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

Supplementary Material

A. SAE Training Details

As outlined in the main paper, we train a separate Sparse Autoencoder for each CLIP backbone (ViT-B/16 and ViT-L/14@336px). Below, we detail the architecture, objective, and optimization hyperparameters used.

Architecture and Objective. We employ the Matryoshka Sparse Autoencoder (MSAE) architecture proposed by Zaigrajew et al. [30]. Unlike standard SAEs, the MSAE is designed to learn hierarchically structured features. We set the total latent dimensionality to 16384. The model is trained to minimize the reconstruction error (MSE) computed at specific nested granularities, specifically $g \in \{256, 512\}$. To enforce the hierarchical structure, we apply Reverse Weighting (RW) to the loss function. This weighting scheme assigns higher importance to errors at lower granularities (*i.e.*, the top-256 features), ensuring that the most salient semantic concepts are captured by the earlier latent dimensions before finer-grained details are learned in the higher dimensions.

Initialization. We use a learned centering parameter b_{pre}, which is subtracted from the input embedding before encoding and added back after decoding. This parameter is initialized to the geometric mean of the training embeddings. For the weights, we follow standard SAE best practices: the decoder weights W_d are initialized using Kaiming uniform initialization and scaled, while the encoder weights W_e are initialized as the transpose of the decoder weights ($W_e = W_d^T$). The encoder bias is initialized to zero.

Optimization and Data. All models are optimized using the AdamW optimizer with a learning rate of 1×10^{-4} and a batch size of 2048. We utilize a linear-decay learning rate scheduler, which maintains a constant learning rate for the initial portion of training before decaying linearly to zero. We use the CC12M-cleaned dataset [25], split into 90% for training and 10% for validation.

Computational Resources. Training was performed on a shared high-performance cluster node equipped with a single NVIDIA A100 GPU (64GB HBM2e), 8 CPU cores, and 128 GB of RAM. Under this setup, training a single SAE takes approximately 1.5 hours.

B. Details on Disentanglement Study

In this section, we provide the full experimental details and results for the disentanglement study presented in Sec. 3.1 of the main paper.

B.1. Experimental Setup

Dataset Generation. To construct the probing dataset, we combine a set of templates with specific attributes. We use:

- **Bias Attributes:**
  - *Gender* (2 classes): 'male', 'female'.
  - *Race* (7 classes): 'Black', 'East Asian', 'Indian', 'Latino/Hispanic', 'Middle Eastern', 'Southeast Asian', 'White'.
- **Main Attribute:** *Profession* (100 classes). The complete list is provided in Tab. 5.
- **Templates:** 20 diverse prompt templates (listed in Tab. 6) that vary syntactic structure while retaining the semantic content slots for {`bias`} and {`profession`}.

We generate all possible combinations of (Template $\times$ Bias $\times$ Profession), resulting in a balanced dataset where every profession is equally represented across all bias classes.

Probing Methodology. We use Logistic Regression classifiers as linear probes. To ensure a rigorous evaluation:

1. **Data Split:** We use 5-fold stratified cross-validation. The splits are stratified by the main task (profession) to ensure all classes are represented in training and testing.
2. **Scaling:** Feature inputs (CLIP embeddings or SAE latents) are standardized (zero mean, unit variance) using statistics computed on the training set of each fold.
3. **Training:** The probes are trained using the L-BFGS solver with a maximum of 1000 iterations to ensure convergence.

B.2. Two-Stage Disentanglement Experiment

We use a sequential probing setup to quantify conceptual entanglement:

1. **Stage 1 (Main Task):** We train a probe P_p to predict the 'profession' label from the features. We report its accuracy as acc_p.
2. **Control (Bias Task):** We train a probe P_b to predict the 'bias' label directly from the features. We report its accuracy as acc_b. This serves as an upper bound on the extractable bias information.
3. **Stage 2 (Entanglement):** We freeze P_p and use it to generate logits for the test set. We then train a second probe $P_{b \leftarrow p}$ to predict the 'bias' label using *only* these profession logits as input. We report its accuracy as $acc_{b \leftarrow p}$.

A high $acc_{b \leftarrow p}$ indicates that the profession classifier relies on features that are entangled with the bias attribute. Ideally, if the embeddings are perfectly disentangled, the profession

Table 5. Complete list of 100 professions used in both the disentanglement and qualitative studies.

