Debiasing CLIP with Neural Interventions

Amelia Gómez-Grabowska[1,2], Jordi Gonzàlez[1,2][0000−0001−8033−0306], and Lluis Gomez[1,2][0000−0003−1408−9803]

[1] Computer Vision Center, Catalonia, Spain.
[2] Universitat Autònoma de Barcelona, Catalonia, Spain.
amelia.gomezg@autonoma.cat
{poal,lgomez}@cvc.uab.cat

Abstract. This paper presents an inference-time method to mitigate demographic bias in CLIP-like cross-modal retrieval models through targeted neural interventions in their internal attention mechanisms. We first identify "expert" attention heads that encode demographic information by systematically analyzing CLIP's internal representations in response to labeled inputs. At inference, we intervene these heads – replacing their activations with demographic prototypes or by neutralizing them (zero ablation). We chose to intervene specifically at the CLS token, as it aggregates information globally across image patches and is directly responsible for the final image embedding. Across fairness benchmarks such as SISPI and So-B-IT, our interventions achieve bias reduction comparable to or exceeding state-of-the-art methods, while being substantially lighter and requiring no retraining.

Keywords: Cross-modal Retrieval · Fairness · Neural Interventions.

1 Introduction

Vision-language models (VLMs) are increasingly vital in today's cross-modal retrieval landscape. Models such as CLIP [20], ALIGN [15], and BLIP [14] have significantly advanced performance in cross-modal retrieval and zero-shot classification, but their gains rely on large, uncurated web datasets which are known to embed social biases [2, 3, 8, 11].

Biases embedded in VLMs can surface in retrieval results – for example, consistently associating certain roles, occupations, or activities with specific genders or ethnicities – thereby reinforcing stereotypes. Such biased associations misrepresent diverse populations and can perpetuate systemic inequities. As VLM-based retrieval is increasingly used in search, recommendation, and decision-support systems, mitigating these biases is essential to ensure fairness, inclusivity, and trustworthiness.

Efforts to mitigate social biases in vision–language models have largely focused on dataset curation and adversarial debiasing [1, 11, 17]. These strategies often require costly retraining or fine-tuning, which is impractical for large-scale foundation models. Moreover, dataset-level interventions have inherent limits:

achieving demographic balance across all factors is infeasible, and even curated data may retain confounding artifacts [18]. There is therefore a need for methods that mitigate biases directly within model representations – without retraining – and that preserve retrieval performance.

In this paper, we propose a novel method to mitigate demographic biases – particularly gender and ethnicity – in vision–language models through targeted neural interventions at inference time. Building on concept steering techniques from large language models [22,24,25], we adapt them to CLIP's visual encoder. Lightweight probes trained on attention-head activations identify components that encode demographic information, which we then adjust via demogaphic replacement or zero-ablation at the CLS token representation. Unlike prior approaches that modify global embedding directions [27], our interventions act on individual attention heads, preserving the model's overall retrieval representation while suppressing biased associations without retraining.

Our contributions are threefold. (1) We introduce an inference-time strategy to mitigate demographic biases in vision–language models without retraining. (2) We present a systematic method to identify and quantify demographic information encoded in CLIP's visual representations. (3) We evaluate our approach on public fairness benchmarks [10,11], showing that it reduces bias while preserving retrieval utility. Our code and data will be made publicly available post-review.

2 Related Work

Neural interventions. Beyond traditional interpretability, recent research has shown that internal activations can be directly edited to steer model behavior. Suau et al. [25] identified "expert neurons" in language models that control specific concepts, while follow-up work demonstrated that soft interventions can mitigate toxicity without harming fluency [24]. More general approaches, such as Activation Transport (AcT) [22], shift activation distributions in a modality-agnostic way, offering a flexible alternative to fine-tuning or prompt engineering. These methods share the advantage of being inference-time, reversible, and computationally efficient. Our work extends this paradigm from topical control and toxicity mitigation to fairness in multimodal systems, showing that demographic bias can be reduced by editing a small set of attention heads.

Bias audits and mitigation in VLMs. Bias in vision–language models is well documented [6, 13, 28], with audits revealing systematic disparities across gender and ethnicity [11, 31]. Mitigation strategies generally fall into three categories: (1) training-time methods such as FairCLIP [17] and Prompt Array [1]; (2) post-processing methods like CLIP-clip [27], which remove embedding directions correlated with protected attributes; and (3) joint debiasing of image and text features [30]. While effective in some cases, these approaches either require expensive retraining or risk degrading retrieval performance.

