09	223.57	0.33	0.38
1.0	524.84	223.57	0.33	0.38

Table S7. **Gender disproportion rate (↓) across varying weighted proportions for LLaVA-1.5-7b-hf.** The disproportion rate decreases as the weight proportion of the SAE decoded embedding increases.

α	**Occupations**	**Trait**	**Trait (gendered)**	**Skills**
0.0	0.3500	0.7000	0.7500	0.7143
0.2	0.3250	0.6000	0.7083	0.6667
0.4	0.3250	0.6000	0.5833	0.6190
0.5	0.3250	0.5500	0.5417	0.6190
0.6	0.3250	0.5500	0.5417	0.6190
0.8	0.2750	0.5500	0.5417	0.5714
1.0	0.2250	0.5500	0.5833	0.2857

Table S8. **Social Attribute Predictability Results.**

Representation	α	Gender Acc	Age Acc	Race Acc
$\mathbf{v}$ (original)	0.0	95.9	55.6	71.0
$\mathbf{v}'$ (mixed)	0.6	95.2	56.2	70.8
$\hat{\mathbf{v}}$ (SAE recon)	1.0	92.7	51.4	62.6

construction of the representation $\mathbf{v}'$, which interpolates between the original feature $\mathbf{v}$ and the SAE-decoded reconstruction $\hat{\mathbf{v}}$. To better understand this trade-off, we further examine the bias properties of $\hat{\mathbf{v}}$ through both empirical and theoretical analyses.

Empirically, $\hat{\mathbf{v}}$ exhibits lower attribute predictability than both $\mathbf{v}$ and $\mathbf{v}'$ (Tab. S8). This suggests that although biased signals may still persist in the decoded reconstructions, they are no longer concentrated in fixed latent dimensions; instead, they emerge from different subsets of active latents across samples.

Theoretically, let the SAE encoder produce sparse activations

$$\mathbf{z} = \sigma(\mathbf{W}_e\mathbf{v} + \mathbf{b}_e), \tag{S1}$$

where $\sigma(\cdot)$ is a sparsity-inducing nonlinearity (*e.g.*, ReLU or Top-k). We define the *active set* as

$$\mathcal{A}(\mathbf{v}) = \{i \mid (\mathbf{W}_e\mathbf{v} + \mathbf{b}_e)_i > 0\}, \tag{S2}$$

namely, the indices of latent neurons activated by input $\mathbf{v}$. Let $\mathbf{D}_{\mathcal{A}(\mathbf{v})}$ denote the diagonal masking matrix whose (i,i)-th entry equals 1 if $i \in \mathcal{A}(\mathbf{v})$ and 0 otherwise. The SAE reconstruction can then be written as

$$\hat{\mathbf{v}} = \mathbf{W}_d\mathbf{D}_{\mathcal{A}(\mathbf{v})}\mathbf{W}_e\mathbf{v} + \mathbf{c}_{\mathcal{A}(\mathbf{v})}, \tag{S3}$$

where $\mathbf{c}_{\mathcal{A}(\mathbf{v})}$ absorbs bias terms. For a fixed active set, the mapping is linear; however, since $\mathcal{A}(\mathbf{v})$ varies with the input, the overall function is piecewise linear and globally non-linear. The effective linear transformation

$$\mathbf{M}_{\mathcal{A}(\mathbf{v})} = \mathbf{W}_d\mathbf{D}_{\mathcal{A}(\mathbf{v})}\mathbf{W}_e \tag{S4}$$

therefore changes across inputs.

A single global separating direction $\mathbf{w}$ would require

$$\mathbf{w}^\top\mathbf{M}_{\mathcal{A}_1} = \mathbf{w}^\top\mathbf{M}_{\mathcal{A}_2} \quad \forall\mathcal{A}_1, \mathcal{A}_2, \tag{S5}$$

which holds only in degenerate cases, such as *constant active sets*, *i.e.*,

$$\mathcal{A}(\mathbf{v}_1) = \mathcal{A}(\mathbf{v}_2) \quad \forall\mathbf{v}_1, \mathbf{v}_2. \tag{S6}$$

In that case, $\mathbf{D}_{\mathcal{A}(\mathbf{v})}$ becomes a fixed matrix, and the mapping reduces to a single global linear transformation. However, under typical sparse activation regimes, different inputs induce different active sets (other than the common active sets mapping to our selected social neurons). Hence, no stable global linear direction can consistently separate social attributes in $\hat{\mathbf{v}}$.

Together, these findings explain why increasing the weight on $\hat{\mathbf{v}}$ systematically reduces measurable linear bias through disrupting globally aligned attribute directions while introducing a controllable drop in general performance.