Accountant	Actor	Architect	Astronaut	Audiologist	Author	Baker	Barber
Biologist	Blacksmith	Bricklayer	Bus driver	Butcher	Carpenter	Chef	Chemist
Civil engineer	Cleaner	Clerk	Coach	Comedian	Computer programmer	Construction worker	Consultant
Counselor	Dancer	Dentist	Designer	Detective	Dietitian	DJ	Doctor
Driver	Economist	Editor	Electrician	Engineer	Entrepreneur	Farmer	Firefighter
Florist	Graphic designer	Hairdresser	Historian	Interpreter	Journalist	Judge	Lawyer
Librarian	Magician	Makeup artist	Manager	Marine biologist	Mathematician	Mechanic	Model
Musician	Nanny	Nurse	Nutritionist	Optician	Painter	Paramedic	Pastry chef
Pediatrician	Pharmacist	Photographer	Physicist	Pilot	Plumber	Police officer	Politician
Professor	Psychologist	Real estate agent	Receptionist	Recruiter	Reporter	Researcher	Sailor
Salesperson	Scientist	Security guard	Singer	Social worker	Software developer	Statistician	Surgeon
Surveyor	Teacher	Technician	Therapist	Tour guide	Translator	Vet	Videographer
Waiter	Web developer	Writer	Zoologist				

Table 6. Prompt templates used for the disentanglement study.

Templates with {bias} and {profession} placeholders	
`A photo of a {bias} {profession}.`	`A {bias} {profession} at work.`
`An image of a {bias} {profession}.`	`An illustration of a {bias} {profession}.`
`A {bias} {profession}.`	`A studio shot of a {bias} {profession}.`
`A portrait of a {bias} {profession}.`	`A {bias} professional who works as a {profession}.`
`This is a {bias} {profession}.`	`A close-up of a {bias} {profession}.`
`Here is a {bias} {profession}.`	`A {bias} {profession} on the job.`
`A picture depicting a {bias} {profession}.`	`A {bias} individual employed as a {profession}.`
`A {bias} person who is a {profession}.`	`We can see a {bias} {profession} here.`
`A {bias} person working as a {profession}.`	`A {bias} {profession} posing for the camera.`
`This image shows a {bias} {profession}.`	`A depiction of a {bias} {profession}.`

classifier should make its predictions without relying on any gender-related information.

B.3. Full Results

Tab. 7 presents the detailed accuracies for all stages. As noted in the main paper, both CLIP and SAE representations allow for near-perfect performance on the main task ($acc_p > 0.99$). However, the sequential probe accuracy ($acc_{b \leftarrow p}$) is significantly lower for the SAE latent space compared to the dense CLIP embedding space. This quantitative gap drives the higher Disentanglement Score (D) reported in the main paper, confirming that the SAE effectively separates bias information from task-relevant semantics.

C. Details on Qualitative Study

In Sec. 4.2 of the main paper, we presented a qualitative analysis of conceptual entanglement. Here, we provide the detailed experimental setup, dataset construction, and formal definitions of the metrics used for that study.

C.1. Dataset Construction

To study the entanglement of bias and content, we constructed a targeted dataset of 100 profession prompts. The professions are the same as those listed in Tab. 5 (*e.g.*, accountant, doctor, engineer).

For each profession p, we generate three prompt variants:

1. **Female:** "A photo of a female {`profession`}."
2. **Male:** "A photo of a male {`profession`}."
3. **Neutral:** "A photo of a {`profession`}."

This results in a total of 300 prompts. This controlled set allows us to isolate the effect of the gender attribute on the profession semantics.

C.2. Methodology

Models and Baselines. We compute embeddings for all 300 prompts using the ViT-L/14@336px backbone, matching the quantitative results reported in Sec. 4.3. We compare three sets of embeddings:

- **BASE CLIP:** The original, unperturbed embeddings.
- **ORTH-PROJ [9]:** Embeddings debiased by projecting out the gender subspace.
- **SEM$_b$:** Embeddings debiased using our proposed sparse modulation. For this specific experiment, to ensure maximum content preservation, the content score S_{concept} was computed using the *neutral* profession prompt as the reference.

PCA Visualization. To generate the visualization in the main paper (Fig. 4), we apply Principal Component Analysis (PCA) to the set of 300 embeddings for each method independently. We project the embeddings onto their first two principal components. This allows us to visualize the geo-

Table 7. **Full Probing Results.** Mean accuracies for profession prediction (acc_p), direct bias prediction (acc_b), and sequential entanglement probe ($acc_{b\leftarrow p}$) across Race and Gender settings. Lower entanglement ($acc_{b\leftarrow p}$) indicates better disentanglement.

Method	ViT-B/16						ViT-L/14@336px					
	Race			Gender			Race			Gender		
	$acc_p(\uparrow)$	$acc_b(\uparrow)$	$acc_{b\leftarrow p}(\downarrow)$	$acc_p(\uparrow)$	$acc_b(\uparrow)$	$acc_{b\leftarrow p}(\downarrow)$	$acc_p(\uparrow)$	$acc_b(\uparrow)$	$acc_{b\leftarrow p}(\downarrow)$	$acc_p(\uparrow)$	$acc_b(\uparrow)$	$acc_{b\leftarrow p}(\downarrow)$
Base CLIP	1.000	1.000	0.957	1.000	1.000	0.923	1.000	1.000	0.949	1.000	1.000	0.852
SAE	0.996	1.000	**0.755**	0.995	0.997	**0.800**	0.994	0.998	**0.710**	0.993	0.996	**0.748**

metric structure of the 'male', 'female', and 'neutral' clusters for each method without the projection being dominated by the global variance of the original space.