In contrast, our method directly targets internal representations at inference time. By locating demographic-encoding heads in CLIP and intervening on

them, we reduce representational bias without retraining and with minimal loss in utility. Compared to CLIP-clip, we operate at the attention-head level, extending recent neural intervention techniques [22, 24] to fairness. This positions our work at the intersection of interpretability and bias mitigation, providing a fine-grained, transparent, and efficient alternative to prior approaches.

In summary, building on the aforementioned works in neural steering and activation patching, in the next Section we emphasize that intervening on attention heads identified via probing offers interpretability, i.e. one can point to specific heads encoding demographic traits, and complements methods such as CLIP-clip and DebiasCLIP by operating at a finer granularity.

3 Methodology

We mitigate demographic bias in CLIP's image encoder by detecting and intervening on attention heads that encode sensitive attributes. Our approach decomposes the CLS token representation into per-head contributions, allowing us to probe which heads encode gender or ethnicity traits. Heads identified as "experts" are then modified at inference time—either by replacing their activations with demographic prototypes or by zeroing them – while leaving the rest of the network untouched.

3.1 Identifying Expert Heads for Demographic Attributes

To identify attention heads that encode demographic attributes – such as gender or ethnicity – we adopt a probing-based approach inspired by prior work on expert subnetworks in large language models [24]. For this, we treat each attention head as a candidate "expert unit" whose activations may correlate with demographic attributes. This is confirmed by training a linear classifier to quantify its predictive capacity for demographic categories.

Given a dataset of images annotated with demographic labels (e.g., male vs. female), we first extract the outputs for each head h in layer l of CLIP's visual encoder. Formally, the contribution of attention head (l, h) to the CLS token is defined as:

$$x^{(l,h)} = \sum_{i=0}^{N} x_i^{l,h} \in \mathbb{R}^d \tag{1}$$

where N is the number of heads in CLIP ViT encoder.

For each attention head (l, h), we train a classifier to predict the demographic label from its head-level representation $x^{(l,h)}$. Given a dataset $\mathcal{D}$ of M labeled images: $\mathcal{D} = \{(I_1, y_1), \ldots, (I_M, y_M)\}, \quad y_m \in \mathcal{A}$; where $\mathcal{A}$ denotes a set of demographic attributes (e.g., gender or ethnicity classes), we extract the head-level output $x_m^{(l,h)} \in \mathbb{R}^d$ for each image I_m and attention head (l, h) (Equation 1). This results in a dataset per head: $\mathcal{D}^{(l,h)} = \{(x_m^{(l,h)}, y_m)\}_{m=1}^{M}$.

We then train a separate classifier $\phi^{(l,h)} : \mathbb{R}^d \to \mathcal{A}$ for each head (l, h) using $\mathcal{D}^{(l,h)}$ as input. We use a linear layer classifier (linear probe). The probes used to identify expert heads are trained offline on a moderate number of annotated images (as few as 500) and are not required at deployment. Once the expert heads are determined and prototype vectors are computed, the inference-time intervention consists only of replacing or zeroing activation vectors, incurring negligible computational overhead. Demographic labels are therefore needed only during probe training and evaluation, not during real-time operation.

To evaluate each head's ability to encode demographic information, we compute classification accuracy, area under the receiver operating characteristic curve (AUROC), average precision (AP), and precision-recall curves on a held-out test set. These metrics provide a robust and comprehensive view of how well each head encodes attribute-specific information. We hypothesize that only a small subset of heads encode demographic concepts with statistical significance. However, we assume a small subset of heads to achieve significantly higher predictive accuracy – these are the "expert heads" we aim to identify, so that we can then intervene on them in our mitigation framework.

Following the findings of [7], we restrict our analysis to the last attention layer of CLIP's ViT encoder. This layer has been shown to contribute the most to the final image representation, while interventions on early blocks and MLPs have negligible direct effect on CLIP's output and zero-shot performance.

3.2 Inference-Time Expert Head Interventions

After identifying expert heads associated with demographic attributes, we intervene them at inference-time by modifying their outputs prior to their contribution to the CLS token. By modifying only the outputs of expert heads at the CLS token, our interventions minimize demographic encoding without disrupting global image-text alignment, preserving retrieval utility. We choose to intervene at the CLS token because it aggregates global information across all image patches and directly determines the final image embedding; modifying earlier layers or patch tokens produced negligible fairness improvements and larger performance drops.