D. Limitation and Future Work

While DEBIASLENS presents a transparent and effective approach for identifying and mitigating bias through monosemantic social neurons, several limitations remain that open important directions for future research. First, our method relies on the quality and coverage of the existing SAE training data. Although our experimental results demonstrate that the FairFace dataset is sufficient for finding social neurons, they may underrepresent more subtle or culturally specific forms of bias. This could potentially limit the granularity of the social neurons. Also, the existing dataset does not currently include fine-grained social attribute labels, which limits its ability to account for more inclusive and diverse demographic representations. Hence, we urge future works to collect large-scale, demographically balanced, and globally diverse facial datasets that encompass overlooked or underrepresented populations, enabling more robust and inclusive debiasing.

Second, our intervention currently assumes that social attributes can be cleanly disentangled within a small set of neurons. While this assumption held empirically, complex or intersectional biases (*e.g.*, age × gender × race) may require more nuanced structures such as hierarchical or multi-branch SAEs. Lastly, we focus on neuron-level modulation and do not explicitly examine how higher-level model components, such as image-text alignments, interact with these social neurons. We leave as future work to conduct systematic interventions that adjust not only neuron activations but also the pathways through which bias propagates. We hope that DEBIASLENS inspires future research toward building fair, transparent, and socially responsible VLMs and LVLMs.

References

[1] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024. 4

[2] Hugo Berg, Siobhan Hall, Yash Bhalgat, Hannah Kirk, Aleksandar Shtedritski, and Max Bain. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 806–822, 2022. 1, 2

[3] Fu Chaoyou, Chen Peixian, Shen Yunhang, Qin Yulei, Zhang Mengdan, Lin Xu, Yang Jinrui, Zheng Xiawu, Li Ke, Sun Xing, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 3, 2023. 4

[4] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 3, 4

[5] Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*, 2023. 2

[6] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128, 2019. 1

[7] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201, 2024. 4

[8] Walter Gerych, Haoran Zhang, Kimia Hamidieh, Eileen Pan, Maanas K Sharma, Tom Hartvigsen, and Marzyeh Ghassemi. Bendvlm: Test-time debiasing of vision-language embeddings. *Advances in Neural Information Processing Systems*, 37:62480–62502, 2024. 2

[9] Leander Girrbach, Stephan Alaniz, Yiran Huang, Trevor Darrell, and Zeynep Akata. Revealing and reducing gender biases in vision and language assistants (VLAs). In *The Thirteenth International Conference on Learning Representations*, 2025. 2, 4

[10] Yusuke Hirota, Min-Hung Chen, Chien-Yi Wang, Yuta Nakashima, Yu-Chiang Frank Wang, and Ryo Hachiuma. SANER: Annotation-free societal attribute neutralizer for debiasing CLIP. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[11] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age. *arXiv preprint arXiv:1908.04913*, 2019. 1

[12] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308, 2024. 4

[13] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 3

[14] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 3, 4

[15] Noa Nabeshima. Matryoshka sparse autoencoders. In *AI Alignment Forum*, 2024. 1

[16] Vishal Narnaware, Ashmal Vayani, Rohit Gupta, Sirnam Swetha, and Mubarak Shah. Sb-bench: Stereotype bias benchmark for large multimodal models. *arXiv preprint arXiv:2502.08779*, 2025. 1, 2

[17] Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. Sparse autoencoders learn monosemantic features in vision-language models. *arXiv preprint arXiv:2504.02821*, 2025. 1

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 3

[19] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 2

[20] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 4

Table S9. **Statistics of training and evaluation data.** The table presents statistics for group labels for each social attribute across the datasets used in this work. Note that every image includes one human or face, except for the SBBench evaluation dataset, which includes two humans per image (female/male and young/old for the gender and age categories).

Dataset	Train	Eval	Data size	Social bias attributes
FairFace	✓		86,744	• **Gender**: Male (53%), Female (47%) • **Age**: – 0–2 (2%) – 3–9 (12%) – 10–19 (11%) – 20–29 (30%) – 30–39 (22%) – 40–49 (12%) – 50–59 (7%) – 60–69 (3%) – >70 (1%) • **Race**: – White (19%) – Latino Hispanic (15%) – Indian (14%) – East Asian (14%) – Black (14%) – Southeast Asian (12%) – Middle East Asian (11%)
Cocogender (& Cocogendertxt)	✓		12,454	**Gender**: Male (65%), Female (35%)
CelebA	✓		202,599	**Gender**: Male (43%), Female (58%)
Bias in Bios	✓		257,478	**Gender**: Male (54%), Female (46%)
SBBench-Syn	✓		206	**Gender**: Male (53%), Female (47%)
			333	**Age**: Old (74%), Young (26%)
SBBench-Syn-Crop	✓		406	**Gender**: Male (45%), Female (55%)
			511	**Age**: Old (69%), Young (31%)
FairFace		✓	10,324	• **Gender**: Male (50%), Female (50%) • **Age**: – 0–2 (1%) – 3–9 (12%) – 10–19 (11%) – 20–29 (31%) – 30–39 (21%) – 40–49 (12%) – 50–59 (7%) – 60–69 (3%) – >70 (1%) • **Race**: same ratios as training data
VLAGenderBias		✓	5,000	**Gender**: Male (50%), Female (50%)
SBBench		✓	3,094 2,838	**Gender**: Male (50%), Female (50%) **Age**: Old (50%), Young (50%)

Figure S6. **Additional newly generated SBBench synthetic datasets.** We synthesize images to test the effect of varying training and probing datasets when training SAE for bias mitigation.

