CABIN: Debiasing Vision-Language Models Using Backdoor Adjustments

Bo Pang[1] , **Tingrui Qiao**[1] , **Caroline Walker**[2] , **Chris Cunningham**[3] and **Yun Sing Koh**[1]
[1]School of Computer Science, University of Auckland, Auckland, New Zealand
[2]The Liggins Institute, University of Auckland, Auckland, New Zealand
[3]Research Centre for Māori Health and Development, Massey University, Wellington, New Zealand
{bpan882, tqia361}@aucklanduni.ac.nz, C.W.Cunningham@massey.ac.nz, {caroline.walker, y.koh}@auckland.ac.nz

Abstract

Vision-language models (VLMs) have demonstrated strong zero-shot inference capabilities but may exhibit stereotypical biases toward certain demographic groups. Consequently, downstream tasks leveraging these models may yield unbalanced performance across different target social groups, potentially reinforcing harmful stereotypes. Mitigating such biases is critical for ensuring fairness in practical applications. Existing debiasing approaches typically rely on curated face-centric datasets for fine-tuning or retraining, risking overfitting and limiting generalizability. To address this issue, we propose a novel framework, **CABIN** (*Causal Adjustment Based INtervention*). It leverages a causal framework to identify sensitive attributes in images as confounding factors. Employing a learned mapper, which is trained on general large-scale image-text pairs rather than face-centric datasets, CABIN may use text to adjust sensitive attributes in the image embedding, ensuring independence between these sensitive attributes and image embeddings. This independence enables a backdoor adjustment for unbiased inference without the drawbacks of additional fine-tuning or retraining on narrowly tailored datasets. Through comprehensive experiments and analyses, we demonstrate that CABIN effectively mitigates biases and improves fairness metrics while preserving the zero-shot strengths of VLMs. The code is available at: https://github.com/ipangbo/causal-debias

1 Introduction

Vision-Language Models (VLMs) such as CLIP [Radford *et al.*, 2021] unify visual and textual representations, enabling them to excel in a broad range of downstream tasks. Training on large-scale image-text pairs allows these models to learn a joint embedding space where visual and linguistic concepts align effectively, exhibiting strong generalization even without extensive task-specific fine-tuning [Zhao *et al.*, 2024]. Their ability to learn generalisable representations makes VLMs attractive for numerous real-world applications, from web-scale image searches [Lülf *et al.*, 2024] to automated content filtering [Hong *et al.*, 2024].

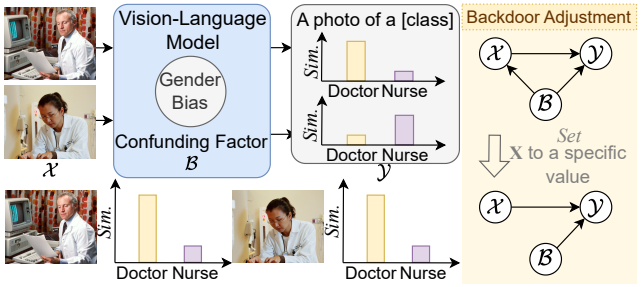

Figure 1: Debiasing VLM using backdoor adjustment. A biased VLM is influenced by gender-based confounding factors ($\mathcal{B}$), which leads to biased prediction based on input ($\mathcal{X}$) and predicted result ($\mathcal{Y}$), although pictures reflect the same occupations. By applying backdoor adjustment to conditioning the value of $\mathcal{X}$ (Set operation) and observing $\mathcal{Y}$, we estimate and mitigate the confounding effect of $\mathcal{B}$, such that models may achieve fairer predictions.

Despite their impressive performance, recent studies have shown that VLMs inherit stereotypical biases in their training data [Birhane *et al.*, 2021]. Such biases may manifest as performance disparities across sensitive attributes (*e.g.* gender, race and age). For instance, a VLM might misclassify or underrepresent certain demographic groups in tasks, *e.g.* face recognition or occupational image classification [Parraga *et al.*, 2023]. For image retrieval, such biases could cause search results to reflect stereotypical associations. For example, querying "nurse" primarily returns female-presenting images. Such bias adversely impacts both the inclusivity and fairness of VLM-driven systems.

Efforts to mitigate bias in machine learning encompass several intervention strategies. Pre-processing methods, *e.g.* data augmentation [Choi *et al.*, 2020; Lim *et al.*, 2023], and in-processing techniques, *e.g.* adversarial learning [Zhang *et al.*, 2018] and feature disentanglement [Locatello *et al.*, 2019], generally require retraining the entire model. This process consumes substantial computational resources and often involves counterfactual data [Zhang *et al.*, 2022; Lee *et al.*, 2023], which can generate unrealistic or distorted results. Post-processing strategies involve adjustments applied after training and include techniques such as feature projec-

tion [Chuang *et al.*, 2023; Ratzlaff *et al.*, 2024] and prompt engineering [Friedrich *et al.*, 2023; Qiao *et al.*, 2025]. Although these methods do not require fine-tuning or retraining, they typically focus only on single modalities, thus failing to use VLMs' shared embedding spaces. Furthermore, many existing approaches rely on training datasets that are closely related to their test datasets, often partitioned from similar face-centric datasets (*e.g.*, FairFace [Karkkainen and Joo, 2021]), which limits their generalizability and hinders consistent performance across diverse datasets.