Metric Definitions. To quantify the visual observations, we defined two metrics based on cosine similarity. Let $z_p^{\text{neut, orig}}$ denote the *original Base CLIP* embedding for the neutral prompt of profession p. Let z_p^g denote the *debiased* embeddings for profession p with gender attribute $g \in \mathcal{G} = \{\text{male}, \text{female}\}$.

- **Content Preservation (CP):** This metric measures how well the gendered embeddings retain the semantics of the original *neutral* concept after debiasing. It is computed as the average cosine similarity between the gendered embeddings and the original neutral anchor:

$$\text{CP} = \frac{1}{|\mathcal{P}||\mathcal{G}|} \sum_{p\in\mathcal{P}} \sum_{g\in\mathcal{G}} \cos(z_p^g, z_p^{\text{neut, orig}}) \tag{12}$$

  A CP value close to the baseline (Base CLIP) indicates that the method has preserved the core semantic identity of the profession. A significant drop indicates concept corruption.
- **Bias Neutralization (BN):** This metric measures the alignment between the male and female representations of the same profession. Higher similarity implies that the gender information distinguishing them has been removed (*i.e.*, the embeddings have merged).

$$\text{BN} = \frac{1}{|\mathcal{P}|} \sum_{p\in\mathcal{P}} \cos(z_p^{\text{male}}, z_p^{\text{fem}}) \tag{13}$$

  An ideal debiasing method should maximize BN (pushing it towards 1.0) while maintaining high CP.

D. Text Prompts

In this section, we provide details on the prompt sets used in our experiments: *bias prompts* ($\mathcal{P}_{\text{bias}}$), *diverse prompts* ($\mathcal{P}_{\text{div}}$), and *augmented query prompts* ($\mathcal{P}_q$). All prompts were generated using Google Gemini 2.5 Pro [12].

D.1. Bias Prompts

For each bias attribute we aim to mitigate (*e.g.*, gender, race), we define a corresponding set of bias classes $\mathcal{C}_a$ (*e.g.*, 'male', 'female' for gender; the seven ethnicity categories used in the main paper for race). To populate $\mathcal{P}_{\text{bias}}$, we prompt the LLM to generate 20 natural language captions for each class that describe the attribute with syntactic variety but without introducing confounding concepts.

For example:

- **Gender:** "A portrait of a man.", "A close-up of a woman's face.".
- **Race:** "A photo of a Black person from the side.", "A person with East Asian facial features.".

D.2. Diverse Prompts

To effectively identify bias neurons, it is crucial to measure activations relative to a neutral baseline rather than in absolute terms. This allows us to distinguish neurons specific to a bias concept from those that activate generally. We generate a set of 328 diverse, neutral text prompts ($\mathcal{P}_{\text{div}}$) designed to cover a broad range of semantic concepts with a roughly uniform distribution. These captions span various scenes, activities, objects, animals, and environments to ensure wide coverage of the semantic space.

Examples are provided in Tab. 8.

Table 8. Examples of diverse prompts used to establish a baseline activation distribution.

Prompt
`A firefighter in full gear holding a water hose.`
`A musician playing a guitar on a dimly lit stage.`
`A group of puppies tumbling and playing together.`
`A modern skyscraper made of glass and steel.`
`A golden retriever fetching a stick in a park.`
`A panoramic skyline of a modern city at night.`
`A rocky canyon carved by a river.`
`A close-up of moss growing on a tree trunk.`

D.3. Augmented Query Prompts

To improve robustness in both retrieval and zero-shot classification, we generate augmented prompts ($\mathcal{P}_q$) for each query using an LLM.

- **Retrieval:** For each input query (*e.g.*, "A photo of a criminal"), the LLM generates 10 paraphrases (*e.g.*, "An image of a criminal", "A person who committed a crime") to enhance semantic diversity and reduce sensitivity to specific wording.

Table 9. **Measuring gender bias for *Stereotype* and *Hair Color* queries on CelebA. Bold**: Best in setting (row group) and better than BASE CLIP. Underline: Best in setting, but not improving over BASE CLIP. Gray: Method is not zero-shot.