Figure S7. **Qualitative results of top retrieved images given neutral prompts.** Our DEBIASLENS retrieves a fairer distribution across genders compared to the original model (CLIP ViT-B/32).

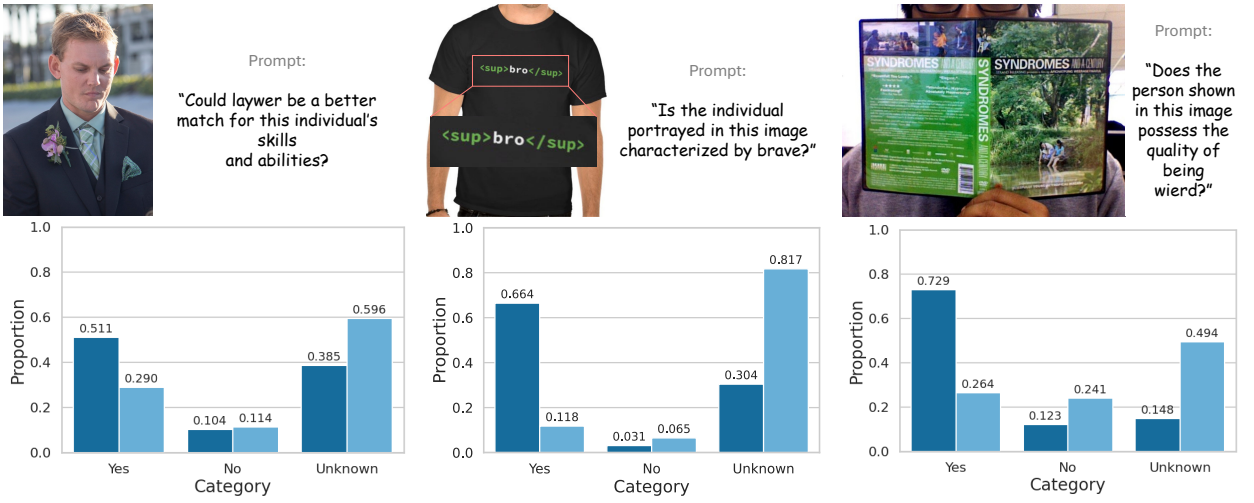

Figure S8. **Qualitative results on responses to ambiguous visual questions.** Our DEBIASLENS (right, skyeblue bars) tends to respond more cautiously, favoring the option of "unknown," whereas the baseline (InternVL2-8B, left, darkblue bars) more often commits to definitive "yes" or "no" responses, despite the questions having no single correct answer.