To mitigate biased behavior, we intervene at inference time by modifying the head-level outputs $x^{(l,h)}$ for a selected set of expert heads identified in the previous step. We experiment with two intervention strategies:

Demographic replacement: we replace the head output $x^{(l,h)}$ with the average activation associated with a specific demographic class. Let $\mathcal{D}_c^{(l,h)} = \{x_m^{(l,h)} \mid y_m = c\}$ be the set of activations for class c (e.g., "female") in our dataset. We compute the prototype vector for class c as:

$$\mu_c^{(l,h)} = \frac{1}{|\mathcal{D}_c^{(l,h)}|} \sum_{x \in \mathcal{D}_c^{(l,h)}} x \tag{2}$$

Demographic prototypes are computed once offline using labeled data; during deployment, no demographic labels are required, and one could alternatively

derive prototypes via unsupervised clustering or continuous demographic embeddings, making the method purely inference-time. Thus, at inference time, we substitute $x^{(l,h)} \leftarrow \mu_c^{(l,h)}$ for the chosen class c (e.g., "neutral" (average across all image representations), or an average across "male" and "female" representations).

Zero Ablation: we set $x^{(l,h)} = 0$, effectively neutralizing the influence of the head on the CLS token.

These interventions are applied only to heads identified as experts via the probing step. All other heads remain unaltered, preserving the bulk of the model's learned representations. This targeted modification enables mitigation of demographic encoding while minimizing impact on overall model behavior. The probing step is performed offline and requires less than one minute per dataset on a single GPU; at inference time, interventions incur negligible overhead because they replace or zero out a few activation vectors without modifying weights.

4 Experiments

We evaluate our intervention approach using three datasets and multiple fairness metrics, reporting results on expert head identification, retrieval fairness, and trade-offs with model generic retrieval utility.

4.1 Datasets and Evaluation Metrics

We make use of three different datasets in our experiments: A curated 30K subset of **MSCOCO** [16] images with gender annotations [18] to probe attention heads correlated with gender; **SISPI** (Social Inclusive Synthetic Professionals Images) [10], a synthetic dataset comprising 49K images with a balanced representation of demographic groups across various professional roles [23]; and **FairFace** [12], a set of 100K cropped face images from the YFCC-100M Flickr dataset [26] that is balanced on gender and race – we use a subset of 500 training images to identify expert heads and 11K validation images to evaluate our interventions. For each dataset we report the head-level activation statistics and demographic prototypes to facilitate replication of our probes and interventions.

Although SISPI is synthetic, it is designed to provide a controlled, balanced benchmark for fairness research. We treat the SISPI dataset as our validation setting for model variants' selection, as it provides both fairness (NDKL) and retrieval utility (NDCG, mAP) metrics within the same dataset and domain. This allows us to jointly assess bias mitigation effectiveness and model utility when exploring intervention variants. We then perform experiments on three real datasets (FairFace, MSCOCO, and Flickr30K) to demonstrate generalizability and compare with state-of-the-art CLIP debiasing methods.

To measure models' fairness on SISPI we use the **NDKL** [9, 21, 29] metric, which evaluates fairness in retrieval by comparing the demographic distribution of ranked images against a uniform reference distribution. It applies KL

divergence with a position-discounting factor, yielding a score of 0 for perfectly balanced rankings and higher values for greater disparity.

For evaluation on FairFace we follow Hamidieh et al. [11] and use **Normalized Entropy**. Their evaluation framework leverages a comprehensive taxonomy – called So-B-IT, Social Bias Implications Taxonomy – of descriptive words that often carry stereotypical connotations. Descriptive words are grouped into multiple socially relevant categories such as "appearance", "behavioral", "political", "religion", "occupation", among others.

Normalized entropy measures the diversity of demographic labels (race, gender) among the top-100 retrieved images for each of the So-B-IT taxonomy descriptive words; values close to 1 indicate uniform distributions, while values near 0 reveal strong bias.

4.2 Identification of Expert Heads and Neurons

We identify attention heads and neurons encoding demographic information in CLIP ViT-B/32 and ViT-B/16 pretrained on WIT-400M [20]. Head-level probes are trained with cross-entropy loss using the Adam optimizer (learning rate 0.001) for 10 epochs.

Table 1 shows the performance of individual attention heads from the final layer (layer 12) of CLIP ViT-B/32 on gender classification across MSCOCO, SISPI, and FairFace. Each head is probed with a linear classifier trained on its activation vector, and performance is reported as accuracy and average precision (AP). Heads 3, 11, and 12 consistently achieve the highest AP across datasets, identifying them as the main carriers of gender-related information. Probes for ethnicity on FairFace reveals overlapping expert heads (3, 10, 11), suggesting that certain heads jointly encode multiple demographic attributes. A similar analysis on CLIP ViT-B/16 yields comparable patterns, with different heads: 1, 2, and 12 for gender; 2, 7, and 11 for ethnicity.