Method	ViT-B/16					ViT-L/14@336px				
	STEREOTYPE		HAIR COLOR			STEREOTYPE		HAIR COLOR		
	KL(↓)	MS(↓)	KL(↓)	MS(↓)	PREC.(↑)	KL(↓)	MS(↓)	KL(↓)	MS(↓)	PREC.(↑)
BASE CLIP	0.314	0.555	0.179	0.409	0.629	0.237	0.536	0.148	0.359	0.622
Bias-agnostic + input-specific prompts										
ROBOSHOT	0.189	**0.355**	**0.144**	**0.244**	0.633	0.195	**0.394**	0.276	0.429	0.675
SEM_i	**0.173**	0.443	0.191	0.345	**0.678**	**0.153**	0.413	0.237	0.458	**0.698**
Bias prompts only										
ORTH-PROJ	**0.188**	**0.382**	0.189	0.378	0.659	**0.099**	**0.355**	0.144	0.373	0.692
PRISM-MINI	0.190	0.384	0.188	**0.377**	0.658	**0.099**	0.357	0.143	0.366	0.696
SEM_b	0.240	0.496	**0.172**	0.395	**0.728**	0.199	0.481	**0.135**	0.366	**0.698**
ZSDEBIAS	*0.196*	*0.441*	*0.193*	*0.377*	*0.522*	*0.256*	*0.556*	*0.118*	*0.353*	*0.509*
Bias prompts + input-specific prompts										
ORTH-CALI	0.236	**0.408**	**0.148**	**0.375**	0.684	**0.054**	**0.266**	**0.107**	**0.312**	0.688
SEM_{bi}	**0.223**	0.488	0.181	0.399	**0.733**	0.209	0.490	0.168	0.402	**0.733**
PRISM	*0.143*	*0.377*	*0.060*	*0.186*	*0.669*	*0.061*	*0.245*	*0.171*	*0.299*	*0.659*
Bias prompts + input-specific prompts + labeled images										
BENDVLM	0.035	0.238	**0.028**	0.173	0.656	**0.030**	**0.217**	**0.028**	**0.164**	0.680
BENDSEM_{bi}	**0.030**	**0.224**	0.029	**0.158**	**0.750**	0.042	0.261	0.032	0.187	**0.685**

- **Zero-Shot Classification:** For each target class label (*e.g.*, "landbird"), the LLM generates 10 descriptive paraphrases (*e.g.*, "This is a picture of a landbird", "A depiction of a bird that lives on land").

We compute the median activation across these augmented sets to obtain a stable, noise-resistant representation of the query content (m_q), improving semantic generalization.

E. Extended Retrieval Results

In this section, we present the additional quantitative results for the retrieval task on CelebA, using both *Stereotype* and *Hair Color* queries, which were omitted from the main paper due to space constraints. We present these results in Tab. 9.

Fairness vs. Precision Trade-off. In the *Bias-agnostic* and *Bias prompts only* settings, our methods (SEM_i and SEM_b) demonstrate a competitive balance. While baselines like ROBOSHOT and ORTH-PROJ sometimes achieve better (lower) fairness scores (KL/MS) on this specific dataset, they often do so at the cost of retrieval quality. In contrast, our methods consistently maintain higher retrieval precision. For instance, on the ViT-B/16 backbone, SEM_i surpasses ROBOSHOT in *Hair Color* precision (0.679 vs. 0.632), and SEM_b outperforms ORTH-PROJ (0.729 vs. 0.660). This indicates that our method prioritizes preserving the query semantics while still reducing bias, avoiding the "over-correction" seen in prior methods that can degrade downstream task performance.

Modularity Improves Semantic Consistency. This advantage is most notable in the *Bias prompts + input-specific prompts + labeled images* setting. Here, the combination of our method with the baseline (BENDSEM_{bi}) provides a distinct advantage in semantic consistency. While BENDSEM_{bi} achieves fairness scores comparable to BENDVLM alone, it boosts retrieval precision by 9.5% (from 0.656 to 0.751) on the ViT-B/16 backbone. This confirms that integrating our sparse, feature-level modulation helps traditional debiasing methods retain critical semantic information, ensuring that the debiased embeddings remain accurate and useful for downstream tasks.

F. Extended Ablation Study

We provide the complete ablation results across all datasets and backbones in Tab. 10 (FairFace/UTKFace Retrieval), Tab. 11 (Zero-Shot Classification), and Tab. 12 (CelebA Retrieval). These results strongly support the design choices discussed in Sec. 4.4 of the main paper, confirming that our full methods, SEM_i and SEM_b, offer the most robust performance across diverse tasks.

Analysis of SEM_i. The zero-shot classification results (Tab. 11) reveal that removing our relevance-based attenuation leads to consistent and substantial drops in Worst-Group (WG) accuracy across all datasets and backbones. For instance, on Waterbirds (ViT-B/16), WG accuracy collapses from 0.498 to 0.210, underscoring the critical role of modulating spurious features. Operating directly in the dense CLIP space ("median CLIP") also proves unreliable. While this baseline performs well on the specific Waterbirds task (ViT-B/16), it is highly unstable elsewhere. It suffers significant performance drops on CelebA ZS (ViT-L/14) and consistently fails to mitigate gender bias in retrieval tasks, particularly on ViT-B/16. Specifically, compared to our SAE-based approach, the dense baseline yields substantially worse gender fairness metrics on FairFace, UTKFace, and CelebA *Stereotype* retrieval for the ViT-B/16 backbone

Table 10. **Extended ablation study for retrieval on FairFace and UTKFace. Bold**: Best in setting.