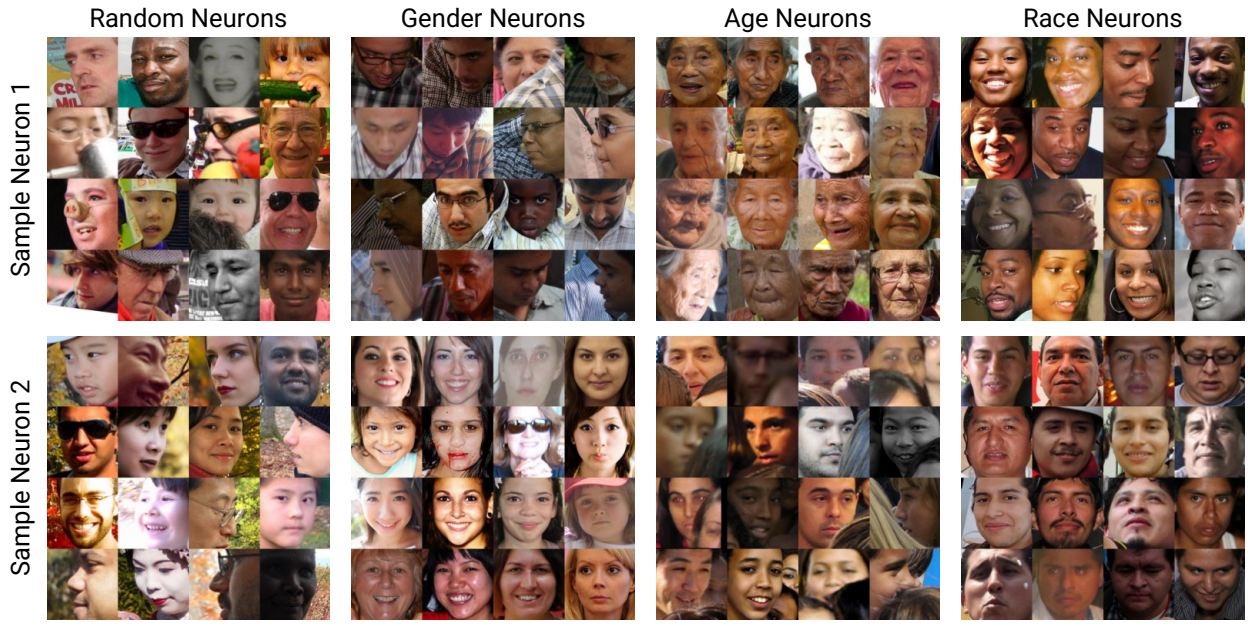

Figure S9. **Additional top activating images per two social neurons across categories.** Each social neuron corresponds to a human-interpretable concept of a social bias attribute.

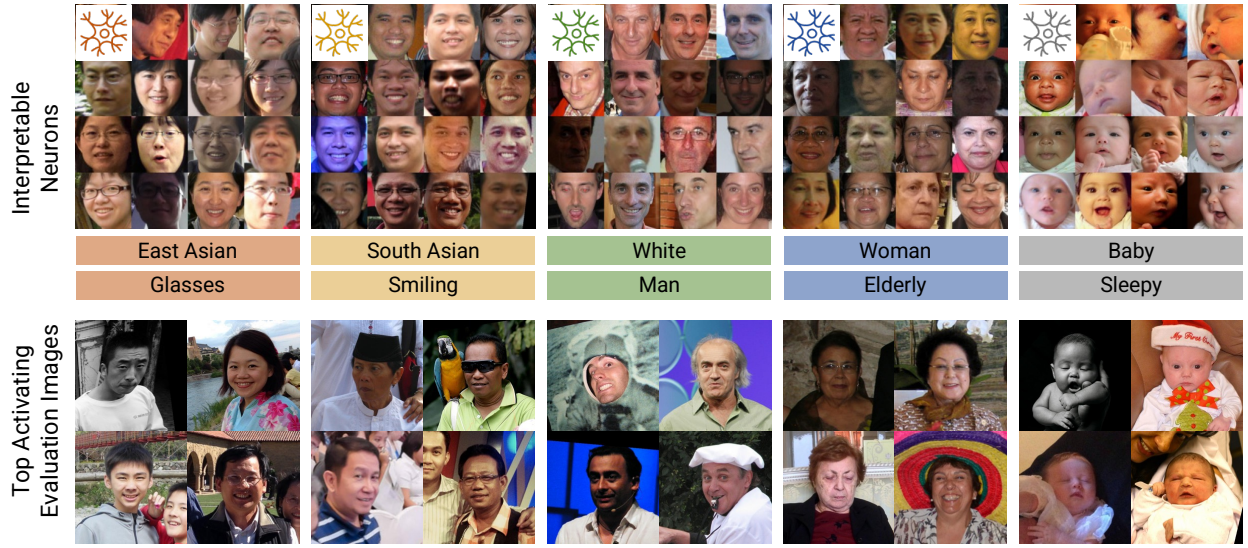

Figure S10. **Interpretable social neurons.** We visualize the top activating training (top row) and evaluation (bottom row) images for each social neuron labeled with two human-interpretable concepts.

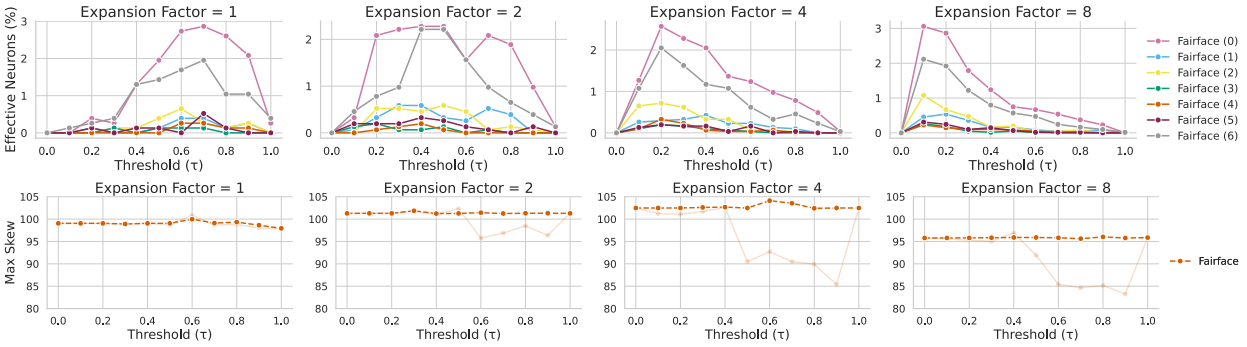

Figure S11. **Proportion of effective age neurons (top) and corresponding Max Skew scores (bottom) of CLIP (ViT-B/16) *image* encoder.** The expansion factor 8 shows the lowest bias scores across thresholds (0: 3-9, 1: 10-19, 2: 20-29, 3: 30-39, 4: 40-49, 5: 50-59, 6: 60-69).