Table 1. Performance of CLIP ViT-B/32 (layer 12) heads on gender classification across datasets. Accuracy (Acc) and Average Precision (AP) are reported. Shaded columns indicate potentially expert heads.

	h:1	h:2	h:3	h:4	h:5	h:6	h:7	h:8	h:9	h:10	h:11	h:12
MSCOCO Acc	0.684	0.682	0.733	0.693	0.695	0.691	0.694	0.693	0.695	0.693	0.738	0.741
MSCOCO AP	0.349	0.362	0.757	0.394	0.365	0.352	0.370	0.381	0.364	0.347	0.745	0.687
SISPI Acc	0.703	0.720	0.738	0.691	0.651	0.668	0.684	0.668	0.687	0.707	0.938	0.831
SISPI AP	0.750	0.776	0.922	0.773	0.767	0.692	0.718	0.660	0.804	0.753	0.983	0.984
FairFace Acc	0.738	0.546	0.689	0.629	0.618	0.547	0.536	0.562	0.576	0.584	0.906	0.856
FairFace AP	0.747	0.754	0.928	0.692	0.874	0.773	0.658	0.753	0.764	0.669	0.964	0.914

Figure 1 depicts the distribution of Average Precision (AP) scores for gender classification across individual neurons within selected attention heads. Expert

heads (top row) show skewed distributions, with many neurons exceeding an AP threshold of 0.42, unlike other non-expert heads (bottom row). This confirms that only a subset of neurons within expert heads encode demographic cues, providing a basis for targeted interventions.

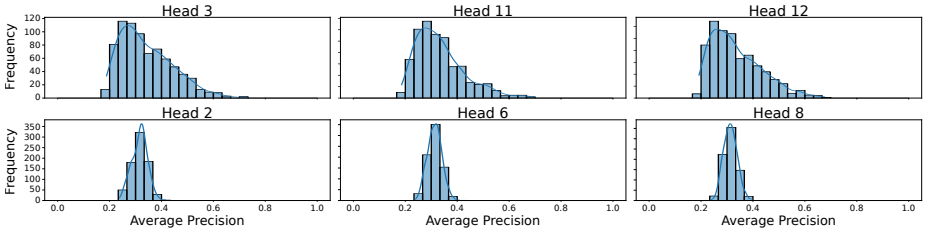

Fig. 1. Distribution of neuron-level Average Precision on gender classification for selected attention heads in layer 12.

4.3 Bias Mitigation with Neural Interventions

We now evaluate our interventions using standard fairness metrics in cross-modal retrieval, comparing them against an unmodified CLIP baseline and state-of-the-art methods. This allows us to assess how effectively the approach reduces demographic bias while maintaining retrieval performance.

Bias Mitigation on Synthetic Data. We first evaluate our framework on the SISPI dataset as a validation setting for model variants' selection. Table 2 reports results for interventions applied to attention heads identified as gender-related. Fairness is measured using NDKL, where lower values indicate smaller demographic skew, and retrieval quality is assessed with mean average precision (mAP) and normalized discounted cumulative gain (NDCG). The baseline CLIP model (without intervention) achieves NDCG = 0.745, mAP = 0.326, and NDKL = 0.114. We then apply multiple interventions to these gender-expert heads and compare their impact on fairness and retrieval performance.

The largest NDKL reduction occurs when jointly modifying all expert heads (3, 11, 12 for ViT-B/32 and 1, 2, 12 for ViT-B/16). Among these interventions, the "male" prototype yields the lowest NDKL, followed by zero ablation. Single-head interventions produce smaller gains, confirming that combining multiple gender-associated heads more effectively mitigates bias in retrieval rankings. A similar analysis for ethnicity shows that using the "white" prototype yields the largest NDKL reduction when jointly modifying all expert heads.

The superior performance of the male and white prototypes in reducing NDKL for gender and ethnicity, respectively, confirms that CLIP inherently privileges white male representations. This reflects the overrepresentation of such

Table 2. Evaluation of attention-head interventions on SISPI text-to-image retrieval using CLIP ViT-B/32 and ViT-B/16 pretrained on WIT-400M. Variants differ by the modified heads and replacement vectors: mean activations for male, female, or gender-neutral (person) images, and zero-vector ablation.

	ViT-B-32				ViT-B-16			
Intervention	Heads	NDCG ↑	mAP ↑	NDKL ↓	Heads	NDCG ↑	mAP ↑	NDKL ↓
n.a.	n.a.	0.745	0.326	0.114	n.a.	0.780	0.396	0.118
Person	3,11,12	0.746	0.330	0.074	1,2,12	0.777	0.389	0.083
Male	3,11,12	0.748	0.333	**0.060**	1,2,12	0.779	0.392	**0.073**
Female	3,11,12	0.746	0.329	0.078	1,2,12	0.777	0.389	0.089
Zero	3,11,12	0.753	0.340	0.066	1,2,12	0.788	0.413	0.082
Person	3	0.744	0.324	0.112	2	0.786	0.407	0.086
Male	3	0.743	0.324	0.108	2	0.787	0.410	0.076
Female	3	0.745	0.324	0.114	2	0.786	0.407	0.090

groups in web-scale training data and the broader societal biases embedded in those datasets.