Method Variant	FairFace								UTKFace							
	ViT-B/16				ViT-L/14@336px				ViT-B/16				ViT-L/14@336px			
	Race		Gender		Race		Gender		Race		Gender		Race		Gender	
	KL(↓)	MS(↓)	KL(↓)	MS(↓)	KL(↓)	MS(↓)	KL(↓)	MS(↓)	KL(↓)	MS(↓)	KL(↓)	MS(↓)	KL(↓)	MS(↓)	KL(↓)	MS(↓)
SEM_i *Variants (Bias-Agnostic)*																
SEM_i (Full)	0.170	0.691	0.088	0.269	0.147	0.625	0.123	0.328	0.096	0.407	**0.065**	**0.245**	**0.059**	0.442	**0.032**	**0.185**
– $M(j) = 1$	**0.139**	**0.659**	**0.078**	**0.243**	**0.124**	**0.573**	0.093	0.288	**0.075**	**0.397**	0.088	0.278	0.061	**0.368**	0.038	0.196
– median CLIP	0.143	0.669	0.131	0.325	0.136	0.626	**0.087**	**0.262**	0.095	0.448	0.131	0.326	0.061	0.420	**0.032**	0.188
SEM_b *Variants (Bias-Aware)*																
SEM_b (Full)	0.232	0.749	0.098	**0.277**	0.194	0.706	0.098	0.298	0.148	0.510	0.123	0.320	0.137	0.445	0.047	0.202
– $M(j) = (1 - S_{\text{bias}})^2$	0.205	**0.738**	**0.095**	0.288	0.298	0.877	0.119	0.343	**0.072**	**0.400**	**0.063**	**0.215**	0.131	0.437	**0.023**	**0.151**
– $S_{\text{bias}} = S_{\text{gen}}$ only	**0.201**	0.754	0.105	0.294	0.211	0.726	**0.092**	**0.285**	0.133	0.501	0.129	0.331	0.158	0.461	0.045	0.201
– $S_{\text{bias}} = S_{\text{spec}}$ only	0.253	0.763	0.102	0.282	**0.185**	**0.700**	0.102	0.303	0.159	0.520	0.129	0.324	**0.111**	**0.435**	0.047	0.200

Table 11. **Extended ablation study for zero-shot classification on CelebA and Waterbirds. Bold**: Best in setting.

Method Variant	CelebA (Gender)						Waterbirds (Background)					
	ViT-B/16			ViT-L/14@336px			ViT-B/16			ViT-L/14@336px		
	Acc.(↑)	WG(↑)	Gap(↓)	Acc.(↑)	WG(↑)	Gap(↓)	Acc.(↑)	WG(↑)	Gap(↓)	Acc.(↑)	WG(↑)	Gap(↓)
SEM_i *Variants (Bias-Agnostic)*												
SEM_i (Full)	**0.736**	**0.611**	**0.125**	**0.791**	**0.745**	**0.046**	0.801	0.498	0.303	0.832	**0.523**	**0.309**
– $M(j) = 1$	0.734	0.609	**0.125**	0.729	0.640	0.089	0.834	0.210	0.624	0.872	0.357	0.515
– median CLIP	0.728	0.601	0.127	0.687	0.558	0.129	**0.840**	**0.563**	**0.277**	**0.879**	0.400	0.479
SEM_b *Variants (Bias-Aware)*												
SEM_b (Full)	0.797	0.711	0.086	**0.856**	**0.824**	0.032	0.825	0.433	0.392	0.855	0.624	0.231
– $M(j) = (1 - S_{\text{bias}})^2$	**0.818**	**0.750**	**0.068**	0.833	0.812	**0.021**	0.788	0.081	0.707	0.848	0.445	0.403
– $S_{\text{bias}} = S_{\text{gen}}$ only	0.809	0.736	0.073	0.846	0.818	0.028	**0.830**	**0.474**	0.356	**0.856**	0.647	0.209
– $S_{\text{bias}} = S_{\text{spec}}$ only	0.789	0.696	0.093	0.853	0.822	0.031	0.822	0.470	**0.352**	0.849	**0.662**	**0.187**

Table 12. **Extended ablation study for retrieval on CelebA. Bold**: Best in setting.