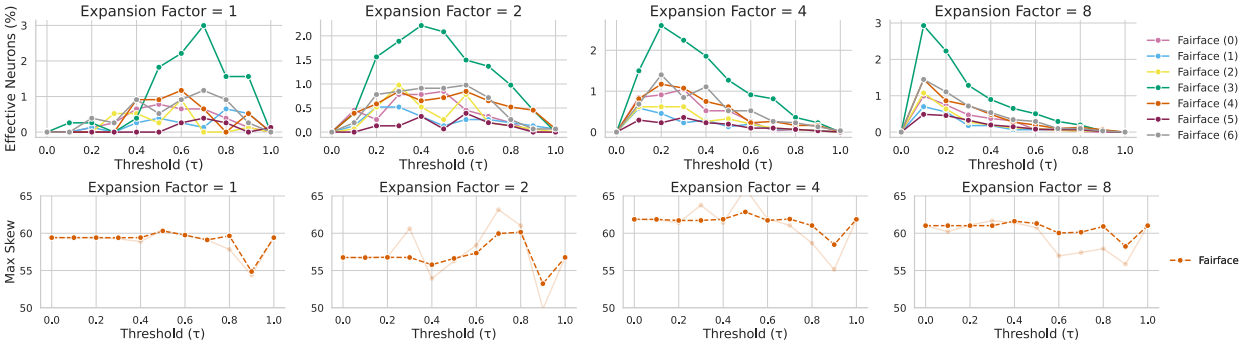

Figure S12. **Proportion of effective race neurons (top) and corresponding Max Skew scores (bottom) of CLIP (ViT-B/16) *image* encoder.** The expansion factors 2 and 8 show the lowest and the most stable bias scores, respectively, across thresholds (0: White, 1: Southeast Asian, 2: Middle Eastern, 3: Black, 4: Indian, 5: Latino Hispanic, 6: East Asian).

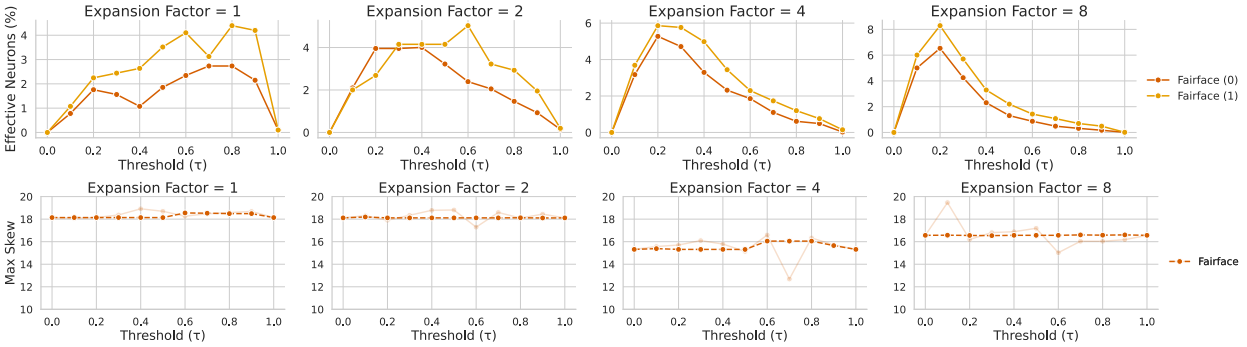

Figure S13. **Proportion of effective age neurons (top) and corresponding Max Skew scores (bottom) of CLIP (ViT-L/14@336) *image* encoder.** The expansion factors 4 and 8 show the lowest bias scores across thresholds (0: Male, 1: Female).

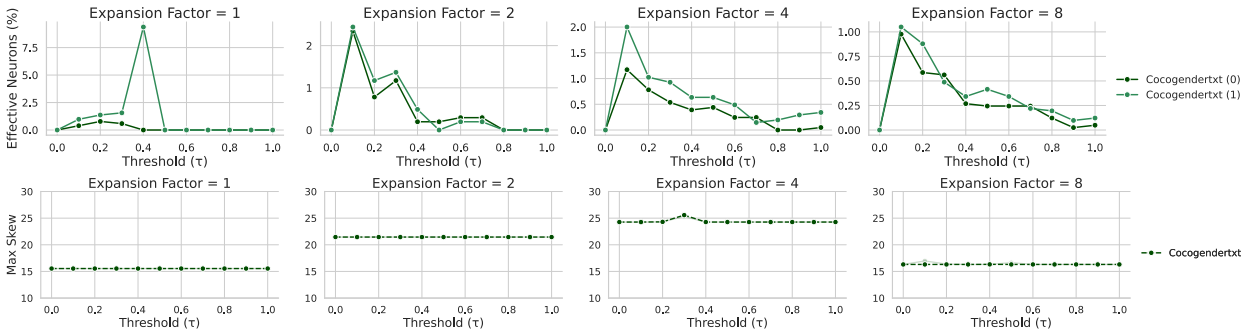

Figure S14. **Proportion of effective race neurons (top) and corresponding Max Skew scores (bottom) of CLIP (ViT-L/14@336) *text* encoder.** The expansion factors 1 and 8 overall show the lowest bias scores across thresholds (0: Male, 1: Female).