Importantly, we appreciate in Table 2 that interventions that reduce bias the most (up to a 42% drop in NDKL) preserve retrieval quality, with mAP and NDCG remaining stable across all variants. This robustness stems from the targeted nature of our edits, which modify only the CLS token activations of gender-associated heads while leaving the rest of the representation intact. These results confirm that focused, head-level interventions can mitigate gender bias without degrading performance, a finding revisited later when analyzing retrieval trade-offs on standard benchmarks (*cf.* Retrieval Performance Trade-offs on Standard Benchmarks).

We also applied interventions at the neuron level, targeting units within each expert head with AP>0.42 in gender classification. Although some neurons showed strong discrimination, these fine-grained edits produced no meaningful bias reduction (Table 3, left), reinforcing the advantage of structural, head-level interventions. Following [24], we also test a damping intervention that scales neuron activations instead of replacing them with demographic prototypes. For each expert neuron, a scaling factor $\alpha \in [0, 1]$ attenuates its contribution to the output embedding. Table 3 (right) reports results on SISPI with $\alpha = 0.1$.

Our findings indicate that attention heads in CLIP function as coherent units with causal influence on the final representation. Head-level interventions act on high-level latent variables that modulate the output through the residual stream while preserving structural consistency. In contrast, neuron-level edits—particularly sparse ones—fail to capture the distributed and interdependent nature of feature encoding within a head. This supports recent causal tracing studies showing that interventions on structurally meaningful components are more effective than isolated unit edits.

Table 3. Evaluation of neuron-level interventions (left) and neural damping interventions (right) on SISPI text-to-image retrieval using CLIP ViT-B-32.

Intervention	mAP ↑	NDKL ↓
n.a.	0.326	0.114
h:3,11,12 (person)	0.331	0.110
h:3,11,12 (male)	0.332	0.103
h:3,11,12 (female)	0.330	0.120
h:3,11,12 (zero)	0.333	0.110

Intervention	mAP ↑	NDKL ↓
n.a.	0.326	0.114
h:3,11,12 (α=0.1)	0.333	0.111
h:3 (α=0.1)	0.329	0.119
h:11 (α=0.1)	0.330	0.116
h:12 (α=0.1)	0.330	0.113

Unless otherwise stated, we use the best-performing intervention variant from Table 2 ("male" and "white" prototypes on three expert heads) in the rest of our experiments, which provides the best fairness–utility balance on SISPI.

Bias Mitigation on So-B-IT Word Associations Next, we evaluate bias mitigation using the FairFace dataset, which serves both to identify demographics-related attention heads and to assess intervention effectiveness. Table 4 reports Normalized Entropy scores for race (R) and gender (G) across five So-B-IT taxonomy categories. We compare our approach against several state-of-the-art debiasing methods, including CLIP-clip [27] (post-processing of biased embedding directions), DebiasCLIP [1] (adversarially trained prompt-based debiasing), Biased-prompts [5] (explicit bias-aware prompt tuning), and VL_Debiasing [30] (joint vision-language debiasing).

Table 4 reports results for interventions using both gender- and ethnicity-based replacements (with "male" and "white" prototypes). Unlike earlier experiments – limited to gender due to missing ethnicity labels in MSCOCO – FairFace includes both attributes, enabling multi-attribute intervention and evaluation.

Table 4 shows that our interventions achieve competitive debiasing performance for both gender (G) and race (R). Unlike other methods, which often increases race bias when reducing gender bias (specially on the ViT-B/32), our method avoids this trade-off. Among variants, the best results are again obtained by jointly modifying all expert heads using the "male" and "white" prototypes respectively, consistent with SISPI findings (Table 2). Notably, our approach matches or surpasses the performance of state-of-the-art methods in several So-B-IT categories (see bold and underlined entries), while remaining significantly lighter—requiring no retraining and minimal computation compared to VL_Debiasing or CLIP-clip. The "(Male+White)" variant corresponds to an intersectional intervention that simultaneously targets gender- and race-encoding heads, achieving balanced improvements across both demographic axes.

Figure 2 illustrates the effect of our intervention on the query "An image of a doctor." Compared to the baseline, which skews heavily toward male-presenting individuals, our method yields a more demographically balanced set of retrieved images, highlighting its qualitative impact on reducing representation bias.