To address these limitations, we propose CABIN, a novel backdoor adjustment framework that mitigates the influence of sensitive attributes on VLM predictions (Fig. 1). CABIN uses a shallow neural network "mapper" to transform text embeddings into image embeddings, allowing us to manipulate sensitive attribute features within the shared embedding space without altering the image. This is a key advantage, as it will enable us to directly address biases within the VLM's core representation. Trained on data similar to the VLMs, this mapper avoids overfitting without requiring annotated face-centric datasets. By manipulating sensitive attribute features in the image embedding through text, CABIN fully leverages the VLMs' shared space. We may identify biased attribute subspaces within this shared embedding space by finding features that vary under different sensitive attributes and locating the features of the image embedding that encode sensitive attributes. By neutralizing these subspaces, *e.g.* removing these attribute-related components, CABIN makes the image representations less sensitive to these attributes. Importantly, CABIN does not require model retraining for different test sets. By neutralizing these subspaces, CABIN enables debiasing via backdoor adjustments, a causal inference technique that allows us to estimate the causal effect of the input on the prediction by controlling for confounding factors (*i.e.*, sensitive attributes), facilitating accurate causal estimation. This procedure is independent of domain-specific architectures and task-specific data, making it applicable to various tasks, including zero-shot classification and image retrieval.

Our key contributions are as follows. (1) We develop CABIN, a new debiasing framework grounded in causal inference. CABIN uses a lightweight mapper to manipulate sensitive attributes within image embeddings by leveraging text, avoiding overfitting specific datasets and ensuring effective utilization of both image and text modalities. (2) We introduce a task-agnostic structural causal model for VLM inference. This model explicitly represents the confounding influence of sensitive attributes on the model's predictions, allowing us to identify and address the relevant backdoor paths for debiasing. (3) We conduct extensive experiments across diverse image classification and retrieval datasets. The results demonstrate that CABIN effectively mitigates bias, as quantified by multiple fairness metrics, while maintaining competitive performance.

2 Related Work

VLMs Bias Evaluation. VLMs, such as CLIP, excel in unifying visual and textual modalities, enabling robust zero-shot generalization across downstream tasks [Parraga *et al.*, 2023].

Study	Tasks	CS	MP	TA	FD	SG
Wang *et al.* [2020]	IR	–	+	–	–	G
Berg *et al.* [2022]	IC, IR	–	+	+	–	GR
Seth *et al.* [2023]	IC, VC	–	+	+	–	GRA
Jung *et al.* [2024]	IC, IR, CP, IG	–	+	+	–	GR
Friedrich *et al.* [2022]	IG	–	+	–	+	GR
Chuang *et al.* [2023]	IC, IR	–	–	+	+	GRA
Ratzlaff *et al.* [2024]	VQA	–	+	+	+	R
Weng *et al.* [2024]	OD	+	–	+	–	G
Cheong *et al.* [2024]	VC, AC	+	+	+	+	G
Patil *et al.* [2023]	VQA	+	+	–	+	-
Zhang *et al.* [2023]	HCR	+	+	+	–	GRA
Our Work	IC, IR	+	+	+	+	GRA

Table 1: Comparison of debiasing approaches in Vision-Language Models. Tasks include IC (Image Classification), IR (Text-to-Image Retrieval), VC (Video Classification), CP (Image Captioning), IG (Image Generation), VQA (Visual Question Answering), OD (Object Detection), AC (Audio Classification), and HCR (Hair Color Recognition). CS indicates whether the approach leverages causal strategies (+: Yes, –: No); MP indicates the absence of pixel-level image modification (+: Yes, –: No); TA denotes task-agnostic design (+: Yes, –: No); FD reflects reliance on face-centric datasets for training (+: No reliance, –: Relies on face data). SG specifies the social groups addressed: G (Gender), R (Race), and A (Age).

These models effectively align image and text embeddings effectively, supporting applications such as image classification and retrieval. However, their reliance on large-scale, uncurated datasets often perpetuates societal biases [Birhane *et al.*, 2021], raising fairness concerns. Early fairness evaluations primarily relied on datasets containing human images [Wang *et al.*, 2021; Weng *et al.*, 2024], with face-centric datasets, FairFace [Karkkainen and Joo, 2021] being a common choice [Zhang *et al.*, 2023; Berg *et al.*, 2022]. Text-image pair datasets with sensitive attribute annotations were also used to assess the fairness of VLMs [Jung *et al.*, 2024].