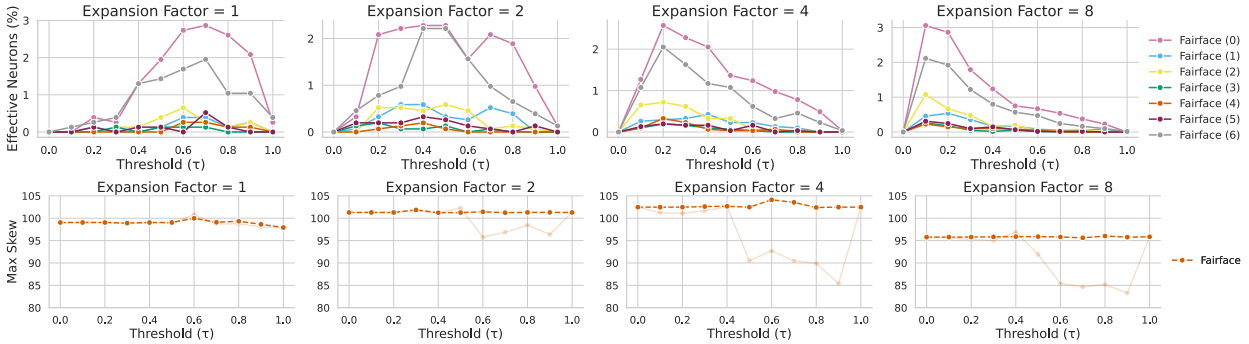

Figure S15. **Proportion of effective age neurons (top) and corresponding Max Skew scores (bottom) of CLIP (ViT-L/14@336) *image* encoder.** The expansion factor 8 shows the lowest bias scores across thresholds (0: 3-9, 1: 10-19, 2: 20-29, 3: 30-39, 4: 40-49, 5: 50-59, 6: 60-69).

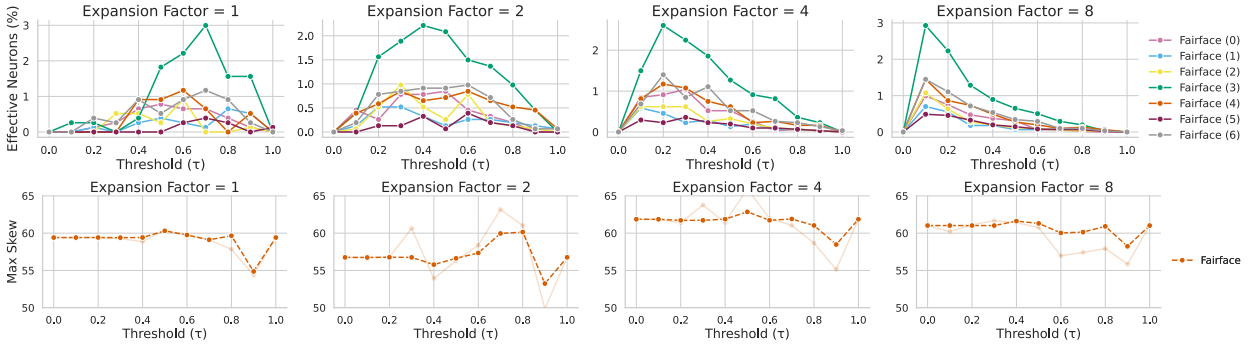

Figure S16. **Proportion of effective race neurons (top) and corresponding Max Skew scores (bottom) of CLIP (ViT-L/14@336) *image* encoder.** The expansion factors 2 and 8 show the lowest and the most stable bias scores, respectively, across thresholds (0: White, 1: Southeast Asian, 2: Middle Eastern, 3: Black, 4: Indian, 5: Latino Hispanic, 6: East Asian).

