CVPR JUNE 3-7, 2026 DENVER COLORADO

UNIVERSITÀ DI TRENTO · FONDAZIONE BRUNO KESSLER · TOYOTA MOTOR EUROPE NV/SA

SEM: Sparse Embedding Modulation for Post-Hoc Debiasing of Vision-Language Models

github.com/mardgui/SEM

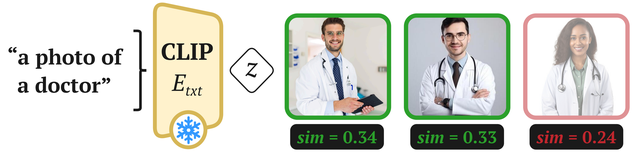
arxiv.org/pdf/2603.19028

py.md/SEM-CVPR

Quentin Guimard[1,*] Federico Bartsch[1,*] Simone Caldarella[1] Rahaf Aljundi[3] Elisa Ricci[1,2] Massimiliano Mancini[1]

[1]University of Trento [2]Fondazione Bruno Kessler [3]Toyota Motor Europe

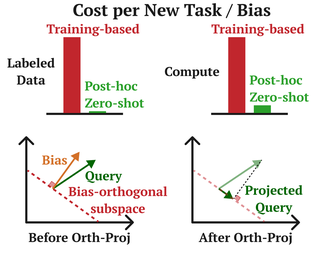
Vision-Language Models are Biased



Spurious correlations and **social biases** are inherited from web-scale training data and affect tasks like **text-to-image retrieval** and **zero-shot classification**.

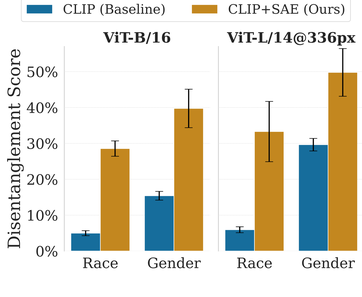
Existing Debiasing Methods are Flawed

- **Training-based methods** require **heavy compute** and **new annotations** for every new task or bias.
- **Prior zero-shot methods damage semantics** because they operate on **dense, entangled** embeddings.

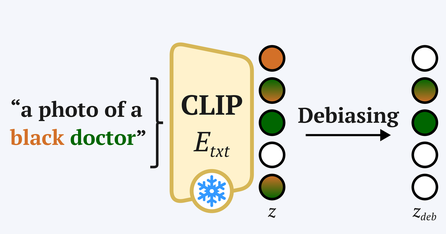


Why Sparse Autoencoders (SAEs)?

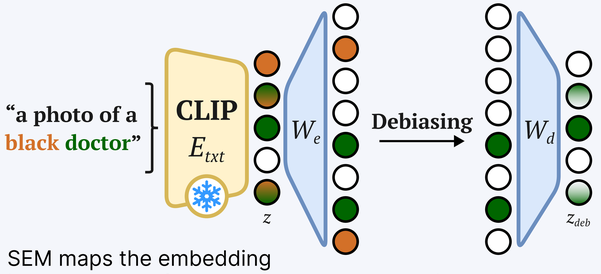
- **Measuring feature entanglement:** *Can predicting a person's profession accidentally reveal their race or gender?*
- **SAEs** improve feature **disentanglement** by up to **5.7x**.



SAEs Decompose Entangled Embeddings for Precise Intervention

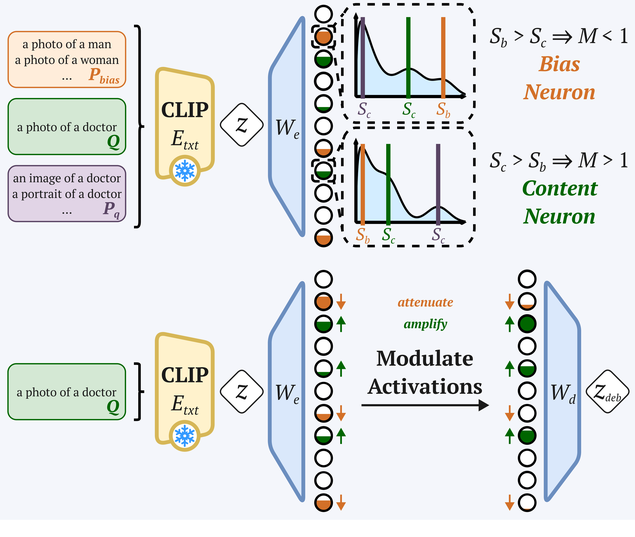


Standard methods operate directly on the **dense, entangled** CLIP embedding space.



SEM maps the embedding into a **sparse, disentangled** latent space via an SAE. This enables a **precise intervention** on specific features.

SEM Scores SAE Neurons to Modulate SAE Activations at Test-Time



1. Scoring SAE Neurons

Neuron activations are ranked against a set of diverse, neutral prompts.

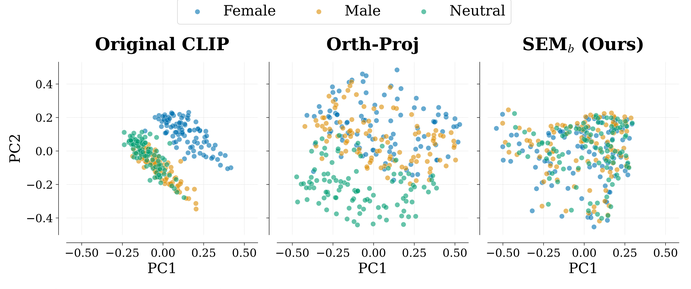
- **Content Relevance** (S_c): Activation rank for the target queries (Q, P_q).
- **Bias Sensitivity** (S_b): Activation rank for a set of bias prompts (P_{bias}).

2. Modulating SAE Activations

Activations are steered using a modulation coefficient M, adapting to available prompts:

- **Bias-Agnostic** (SEM_i): Uses only augmented task queries P_q ($M_i = S_c^2$).
- **Bias-Aware** (SEM_b): Uses bias prompts P_{bias} and the query Q ($M_b = (1 + S_c - S_b)^2$).
- **Full** (SEM_{bi}): Uses bias prompts P_{bias} and augmented task queries P_q ($M_{bi} = M_b$).

SEM Disentangles Bias from Content



Unlike Orth-Proj which disrupts the underlying data structure, SEM cleanly unifies clusters to **neutralize bias** (BN) while **preserving the original content** (CP).

Method	CP (↑)	BN (↑)
Orth-Proj	0.415	0.916
SEM_b (Ours)	**0.878**	**0.974**

SEM Achieves SOTA Fairness and Robustness

- **Zero-Shot Robustness:** SEM improves worst-group accuracy across zero-shot classification benchmarks.
- **Retrieval Fairness:** SEM achieves state-of-the-art fairness for demographic attributes in text-to-image retrieval.
- **Modularity:** SEM combines with existing methods like BendVLM to further improve fairness and robustness.

Method	Waterbirds WG Acc. (↑)	FairFace Race KL (↓)
Base CLIP	0.396	0.244
Bias-agnostic + input-specific prompts		
RoboShot	0.485	0.304
SEM_i (Ours)	**0.523**	**0.146**
Bias prompts only		
Orth-Proj	0.477	0.213
SEM_b (Ours)	**0.631**	**0.194**
Bias prompts + input-specific prompts		
Orth-Cali	0.482	0.169
SEM_{bi} (Ours)	**0.684**	**0.155**
Bias + input-spec. prompts + images		
BendVLM	0.421	0.106
BendSEM_{bi}	**0.741**	**0.063**

Results for ViT-L/14@336px. More experiments in the paper.