Debiasing VLMs. Efforts to mitigate biases in multimodal models typically adapt single-modality techniques such as adversarial training [Zhang *et al.*, 2018], feature disentanglement [Locatello *et al.*, 2019], and data augmentation [Choi *et al.*, 2020; Lim *et al.*, 2023]. However, extending these methods to VLMs is challenging due to the complexity of joint embedding spaces and the high cost of retraining. Post-processing approaches, such as feature projection [Chuang *et al.*, 2023; Ratzlaff *et al.*, 2024] and prompt engineering [Friedrich *et al.*, 2023], avoid retraining but often fail to address cross-modal biases. Other methods [Jung *et al.*, 2024; Seth *et al.*, 2023] disentangle visual and textual features but rely on curated datasets, risking overfitting and limiting generalization. Our approach leverages post-processing by manipulating embeddings without altering image pixels or model architecture. Trained on open-source image-text pairs [Schuhmann *et al.*, 2021], CABIN reduces overfitting while identifying sensitive attribute subspaces in the joint embedding space, which generalises across tasks, including classification and retrieval.

Debias using Causal Inference. Causal inference offers an approach to disentangle spurious associations from genuine causal relationships by introducing interventions that modify variable distributions while keeping others unchanged. Re-

cent research [Weng *et al.*, 2024; Patil *et al.*, 2023; Zhang *et al.*, 2023] apply causal inference to debias machine learning models but often target specific tasks or single modalities, limiting generalizability. We propose a task-agnostic framework for modelling causal relationships in VLM predictions. Our approach removes spurious correlations without task-specific adaptations by treating sensitive attributes as confounders and applying backdoor adjustment. Experiments demonstrate its effectiveness in mitigating bias across diverse tasks while preserving model performance. Table 1 summarizes the main characteristics of existing compared to ours.

3 Preliminaries

Vision-Language Models (VLMs) such as CLIP [Radford *et al.*, 2021] represent a class of foundational models trained on large-scale image-text pairs. These models learn a shared latent space where both images and text are embedded into a common representation. VLMs are trained to align these embeddings such that related image-text pairs are close in the joint embedding space.

Bias in Vision-Language Models. Although VLMs demonstrate strong zero-shot capabilities, they may inherit biases from their pre-training data [Birhane *et al.*, 2021; Gustafson *et al.*, 2023]. Such biases may manifest as systemic performance disparities across different demographic attributes. Consider a set of m sensitive attribute groups $\mathcal{A} = \{A_1, A_2, \ldots, A_m\}$, and each $A_i = \{a_i^1, a_i^2, \ldots, a_i^n\}$ is a set of categories corresponding to a particular sensitive attribute dimension where n is the number of sensitive attributes in sensitive attribute group A_i. Let $D = \{(x_1, \{a_i^j\}_{i=1}^m), (x_2, \{a_i^j\}_{i=1}^m), \ldots, (x_k, \{a_i^j\}_{i=1}^m)\}$ be an image dataset with sensitive attribute annotations, *i.e.*, each image x_k has a corresponding annotation set $\{a_i^j\}_{i=1}^m$. For instance, if A_1 corresponds to race with categories $\{a_1^1, a_1^2, a_1^3\}$ (*e.g.*, Black, Asian, White) and A_2 corresponds to age with categories $\{a_2^1, a_2^2\}$ (*e.g.*, young, old), then annotation $\{a_1^2, a_2^1\}$ might indicate that x_i is an image of a young Asian person. We focus on the event where the VLM's top similarity score for an image x_i is at least a certain threshold ϵ. Among all the textual prompts $\mathcal{T}$ used to compute similarity, let:

$$M(k) = \mathbb{I}\big(\max_{t \in \mathcal{T}} \text{sim}(x_k, t) \geq \epsilon\big), \tag{1}$$

where $\mathbb{I}(\cdot)$ is an indicator function. The binary result $M(k)$ indicates whether the model's confidence, as measured by the maximum similarity to any class, surpasses the threshold ϵ for the image x_k. This may be interpreted as bias if the proportion of images meeting this criterion differs significantly across categories within an attribute group. Alternatively, ϵ could be tuned as a hyperparameter to balance the tradeoff between the number of high-confidence results and the sensitivity of bias measurement. To obtain high-confidence results for the model, we set ϵ to 0.5.