Table 4. Normalized Entropy for race (R) and gender (G) biases across So-B-IT categories. Cell color indicates deviation from the baseline model: blue denotes fairer (lower bias) and red denotes more bias. Best results in bold, and second best underlined.

	appearance		behavioral		education		criminal		healthcare	
	R	G	R	G	R	G	R	G	R	G
OAI ViT-B-32	0.85	0.71	0.86	0.88	0.86	0.84	0.90	0.79	0.90	0.87
CLIP-clip [27]	0.82	**0.94**	0.86	**0.97**	0.87	**0.96**	0.85	**0.96**	0.85	**0.94**
DebiasCLIP [1]	0.83	0.87	0.83	0.92	0.85	0.93	0.86	0.85	0.83	0.92
Biased-prompts [5]	0.83	0.84	0.87	0.90	**0.90**	0.92	0.89	0.93	0.88	0.90
VL_Debiasing [30]	0.85	0.70	**0.89**	0.89	0.89	0.93	0.89	0.89	0.88	0.91
Ours (Male)	0.86	0.84	0.86	0.92	0.88	0.91	0.90	0.85	0.90	**0.94**
Ours (White)	**0.88**	0.80	0.88	0.90	**0.90**	0.86	**0.93**	0.81	**0.92**	0.91
Ours (Male+White)	0.86	0.86	0.88	0.93	0.89	0.94	0.91	0.84	0.91	0.92
OAI ViT-B-16	0.84	0.77	0.83	0.87	0.82	0.85	0.87	0.85	0.86	0.84
CLIP-clip [27]	0.89	0.79	0.88	0.90	0.88	0.86	0.91	0.89	**0.91**	0.86
Biased-prompts [5]	0.86	**0.91**	0.89	0.92	0.90	0.81	0.90	0.83	0.89	0.86
VL_Debiasing [30]	0.85	0.71	0.87	0.80	0.87	0.92	0.87	0.85	0.85	0.77
Ours (Male)	0.87	0.86	0.87	**0.94**	0.87	**0.96**	0.87	0.91	0.85	0.89
Ours (White)	0.90	0.85	**0.91**	0.93	**0.91**	0.93	**0.92**	0.89	**0.91**	**0.90**
Ours (Male+White)	**0.91**	0.85	**0.91**	**0.94**	**0.91**	0.95	**0.92**	**0.92**	0.90	0.89

4.4 Retrieval Performance on Standard Benchmarks

In this section, we evaluate the effect of our interventions on two standard retrieval benchmarks: MSCOCO [16] and Flickr30K [19]. Table 5 compares our method with the ViT-B/32 baseline and three state-of-the-art methods.

Across both MSCOCO and Flickr30K, our interventions cause only minor drops in retrieval performance. On MSCOCO, Recall@K scores change by at most two points, while Flickr30K shows slightly larger but still modest decreases, mainly in R@1. This reflects the typical fairness–utility trade-off observed in prior debiasing work [9]. Overall, retrieval performance remains comparable to Biased-prompts and VL_Debiasing, and notably superior to CLIP-clip, confirming that our strategy mitigates bias with negligible impact on retrieval utility.

5 Theory of Change

5.1 Societal Need and Contribution

Vision–language models (VLMs) such as CLIP are increasingly embedded in information retrieval systems. These systems can reproduce and amplify social biases leading to stereotyping, exclusion, or misrepresentation of marginalized groups. The societal need motivating this work is therefore to ensure that multimodal retrieval systems operate equitably and transparently.

Fig. 2. Top-8 retrieved images for the query "An image of a doctor" on the FairFace dataset. The top row shows results from the baseline CLIP model; the bottom row shows results after applying our "male" intervention, yielding more diverse results.

Table 5. Zero-shot retrieval results on MSCOCO and Flickr30K. Recall@K for text→image (Text) and image→text (Image). Best results in bold, and second best underlined.

	MSCOCO						**Flickr30K**					
	Text			**Image**			**Text**			**Image**		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
OAI ViT-B-32	**0.50**	**0.75**	**0.83**	**0.30**	**0.56**	**0.67**	**0.79**	**0.95**	**0.98**	**0.59**	**0.84**	**0.90**
CLIP-clip [27]	0.44	0.69	0.78	0.27	0.50	0.62	0.70	0.90	0.94	0.52	0.77	0.85
Biased-prompts [5]	0.47	0.72	0.82	**0.30**	0.55	0.66	0.76	0.94	0.97	0.58	0.83	0.89
VL_Debiasing [30]	**0.50**	**0.75**	**0.83**	0.29	0.55	0.66	**0.79**	**0.95**	**0.98**	0.57	0.83	**0.90**
Ours (Male)	0.48	0.74	0.82	0.29	0.54	0.66	0.75	0.94	0.97	0.57	0.82	0.89

Our contribution addresses this need by introducing a lightweight, inference-time strategy for mitigating demographic bias in VLMs without retraining. By identifying and intervening on attention heads that encode demographic information, we offer a practical path for reducing representational harms in deployed retrieval systems. Because our method requires only minimal computational resources and no retraining, it lowers the barrier for practitioners to integrate fairness-aware adjustments into real-world pipelines.