Method Variant	ViT-B/16					ViT-L/14@336px				
	Stereotype		Hair Color			Stereotype		Hair Color		
	KL(↓)	MS(↓)	KL(↓)	MS(↓)	Prec.(↑)	KL(↓)	MS(↓)	KL(↓)	MS(↓)	Prec.(↑)
SEM_i *Variants (Bias-Agnostic)*										
SEM_i (Full)	**0.173**	**0.443**	0.191	**0.344**	0.679	**0.153**	0.413	**0.236**	**0.458**	0.698
– $M(j) = 1$	0.185	0.456	0.193	0.371	0.672	0.195	0.490	0.274	0.484	**0.709**
– median CLIP	0.250	0.503	**0.181**	0.382	**0.689**	0.155	**0.411**	0.299	0.500	0.708
SEM_b *Variants (Bias-Aware)*										
SEM_b (Full)	0.240	0.495	0.172	0.396	0.729	0.199	0.481	0.136	0.366	0.699
– $M(j) = (1 - S_{\text{bias}})^2$	**0.110**	**0.334**	**0.122**	**0.312**	0.641	0.193	0.486	0.152	**0.338**	0.545
– $S_{\text{bias}} = S_{\text{gen}}$ only	0.238	0.493	0.182	0.405	**0.735**	**0.185**	**0.462**	0.142	0.372	**0.712**
– $S_{\text{bias}} = S_{\text{spec}}$ only	0.248	0.501	0.181	0.405	0.726	0.199	0.480	**0.135**	0.355	0.688

(Tabs. 10 and 12), as well as on CelebA *Hair Color* retrieval for both backbones (Tab. 12). In contrast, our full SEM_i method consistently achieves the best balance of fairness and performance across all benchmarks.

Analysis of SEM_b. The extended ablations highlight the necessity of our content-boosting term. While removing content boosting can sometimes improve retrieval fairness (notably on ViT-B/16), it leads to severe failures in several instances. For example, on Waterbirds (ViT-B/16), its WG accuracy plummets to 0.081 (Tab. 11); on FairFace (ViT-L/14), its KL divergence for the race attribute worsens significantly compared to the full method (0.298 vs. 0.194, Tab. 10); and crucially, removing content boosting severely degrades retrieval precision on CelebA across both backbones (dropping from 0.729 to 0.641 on ViT-B/16, and 0.699 to 0.545 on ViT-L/14). Furthermore, relying solely on either the *general* or *specific* bias score leads to inconsistent results. The "general only" variant often degrades social bias fairness (*e.g.*, race debiasing on ViT-L/14 or gender debiasing on ViT-B/16, Tab. 10), while the "specific only" variant struggles with semantic consistency in some settings (*e.g.*, yielding the worst CelebA accuracy and WG accuracy for ViT-B/16). Our full SEM_b formulation, which combines these scores, avoids these pitfalls, maintaining robust performance across both classification and retrieval.

Table 13. **Measuring race and gender bias for *Stereotype* queries on FairFace and UTKFace (ResNet backbones). Bold**: Best in setting (row group) and better than BASE CLIP. Underline: Best in setting, but not improving over BASE CLIP. Gray: Method is not zero-shot.

Method	FairFace								UTKFace							
	ResNet-50				ResNet-101				ResNet-50				ResNet-101			
	RACE		GENDER		RACE		GENDER		RACE		GENDER		RACE		GENDER	
	KL(↓)	MS(↓)	KL(↓)	MS(↓)	KL(↓)	MS(↓)	KL(↓)	MS(↓)	KL(↓)	MS(↓)	KL(↓)	MS(↓)	KL(↓)	MS(↓)	KL(↓)	MS(↓)
BASE CLIP	0.215	0.735	0.170	0.351	0.203	0.744	0.144	0.335	0.127	0.477	0.153	0.340	0.152	0.496	0.136	0.333
Bias-agnostic + input-specific prompts																
ROBOSHOT	0.215	0.706	0.299	0.445	0.222	0.798	0.338	0.494	0.152	0.586	0.258	0.414	0.206	0.652	0.323	0.492
SEM_i	**0.126**	**0.563**	**0.031**	**0.206**	**0.111**	**0.566**	**0.037**	**0.201**	**0.039**	**0.265**	**0.111**	0.383	**0.110**	**0.401**	**0.041**	**0.214**
Bias prompts only																
ORTH-PROJ	0.464	0.996	0.111	0.288	0.322	0.843	0.213	0.409	0.340	0.609	0.117	0.312	0.322	0.583	0.163	0.360
PRISM-MINI	0.454	0.983	0.113	0.291	0.313	0.837	0.215	0.411	0.336	0.608	0.117	0.311	0.315	0.582	0.168	0.363
SEM_b	**0.171**	**0.652**	**0.041**	**0.196**	**0.152**	**0.638**	**0.079**	**0.258**	**0.107**	**0.411**	**0.077**	**0.283**	**0.084**	**0.340**	**0.070**	**0.245**
ZSDEBIAS	*0.046*	*0.383*	*0.049*	*0.217*	*0.082*	*0.588*	*0.030*	*0.186*	*0.027*	*0.339*	*0.036*	*0.183*	*0.091*	*0.567*	*0.022*	*0.164*
Bias prompts + input-specific prompts																
ORTH-CALI	0.411	0.910	0.141	0.357	0.297	0.842	0.278	0.470	0.307	0.582	0.086	**0.257**	0.302	0.574	0.204	0.397
SEM_{bi}	**0.153**	**0.626**	**0.044**	**0.193**	**0.140**	**0.623**	**0.079**	**0.259**	**0.107**	**0.406**	**0.081**	0.281	**0.085**	**0.348**	**0.069**	**0.245**
PRISM	*0.157*	*0.632*	*0.069*	*0.245*	*0.152*	*0.594*	*0.107*	*0.282*	*0.134*	*0.523*	*0.088*	*0.265*	*0.133*	*0.532*	*0.127*	*0.314*
Bias prompts + input-specific prompts + labeled images																
BENDVLM	0.150	0.581	0.006	0.081	0.125	0.583	0.010	0.107	0.101	0.444	**0.008**	**0.093**	0.126	0.542	0.013	**0.123**
$BENDSEM_{bi}$	**0.067**	**0.455**	**0.005**	**0.079**	**0.059**	**0.425**	**0.006**	**0.087**	**0.042**	**0.371**	0.009	0.102	**0.035**	**0.367**	**0.012**	0.126