Fairness Metrics. To quantify bias more interpretably, we examine how high-confidence predictions are distributed across the categories of each attribute group $A_i \in \mathcal{A}$ by measuring the relative representation of these predictions for each category within an attribute group. This approach enables us to identify disparities in how the model treats different categories, such as age or race. Consider one attribute group $A_i = \{a_i^1, a_i^2, \ldots, a_i^n\}$, let the subset of the dataset D containing a_i^j attribute be $D_{a_i^j}$. For each category a_i^j in A_i, where $1 \leq j \leq n$, define:

$$f_i^j = \frac{|D_{a_i^j}|}{|D|}, \tag{2}$$

as the proportion of images in D that belong to category a_i^j of the i-th attribute group. Similarly, let

$$f_i'^j = \frac{|\{x_k \in D_{a_i^j} \mid M(k) = 1\}|}{|\{x_k \in D \mid M(k) = 1\}|}, \tag{3}$$

be the proportion of images that both belong to category a_i^j and meet the confidence criterion ($M(k) = 1$) among all high-confidence classifications. We compute the disparity using the Skew value for each category a_i^j:

$$\text{Skew}(a_i^j) = \log\left(\frac{f_i^j}{f_i'^j}\right). \tag{4}$$

A positive $\text{Skew}(a_i^j)$ indicates that the model is disproportionately favouring category a_i^j relative to its baseline frequency in the dataset. In contrast, a negative $\text{Skew}(a_i^j)$ suggests that a_i^j is underrepresented or disadvantaged.

For a given attribute group A_i, we summarise the skew values across all categories $\{a_i^j\}$ using aggregate metrics as follows. The **MaxSkew** is defined as $\max_j \text{Skew}(a_i^j)$, which captures the most favoured category within the attribute group. Conversely, the **MinSkew**, defined as $\min_j \text{Skew}(a_i^j)$, identifies the most disfavored category. Additionally, we use **MaxSkew@k**, which represents the average MaxSkew values of the top-k results within A_i, providing insights into which categories are most favoured collectively. Similarly, **MinSkew@k** computes the average MinSkew values of the top-k results, highlighting the categories that are most disfavored on aggregate.

4 CABIN Framework

This section describes our debiasing approach grounded in causal inference to mitigate stereotyped biases in VLMs. First, we formulate the causal inference framework for VLMs to make predictions. To effectively mitigate biases in VLMs, we employ a causal inference strategy known as backdoor adjustment in Sec. 4.1 and introduce a lightweight mapper that alters image embeddings based on text prompts introduced in Sec. 4.2. By combining the mappers and backdoor adjustment, we effectively identify biased features in embedding VLMs, described in Sec. 4.3.

4.1 Causal Problem Formulation

Our method selectively incorporates only those causal factors relevant to understanding and mitigating bias, omitting other causal relationships that do not pertain to the debiasing objectives. To define the relationships between different

variables in a VLM-based prediction task, *e.g.*, zero-shot image classification, we use a Structural Causal Model (SCM) to model the causal relationships. The model includes three primary variables: $\mathcal{X}$, the input to the VLM, exemplified by an image; $\mathcal{Y}$, the prediction of the VLM, which could be the predicted class label; and $\mathcal{B}$, which represents the sensitive attributes such as gender and age. The causal relationship between these three variables may be graphically represented using SCM in Fig. 2(c).

The VLM learns complex associations from large-scale image-text data. Thus, the input data $\mathcal{X}$ may naturally encode sensitive attributes $\mathcal{B}$ when making predictions. The model will predict the stereotypical bias inherited from training data. Thus, $\mathcal{B}$ also affects the prediction of $\mathcal{Y}$. Since $\mathcal{B}$ affects both $\mathcal{X}$ and $\mathcal{Y}$, we identify $\mathcal{B}$ as a confounding sensitive factor using a backdoor path $\mathcal{X} \leftarrow \mathcal{B} \rightarrow \mathcal{Y}$, which may render the prediction biased. Specifically, the model may favour certain demographic groups or produce systematically different predictions for images with different sensitive attributes. To mitigate this bias, we isolate the causal effect of $\mathcal{X}$ on $\mathcal{Y}$ by removing the influence of the confounding factor $\mathcal{B}$. Specifically, we identify a set of variables, known as a backdoor adjustment set, that blocks all non-causal paths between $\mathcal{X}$ and $\mathcal{Y}$. By conditioning on this set, we eliminate spurious correlations introduced by $\mathcal{B}$, enabling the computation of the genuine causal effect of $\mathcal{X}$ on $\mathcal{Y}$:

$$P(\mathcal{Y} \mid \text{set}(\mathcal{X})) = \sum_{z} P(\mathcal{Y} \mid \mathcal{X}, z) P(z), \tag{5}$$

where $P(\mathcal{Y} \mid \text{set}(\mathcal{X}))$ denotes the interventional distribution of $\mathcal{Y}$ when $\mathcal{X}$ is set at a particular value.

4.2 Text Embedding to Image Embedding Mapper

We use backdoor adjustment to block the backdoor path $\mathcal{X} \leftarrow \mathcal{B} \rightarrow \mathcal{Y}$. Applying the backdoor adjustment formula requires that $\mathcal{X}$ and $\mathcal{B}$ be independent so that conditioning on $\mathcal{B}$ removes the confounding influence. However, $\mathcal{B}$ is typically visibly encoded in the image $\mathcal{X}$ in vision-based tasks. For example, facial features in a nurse's image may reveal gender information, implying that $\mathcal{X}$ and $\mathcal{B}$ are not independent.