5.2 Preconditions and Assumptions

For our approach to have its intended positive effect, several preconditions must hold. First, demographic annotations used for probing must be reliable, contextually appropriate, and representative of the populations affected by downstream retrieval systems. Second, the institutions deploying such systems must be motivated to measure and mitigate bias, and to incorporate fairness evaluation into their model monitoring practices. Third, users of these systems must have avenues for contestation and oversight so that fairness interventions align with community expectations and local norms.

Our method assumes that bias can be localized to identifiable components (e.g., attention heads) and that such edits generalize beyond our evaluation

datasets. It further presumes that demographic information is well-defined enough to enable partial mitigation. While these assumptions hold empirically in our setting, they may not fully extend to more complex or intersectional contexts.

5.3 Potential Negative Externalities

While our approach aims to promote fairness, several risks remain. First, interventions based on coarse demographic categories may inadvertently reinforce the same categorical boundaries they seek to mitigate. This limitation stems from the demographic labels available in current datasets and does not capture the fluid, intersectional nature of identity. Extending our framework to self-identified, culturally responsive, and multi-label annotations would mitigate this risk.

Second, fairness interventions could be misapplied as a form of "fairness washing," providing an appearance of equity without addressing deeper data collection or governance issues. To avoid this, our method should be viewed as one component within a broader pipeline of participatory dataset design, bias auditing, and accountability practices.

Finally, any intervention that modifies model representations can alter downstream behaviors in unpredictable ways, including shifts in retrieval relevance or new forms of under-representation. Continuous monitoring and community feedback are therefore essential to ensure that mitigation aligns with social values.

5.4 Ethical and Sociocultural Considerations

Our study follows prior work in using the gender categories *Male* and *Female* and broad ethnic labels such as *White*, *Black*, *East Asian*, *Middle Eastern*, and *Latino/Hispanic*, as defined in the SISPI and FairFace datasets. These simplifications are common in demographic fairness research but inevitably oversimplify complex identities. We acknowledge that ethnicity and gender are socially constructed and context-dependent.

We approach these classifications with cultural sensitivity and awareness of their limitations. Future work should adopt more granular, self-identified demographic annotations and community-driven labeling practices [4].

6 Conclusion

We show that editing attention heads associated with demographic attributes at inference time provides a lightweight and interpretable path to improving fairness in VLMs. By identifying and modifying heads that encode gender and ethnicity information, we reduce representational and retrieval bias without retraining or compromising performance. Because interventions minimally affect utility, they can be applied at inference as an optional fairness correction, enabling users to toggle them based on application needs. Experiments on SISPI and So-B-IT benchmarks confirm that structured activation edits – particularly

at the head level – achieve fairness gains comparable to or exceeding state-of-the-art methods, while remaining lighter and requiring no retraining. Our analysis of neuron-level interventions further highlights the importance of structural coherence and targeted design when modifying internal representations. Overall, these results demonstrate the potential of activation-level control as a scalable, modular, and performance-preserving approach to fairness in multimodal retrieval systems. Our code and data will be made publicly available post-review.