G. Extended Results on ResNet Backbones

To demonstrate that our feature-level debiasing framework generalizes beyond Vision Transformer (ViT) architectures, we extend our evaluation to convolutional neural networks. In this section, we benchmark our methods using the ResNet-50 and ResNet-101 CLIP backbones. The experimental setup, datasets, and metrics remain identical to those used for the ViT evaluations in the main paper.

The results are presented in Tab. 13 (FairFace and UTKFace Retrieval), Tab. 14 (Zero-Shot Classification), and Tab. 15 (CelebA Retrieval).

Consistent State-of-the-Art Fairness in Retrieval. The retrieval results in Tab. 13 confirm that our methods maintain their state-of-the-art fairness mitigation on convolutional backbones. In the bias-agnostic setting, SEM_i drastically reduces KL Divergence and MaxSkew compared to both the baseline and ROBOSHOT. For example, on FairFace Race (ResNet-50), SEM_i lowers KL divergence to 0.126 (compared to 0.215 for BASE CLIP and ROBOSHOT). In the bias-aware settings, SEM_b and SEM_{bi} reliably achieve the best fairness metrics across almost all evaluated demographics and datasets, outperforming projection-based baselines like ORTH-PROJ.

SEM Significantly Improves Zero-Shot Robustness. As shown in Tab. 14, SEM exhibits exceptional performance on zero-shot classification with ResNet backbones. Most notably, almost every single "best in setting" result achieved by a SEM variant strictly improves over the BASE CLIP baseline, effectively addressing both social biases (CelebA) and spurious correlations (Waterbirds). For instance, on Waterbirds (ResNet-50), SEM_b improves WG accuracy from 0.394 (BASE CLIP) to 0.577 (+18.3 points), substantially outperforming both ROBOSHOT (0.458) and ORTH-PROJ (0.457). Similarly, SEM_{bi} consistently achieves the lowest fairness Gap on CelebA across both ResNet models while maintaining high overall accuracy.

Maintaining the Fairness vs. Precision Trade-off. Tab. 15 details the performance on CelebA utilizing both *Stereotype* and *Hair Color* queries. While SEM_i achieves exceptional fairness scores (lowering *Stereotype* KL to 0.050 on ResNet-50), it does exhibit a drop in *Hair Color* precision (0.508). However, our bias-aware variants, SEM_b and SEM_{bi}, successfully navigate this trade-off. They significantly reduce *Stereotype* bias metrics compared to BASE CLIP while maintaining highly competitive precision scores (*e.g.*, 0.700 precision for SEM_b on ResNet-50, matching or nearing the baseline precision of 0.735).

Modularity with ResNets. Consistent with our ViT findings, our sparse modulation is highly complementary to existing methods when applied to ResNets. When integrating our SEM_{bi} embeddings into the BENDVLM framework, the resulting $BENDSEM_{bi}$ approach establishes new state-of-the-art results in the labeled images setting. On ResNet-101 zero-shot classification (Tab. 14), $BENDSEM_{bi}$ pushes Waterbirds WG accuracy to 0.638, significantly outperforming BENDVLM alone (0.194). Similarly, it provides the lowest social bias metrics across nearly all retrieval benchmarks (Tabs. 13 and 15).

Table 14. **Measuring zero-shot classification fairness on CelebA and Waterbirds (ResNet Backbones). Bold**: Best in setting (row group) and better than BASE CLIP. Underline: Best in setting, but not improving over BASE CLIP. Gray: Method is not zero-shot.