To resolve that $\mathcal{X}$ and $\mathcal{B}$ are not independent, we propose training a mapper that transforms from text modality to image modality. This mapper is trained on a large-scale text-image pairs dataset LAION-400-MILLION [Schuhmann *et al.*, 2021]. Since the mapper learns the transformative relationship between text and image embedding of the same sense, we only need a neural network with a small number of parameters, such as Bottleneck [He *et al.*, 2016], to leverage the text encoder to generate embeddings for variants of textual descriptions that differ only in their sensitive attribute terms. By applying the mapper to these textual embeddings, we simulate how changes in sensitive attributes affect image embeddings in the VLM's joint space.

Here, we define the mapping process for aligning sensitive attributes. Let $\phi : \mathcal{I} \rightarrow \mathbb{R}^d$ and $\psi : \mathcal{T} \rightarrow \mathbb{R}^d$ represent the image and text encoders, mapping inputs $x \in \mathcal{I}$ and $t \in \mathcal{T}$ to d-dimensional embeddings $\phi(x)$ and $\psi(t)$, respectively. The mapper $M : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is trained so that $M(\psi(t)) \approx \phi(x)$ for corresponding image-text pairs. Once trained, we identify dimensions or subspaces of the image embedding space that are responsible for encoding the sensitive attribute by analyzing how substituting one attribute (*e.g.*, male) for another (*e.g.*, female) changes $M(\psi(t))$.

As shown in Fig. 2(a), we design a training loss with a combination of two loss terms: the first term $\mathcal{L}_{\text{align}}$ measures the alignment between text embeddings and image embeddings, while the second contrastive term $\mathcal{L}$diff encourages mismatched image-text pairs to be pushed farther apart, enhancing the discriminative ability of the learned mapper. To balance two loss items, we use λ to adjust the weights of loss items. In order to find a proper λ to balance two terms in the loss function, we use PCGrad [Yu *et al.*, 2020], which dynamically adjusts the gradient directions by projecting conflicting gradients onto orthogonal subspaces to reduce interference between losses and ensure a more stable optimization process. This approach aligns with our requirement to simultaneously enhance the mapper's alignment capacity and discriminative ability. Let (x_k, t_k) be a matched image-text pair. We define an arbitrary prompt that does not correspond to x_k in the dataset as $t_{uk} = \{t \in \mathcal{T} \mid t \neq t_k\}$ and let (x_k, t') be a non-matching pair where $t' \in t_{uk}$. We propose the following loss function:

$$\mathcal{L}_{\text{mapper}} = \mathcal{L}_{\text{align}} + \lambda \mathcal{L}_{\text{diff}} \tag{6}$$

where λ is a learnable parameter determined by PCGrad.

Alignment Loss ($\mathcal{L}_{\text{align}}$): The primary goal is to ensure that $M(\psi(t))$ aligns closely with $\phi(x)$. Intuitively, we minimize the distance measure between text and image embeddings:

$$\mathcal{L}^k_{\text{align}} = \frac{1}{d} \sum \big(M(\psi(t_k)) - \phi(x_k)\big)^2. \tag{7}$$

Difference Loss ($\mathcal{L}_{\text{diff}}$): Additionally, to guide the mapper to distinguish matched text-image pairs from mismatched pairs, we introduce a contrastive term:

$$\begin{aligned} \mathcal{L}^k_{\text{diff}} = \max(0, \mu - \text{sim}(M(\psi(t_k)), \phi(x_k)) \\ + \text{sim}(M(\psi(t')), \phi(x_k))) \end{aligned} \tag{8}$$

where $\mu > 0$ is a contrastive loss margin. Incorporating $\mathcal{L}_{\text{diff}}$ into $\mathcal{L}_{\text{mapper}}$ ensures that matched pairs are mapped closer together than mismatched pairs, fostering a more discriminative mapping.

By analyzing pairs of textual descriptions that differ only in their sensitive attribute categories, we identify embedding directions corresponding to specific attribute groups. For a given attribute group A_i (*e.g.* gender), we compute all pairwise differences among the categories (*e.g.* male and female) to form a set of directional vectors:

$$S_{A_i} = \{M(\psi(t_{a_i^p})) - M(\psi(t_{a_i^q})) \mid \forall a_i^p, a_i^q \in A_i, a_i^p \neq a_i^q\}. \tag{9}$$

The S_{A_i} defines a subspace that captures the variations across categories within the attribute group, effectively characterizing biased directions.