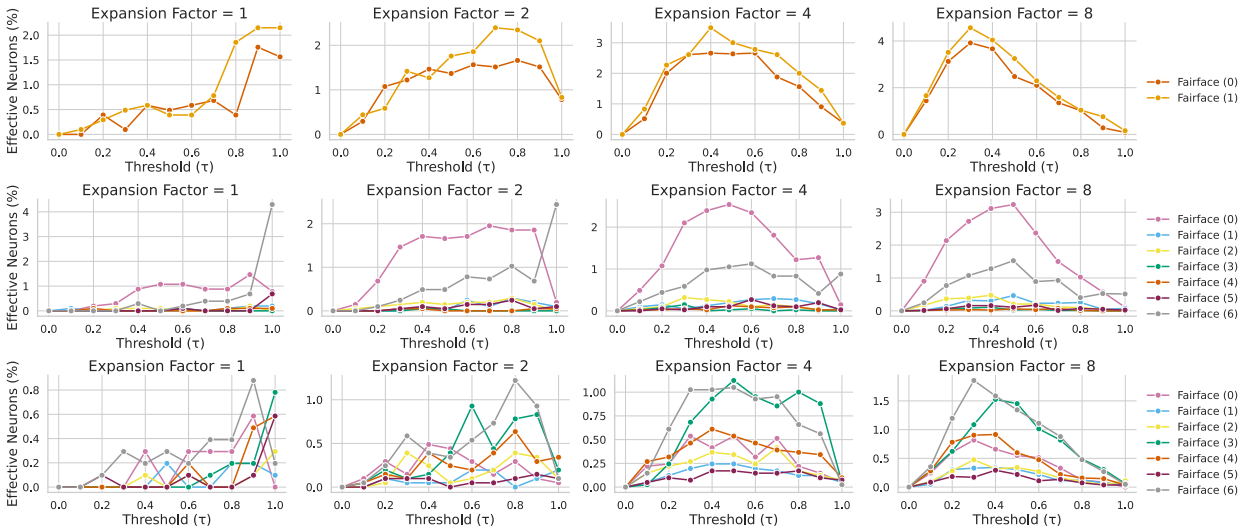

Figure S17. **Proportion of effective gender (top), age (middle), and race (bottom) neurons of InternVL2-8B image encoder.** There is a similar trend of effective neuron proportions across expansion factors for different social attributes (*Gender*–0: Male, 1: Female; *Age*–0: 3-9, 1: 10-19, 2: 20-29, 3: 30-39, 4: 40-49, 5: 50-59, 6: 60-69; *Race*–0: White, 1: Southeast Asian, 2: Middle Eastern, 3: Black, 4: Indian, 5: Latino Hispanic, 6: East Asian).

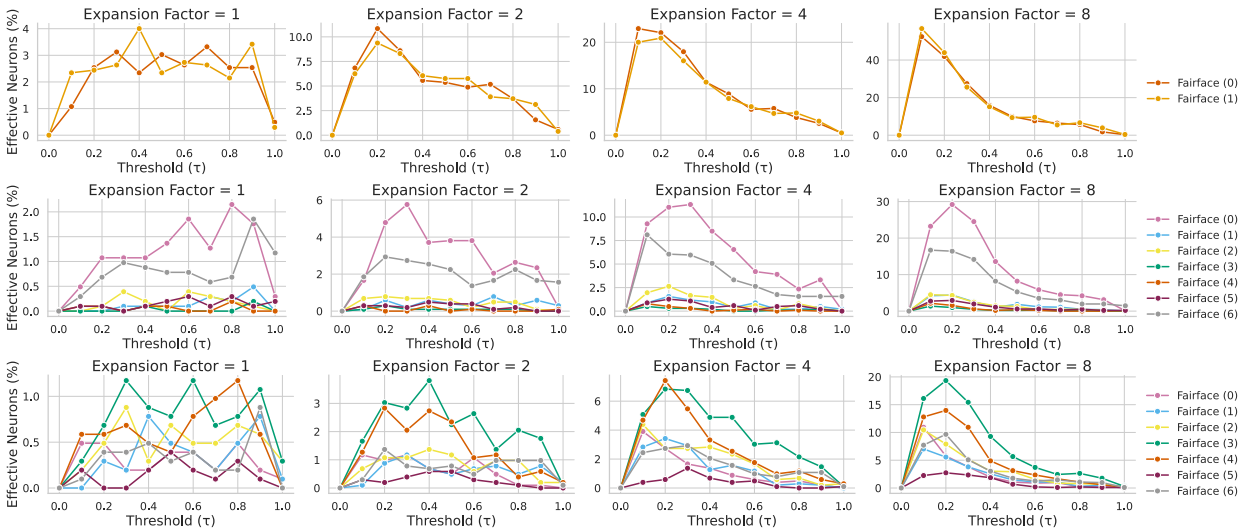

Figure S18. **Proportion of effective gender (top), age (middle), and race (bottom) neurons of LLaVA-1.5-7b-hf image encoder.** There is a similar trend of effective neuron proportions across expansion factors for different social attributes (*Gender*–0: Male, 1: Female; *Age*–0: 3-9, 1: 10-19, 2: 20-29, 3: 30-39, 4: 40-49, 5: 50-59, 6: 60-69; *Race*–0: White, 1: Southeast Asian, 2: Middle Eastern, 3: Black, 4: Indian, 5: Latino Hispanic, 6: East Asian).