Acknowledgments. This work has been supported by the Ramon y Cajal research fellowship RYC2020-030777-I/AEI/ 10.13039/501100011033, Spanish grants PID2020-120611RB-I00, PID2024-162984NB-I00 and PID2023-146426NB-100 (funded by FEDER/UE and MICIU/AEI/10.13039/501100011033), the European Union under the CERV programme (Call: CERV-2024-CHAR-LITI-CHARTER, Project ID: 101214711), the European Lighthouse on Safe and Secure AI - ELSA funded by European Union's Horizon Europe programme under grant agreement No 101070617, and the Consolidated Research Group 2021 SGR 01559 from the Research and University Department of the Catalan Government.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Berg, H., Hall, S.M., Bhalgat, Y., Yang, W., Kirk, H.R., Shtedritski, A., Bain, M.: 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). pp. 806–822. Association for Computational Linguistics (nov 2022), https://aclanthology.org/2022.aacl-main.61
2. Birhane, A., Han, S., Boddeti, V., Luccioni, S., et al.: Into the laion's den: Investigating hate in multimodal datasets. In: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2023)
3. Birhane, A., Prabhu, V.U., Kahembwe, E.: Multimodal datasets: misogyny, pornography, and malignant stereotypes. arXiv preprint arXiv:2110.01963 (2021)
4. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Proceedings of the 1st Conference on Fairness, Accountability and Transparency. pp. 77–91 (2018)
5. Chuang, C.Y., Jampani, V., Li, Y., Torralba, A., Jegelka, S.: Debiasing vision-language models via biased prompts. arXiv preprint arXiv:2302.00070 (2023)
6. De-Arteaga, M., Romanov, A., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., Geyik, S., Kenthapadi, K., Kalai, A.T.: 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. p. 120–128. FAT* '19, ACM (Jan 2019). https://doi.org/10.1145/3287560.3287572, http://dx.doi.org/10.1145/3287560.3287572
7. Gandelsman, Y., Efros, A.A., Steinhardt, J.: Interpreting clip's image representation via text-based decomposition. In: The Twelfth International Conference on Learning Representations (2024)

8. Garcia, N., Hirota, Y., Wu, Y., Nakashima, Y.: Uncurated image-text datasets: Shedding light on demographic bias. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6957–6966 (2023)
9. Geyik, S.C., Ambler, S., Kenthapadi, K.: Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In: Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining. pp. 2221–2231 (2019)
10. Gomez, L.: Measuring text-image retrieval fairness with synthetic data. In: Proceedings of the SIGIR Conference on Research and Development in Information Retrieval. ACM (2025)
11. Hamidieh, K., Zhang, H., Gerych, W., Hartvigsen, T., Ghassemi, M.: Identifying implicit social biases in vision-language models. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. vol. 7, pp. 547–561 (2024)
12. Karkkainen, K., Joo, J.: Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1548–1558 (2021)
13. Lee, N., Bang, Y., Lovenia, H., Cahyawijaya, S., Dai, W., Fung, P.: Survey of social bias in vision-language models. arXiv preprint arXiv:2309.14381 (2023)
14. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022)
15. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems **34**, 9694–9705 (2021)
16. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
17. Luo, Y., Shi, M., Khan, M.O., Afzal, M.M., Huang, H., Yuan, S., Tian, Y., Song, L., Kouhana, A., Elze, T., Fang, Y., Wang, M.: Fairclip: Harnessing fairness in vision-language learning (2024), https://arxiv.org/abs/2403.19949
18. Meister, N., Zhao, D., Wang, A., Ramaswamy, V.V., Fong, R., Russakovsky, O.: Gender artifacts in visual datasets. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4837–4848 (2023)
19. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: IEEE International Conference on Computer Vision (2015)
20. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
21. Raj, A., Ekstrand, M.D.: Measuring fairness in ranked results: An analytical and empirical comparison. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 726–736 (2022)
22. Rodriguez, P., Blaas, A., Klein, M., Zappella, L., Apostoloff, N., marco cuturi, Suau, X.: Controlling language and diffusion models by transporting activations. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=l2zFn6TIQi

23. Saunders, D., Byrne, B.: Reducing gender bias in neural machine translation as a domain adaptation problem. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7724–7736 (2020)
24. Suau, X., Delobelle, P., Metcalf, K., Joulin, A., Apostoloff, N., Zappella, L., Rodríguez, P.: Whispering experts: neural interventions for toxicity mitigation in language models. In: Proceedings of the 41st International Conference on Machine Learning. pp. 46843–46867 (2024)
25. Suau, X., Zappella, L., Apostoloff, N.: Self-conditioning pre-trained language models. In: International Conference on Machine Learning. pp. 4455–4473. PMLR (2022)
26. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: Yfcc100m: The new data in multimedia research. Communications of the ACM **59**(2), 64–73 (2016)
27. Wang, J., Liu, Y., Wang, X.: Are gender-neutral queries really gender-neutral? mitigating gender bias in image search. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (2021)
28. Wu, X., Wang, Y., Wu, H.T., Tao, Z., Fang, Y.: Evaluating fairness in large vision-language models across diverse demographic attributes and prompts (2024), https://arxiv.org/abs/2406.17974
29. Yang, K., Stoyanovich, J.: Measuring fairness in ranked outputs. In: Proceedings of the 29th international conference on scientific and statistical database management. pp. 1–6 (2017)
30. Zhang, H., Guo, Y., Kankanhalli, M.: Joint vision-language social bias removal for clip. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4246–4255 (June 2025)
31. Zhang, M., Chunara, R.: Leveraging vision-language models for fair facial attribute classification (2024), https://arxiv.org/abs/2403.10624