To ensure the independence of $\mathcal{X}$ and $\mathcal{B}$ so that the image embedding does not contain information related to sensitive

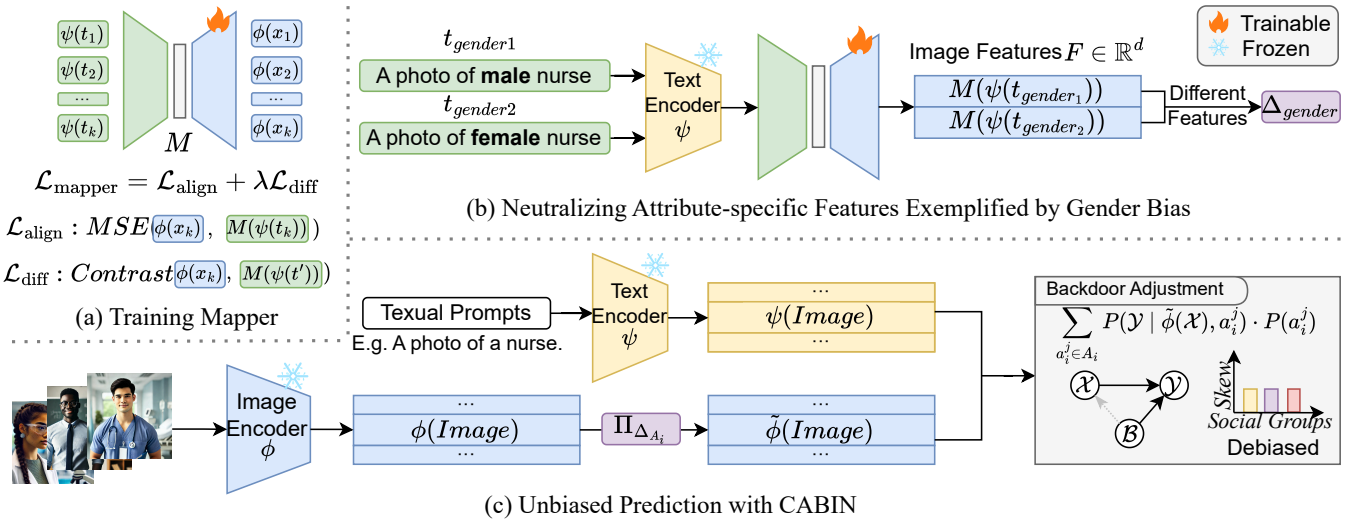

Figure 2: Overview of our proposed debiasing approach. (a) Training the Mapper: We train a lightweight mapper M to align textual and image embeddings. The mapper is guided by an alignment loss $\mathcal{L}_{\text{align}}$ and a difference loss $\mathcal{L}_{\text{diff}}$ to ensure that $M(\psi(t_n))$ closely matches $\phi(i_n)$. (b) Neutralizing Attribute-specific Features Exemplified by Gender Bias: Take gender bias as an example. By examining pairs of textual prompts differing only in a sensitive attribute (*e.g.*, "male nurse" vs. "female nurse"), we extract the attribute embedding direction Δ_{gender} using the mapped textual embeddings and remove the projection of image embeddings onto these attribute directions results in neutralized features $\tilde{\phi}(X)$ that are less sensitive-attribute-dependent. (c) Unbiased Prediction: With neutralized image embeddings and known attribute distributions $P(a_i^j)$, we apply the backdoor adjustment formula. It provides unbiased predictions that are fairer to different social groups.

attributes, we neutralize these attribute-specific components from the image representation:

$$\tilde{\phi}(x) = \phi(x) - \sum_{A_i \in \mathcal{A}} \sum_{\Delta \in S_{A_i}} \Pi_{\Delta}(\phi(x)), \tag{10}$$

where $\Pi_{\Delta}(\cdot)$ denotes the projection onto the subspace defined by all difference vectors Δ in S_{A_i}. After this transformation, $\tilde{\phi}(x)$ is less dependent on sensitive attributes, moving us closer to a scenario where X and these sensitive attributes are independent.

4.3 Applying Backdoor Adjustment

As we have ensured that $\mathcal{X}$ and the sensitive attributes are approximately independent by neutralizing the corresponding attribute subspaces in the image embeddings, we apply the backdoor adjustment to estimate the causal effect of $\mathcal{X}$ on $\mathcal{Y}$, which represents the genuine causal relationship, *i.e.,* the unbiased prediction.