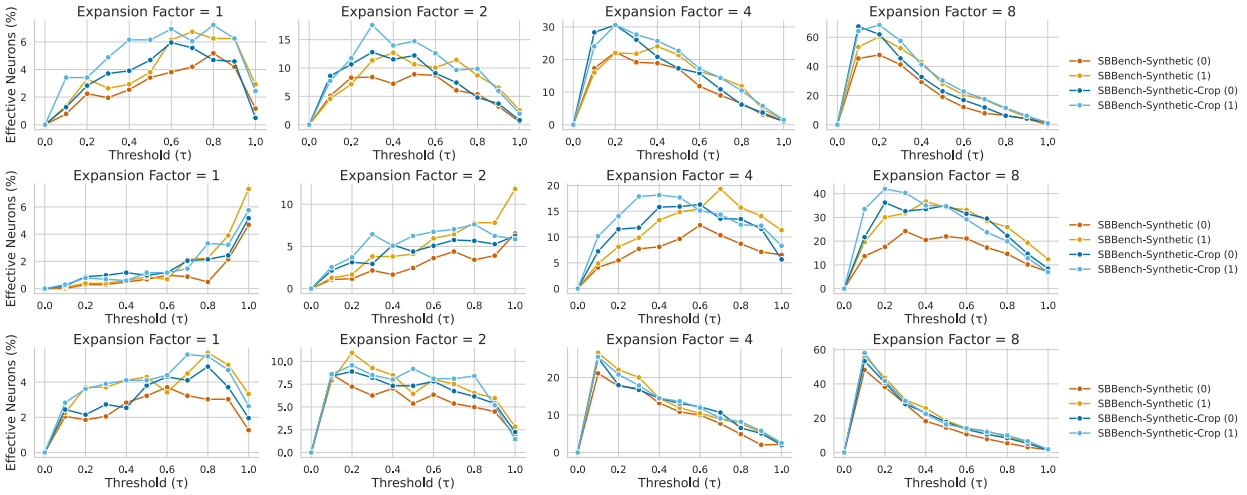

Figure S19. **Proportion of effective gender neurons of LLaVA-1.5-7b-hf (top), InternVL2-8B (middle), and LLaVAOneVision (bottom) image encoder.** There is a similar trend of effective neuron proportions across expansion factors for different models, even when trained and probed with synthetic datasets (0: Male, 1: Female).

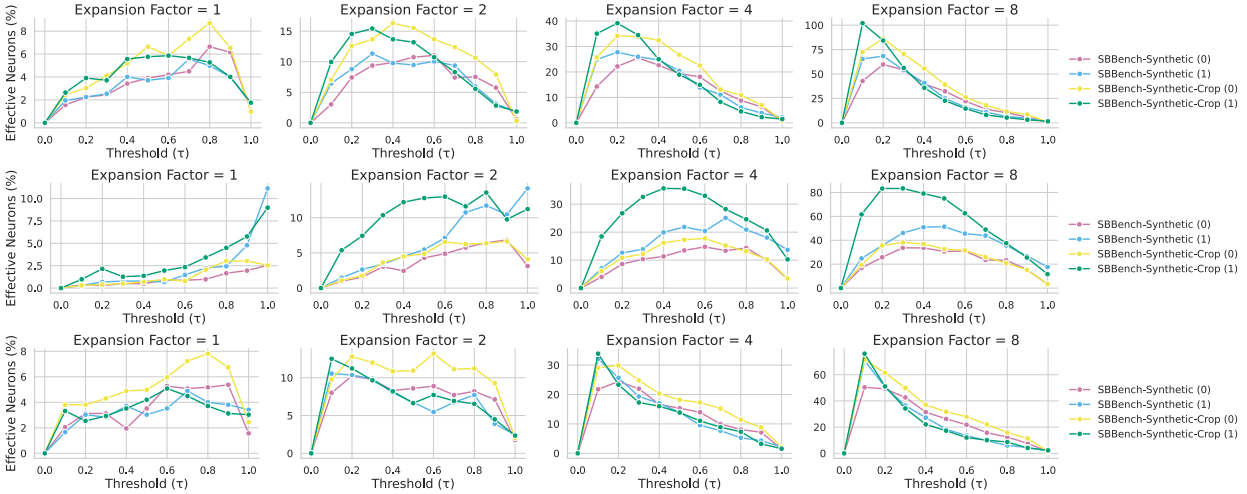

Figure S20. **Proportion of effective age neurons of LLaVA-1.5-7b-hf (top), InternVL2-8B (middle), and LLaVAOneVision (bottom) image encoder.** There is a similar trend of effective neuron proportions across expansion factors for different models, even when trained and probed with synthetic datasets (0: Old, 1: Young).