Method	CelebA (Gender)						Waterbirds (Background)					
	ResNet-50			ResNet-101			ResNet-50			ResNet-101		
	ACC.(↑)	WG(↑)	GAP(↓)	ACC.(↑)	WG(↑)	GAP(↓)	ACC.(↑)	WG(↑)	GAP(↓)	ACC.(↑)	WG(↑)	GAP(↓)
BASE CLIP	0.820	0.768	0.053	0.689	0.502	0.188	0.837	0.394	0.442	0.801	0.499	0.301
Bias-agnostic + input-specific prompts												
ROBOSHOT	**0.841**	**0.806**	**0.035**	0.737	0.596	0.140	0.762	0.458	0.304	0.761	0.450	0.310
SEM_i	0.835	0.799	0.036	**0.811**	**0.758**	**0.052**	**0.851**	**0.557**	**0.295**	**0.843**	**0.581**	**0.262**
Bias prompts only												
ORTH-PROJ	0.795	0.722	0.073	0.675	0.486	0.189	**0.859**	0.457	0.402	**0.858**	0.401	0.457
PRISM-MINI	0.795	0.722	0.073	0.675	0.486	0.189	**0.859**	0.457	0.402	**0.858**	0.401	0.457
SEM_b	**0.847**	**0.798**	**0.049**	**0.795**	**0.750**	**0.045**	0.845	**0.577**	**0.269**	0.846	**0.588**	**0.258**
ZSDEBIAS	*0.695*	*0.589*	*0.106*	*0.565*	*0.460*	*0.106*	*0.802*	*0.148*	*0.654*	*0.774*	*0.398*	*0.376*
Bias prompts + input-specific prompts												
ORTH-CALI	0.831	0.801	**0.030**	0.679	0.505	0.174	0.808	**0.704**	**0.104**	0.823	**0.554**	**0.269**
SEM_{bi}	**0.851**	**0.803**	0.048	**0.791**	**0.741**	**0.049**	**0.864**	0.525	0.338	**0.871**	0.541	0.330
PRISM	*0.824*	*0.763*	*0.061*	*0.788*	*0.688*	*0.100*	*0.886*	*0.634*	*0.252*	*0.840*	*0.672*	*0.168*
Bias prompts + input-specific prompts + labeled images												
BENDVLM	0.809	0.715	0.094	0.702	0.490	0.212	0.826	0.611	0.215	0.812	0.194	0.618
BENDSEM_{bi}	**0.848**	**0.815**	**0.033**	**0.784**	**0.699**	**0.086**	**0.856**	**0.648**	**0.208**	**0.881**	**0.638**	**0.243**

Table 15. **Measuring gender bias for *Stereotype* and *Hair Color* queries on CelebA (ResNet Backbones). Bold**: Best in setting (row group) and better than BASE CLIP. Underline: Best in setting, but not improving over BASE CLIP. Gray: Method is not zero-shot.

Method	ResNet-50					ResNet-101				
	STEREOTYPE		HAIR COLOR			STEREOTYPE		HAIR COLOR		
	KL(↓)	MS(↓)	KL(↓)	MS(↓)	PREC.(↑)	KL(↓)	MS(↓)	KL(↓)	MS(↓)	PREC.(↑)
BASE CLIP	0.389	0.622	0.187	0.367	0.735	0.300	0.560	0.205	0.414	0.718
Bias-agnostic + input-specific prompts										
ROBOSHOT	0.190	0.337	0.364	0.550	**0.762**	0.294	0.454	0.274	0.459	**0.723**
SEM_i	**0.050**	**0.193**	0.246	0.369	0.508	**0.041**	**0.185**	0.301	0.508	0.688
Bias prompts only										
ORTH-PROJ	0.145	0.383	**0.136**	0.343	**0.783**	**0.171**	**0.372**	0.325	0.506	0.752
PRISM-MINI	0.143	0.379	**0.136**	**0.339**	**0.783**	0.172	0.374	0.321	0.499	0.752
SEM_b	**0.111**	**0.311**	0.263	0.453	0.700	0.195	0.448	0.232	0.447	**0.767**
ZSDEBIAS	*0.058*	*0.237*	*0.129*	*0.291*	*0.436*	*0.016*	*0.119*	*0.046*	*0.187*	*0.317*
Bias prompts + input-specific prompts										
ORTH-CALI	**0.069**	**0.239**	**0.116**	**0.305**	**0.774**	0.191	**0.352**	0.313	0.502	0.751
SEM_{bi}	0.110	0.307	0.283	0.468	0.700	**0.165**	0.395	0.240	0.444	**0.766**
PRISM	*0.170*	*0.397*	*0.187*	*0.330*	*0.679*	*0.162*	*0.361*	*0.187*	*0.342*	*0.707*
Bias prompts + input-specific prompts + labeled images										
BENDVLM	0.029	0.218	0.025	0.169	**0.754**	**0.019**	**0.173**	**0.013**	**0.125**	0.704
BENDSEM_{bi}	**0.010**	**0.119**	**0.010**	**0.086**	0.619	0.021	0.184	0.018	0.140	**0.723**