To implement the backdoor adjustment, we first compute a generalised outcome distribution conditioned on the neutralized embedding $\tilde{\phi}(\mathcal{X})$ under each category $a_i^j \in A_i$ for every attribute group $A_i \in \mathcal{A}$. This involves providing the VLM with the neutralized visual representation $\tilde{\phi}(\mathcal{X})$ as well as the textual inputs $\{\psi(t)\}$ that vary depending on the downstream task. The VLM's inference function $f(\tilde{\phi}(X), \{\psi(t)\}, \theta)$, parameterised by θ, then produces a conditional outcome distribution. For each $a_i^j \in A_i$:

$$P(\mathcal{Y} \mid \tilde{\phi}(\mathcal{X}), a_i^j) = \text{Norm}(f(\tilde{\phi}(\mathcal{X}), \{\psi(t)\}, \theta)), \tag{11}$$

where $\text{Norm}(\cdot)$ is a normalizing function (*e.g.*, softmax for categorical outcomes) chosen according to the requirements of the downstream task. Next, the probabilities $P(a_i^j)$ for each category a_i^j are estimated from the distribution of attributes in the test data D_{test}:

$$P(a_i^j) = \frac{|D_{test}^{a_i^j}|}{|D_{test}|}, \tag{12}$$

where $D_{test}^{a_i^j}$ represents the subset of D_{test} containing a_i^j attribute.

Finally, we aggregate over all weighted conditional outcomes for each attribute group. For a single attribute group A_i:

$$P(\mathcal{Y} \mid \text{set}(\mathcal{X})) = \sum_{a_i^j \in A_i} P(\mathcal{Y} \mid \tilde{\phi}(\mathcal{X}), a_i^j) \cdot P(a_i^j). \tag{13}$$

Extending to multiple attribute groups, each A_i is adjusted independently. Fig. 2(c) illustrates the process of applying backdoor adjustment. This approach is agnostic to the specific downstream task, relying only on the availability of textual inputs $\{\psi(t)\}$ to guide the VLM's prediction. By performing a backdoor adjustment on the neutralized embeddings, the outcome reflects the genuine causal relationships free from confounding sensitive attributes.

5 Experiments

We structure our experimental study around two key research questions (RQs): How effectively does our method reduce bias? **(RQ1)** and How does our approach balance debiasing with preserving task performance? **(RQ2)**. We focus on two representative tasks, image classification and image retrieval,

			MaxSkew↓				MinSkew↑				MaxSkew@k↓				MinSkew@k↑			
			VA	DR	SF	CB	VA	DR	SF	CB	VA	DR	SF	CB	VA	DR	SF	CB
FACET	ResNet50	Gender	0.36	0.19	0.22	**0.15**	-0.25	-0.15	-0.17	**-0.10**	0.48	0.35	0.32	**0.20**	-2.58	-2.30	-2.42	**-1.85**
		Race	0.54	0.33	0.30	**0.25**	-0.53	-0.35	-0.33	**-0.20**	0.55	0.40	0.38	**0.25**	-3.19	-2.50	-2.51	**-1.65**
		Age	0.49	0.37	0.35	**0.20**	-0.85	-0.40	-0.42	**-0.15**	0.42	0.30	0.34	**0.18**	-3.47	-2.89	-2.90	**-2.00**
	ViT-B/32	Gender	0.35	0.18	0.21	**0.13**	-0.24	-0.14	-0.16	**-0.09**	0.47	0.33	0.31	**0.19**	-2.57	-2.28	-2.40	**-1.83**
		Race	0.52	0.32	0.29	**0.24**	-0.51	-0.34	-0.32	**-0.19**	0.54	0.39	0.37	**0.24**	-3.17	-2.48	-2.49	**-1.63**
		Age	0.48	0.36	0.34	**0.09**	-0.83	-0.39	-0.41	**-0.14**	0.41	0.29	0.33	**0.17**	-3.45	-2.87	-2.88	**-1.98**
PATA	ResNet50	Gender	0.38	0.20	0.24	**0.15**	-0.27	-0.16	-0.18	**-0.15**	0.50	0.36	0.30	**0.25**	-2.60	-2.35	-2.45	**-2.30**
		Race	0.60	0.35	**0.33**	0.34	-0.55	-0.38	-0.35	**-0.33**	0.60	0.42	0.40	**0.38**	-3.25	-2.60	-2.55	**-2.50**
		Age	0.52	0.38	0.36	**0.30**	-0.90	-0.42	-0.40	**-0.35**	0.45	0.32	0.34	**0.28**	-3.50	-2.90	-2.95	**-2.85**
	ViT-B/32	Gender	0.34	**0.18**	0.22	**0.18**	-0.23	-0.14	-0.17	**-0.12**	0.48	0.34	0.29	**0.24**	-2.55	-2.25	-2.35	**-2.20**
		Race	0.58	0.34	0.31	**0.25**	-0.50	-0.33	-0.30	**-0.28**	0.53	0.38	0.35	**0.31**	-3.15	-2.45	-2.40	**-2.35**
		Age	0.50	0.35	0.33	**0.29**	-0.80	-0.38	-0.40	**-0.37**	0.40	0.28	0.26	**0.20**	-3.40	-2.80	-2.85	**-2.70**
Flickr30K	ResNet50	Gender	0.47	0.20	0.29	**0.15**	-0.52	-0.25	-0.27	**-0.23**	0.51	0.29	0.25	**0.20**	-2.98	-2.76	**-2.00**	-2.05
		Race	0.56	0.30	0.26	**0.21**	-0.78	-0.40	-0.43	**-0.31**	0.43	0.35	**0.33**	0.35	-3.09	-2.71	-2.74	**-2.40**
		Age	0.51	0.21	0.28	**0.20**	-0.65	-0.45	**-0.29**	-0.35	0.59	0.20	0.24	**0.17**	-3.45	-2.80	-2.91	**-2.70**
	ViT-B/32	Gender	0.37	0.26	**0.18**	0.22	-0.19	-0.16	-0.20	**-0.15**	0.60	0.20	0.21	**0.10**	-2.55	**-2.25**	-2.39	-2.45
		Race	0.45	0.34	0.31	**0.28**	-0.34	-0.28	-0.30	**-0.23**	0.53	0.40	0.38	**0.31**	-3.15	-2.45	-2.38	**-2.36**
		Age	0.44	0.38	0.35	**0.31**	-0.38	-0.42	-0.37	**-0.29**	0.57	0.28	0.30	**0.27**	-3.45	-2.80	-2.83	**-2.69**
FairFace	ResNet50	Gender	0.42	0.25	0.23	**0.18**	-0.50	-0.35	-0.32	**-0.20**	0.47	0.33	0.29	**0.23**	-2.70	-2.40	-2.20	**-2.10**
		Race	0.55	0.32	0.30	**0.27**	-0.62	-0.38	-0.40	**-0.25**	0.50	0.35	0.33	**0.29**	-3.10	-2.60	-2.50	**-2.40**
		Age	0.50	0.35	0.31	**0.28**	-0.80	-0.38	-0.36	**-0.28**	0.49	0.30	0.29	**0.22**	-3.25	-2.70	-2.65	**-2.50**
	ViT-B/32	Gender	0.40	**0.22**	0.28	0.25	-0.45	-0.30	-0.33	**-0.25**	0.48	**0.34**	0.36	0.37	-2.60	-2.40	-2.35	**-2.30**
		Race	0.50	0.35	0.39	**0.32**	-0.48	-0.30	-0.28	**-0.20**	0.46	0.36	0.40	**0.33**	-3.00	-2.60	**-2.50**	-2.55
		Age	0.44	0.33	0.35	**0.26**	-0.44	-0.38	-0.42	**-0.24**	0.40	0.29	**0.27**	0.31	-3.20	-2.70	-2.75	**-2.60**
MS	ResNet50	Gender	0.41	0.25	0.28	**0.18**	-0.51	-0.32	-0.29	**-0.20**	0.47	0.31	0.32	**0.22**	-2.70	-2.50	-2.55	**-2.40**
	ViT-B/32	Gender	0.42	0.27	0.29	**0.23**	-0.46	-0.30	-0.28	**-0.23**	0.44	0.32	0.35	**0.25**	-2.62	-2.50	-2.40	**-2.30**
PS	ResNet50	Gender	0.49	0.34	0.36	**0.26**	-0.59	-0.32	-0.30	**-0.24**	0.53	0.38	0.35	**0.29**	-2.90	-2.60	-2.52	**-2.41**
	ViT-B/32	Gender	0.39	0.25	0.27	**0.19**	-0.49	-0.34	-0.31	**-0.20**	0.48	0.33	0.30	**0.21**	-2.80	-2.56	-2.50	**-2.35**

Table 2: Comparison of Skew metrics across diverse datasets for fairness evaluation. Results include MaxSkew, MinSkew, MaxSkew@k, and MinSkew@k metrics for Vanilla CLIP models (VA), DeAR (DR), SFID (SF), and CABIN (CB). Standard deviations for results are below 0.01, which is less than the precision shown, omitted for clarity. We conducted paired t-tests for each metric across repeated runs, applying Holm–Bonferroni correction. The performance differences are statistically significant (all adjusted $p < 0.05$). **Bold** numbers denote the best (lower for MaxSkew, higher for MinSkew), and underlined numbers denote the second best.

to provide a holistic perspective on how bias arises and is mitigated in VLMs.

Baseline and Architecture. All experiments use Vanilla CLIP encoders with ResNet50 and ViT-B/32 backbones as baselines. DeAR [Seth *et al.*, 2023] and SFID [Jung *et al.*, 2024] serve as comparative methods. We report Skew-based metrics mentioned in Sec. 3.

Data and Training. As mainstream VLMs such as CLIP's training data have not been published, we use a large-scale open-source image-text dataset LAION-400-MILLION [Schuhmann *et al.*, 2021] to train our mapper M. We randomly sampled 10 million paired image-text data from the dataset to ensure $M(\psi(t))$ approximates $\phi(x)$ accurately.

Evaluation Datasets. We use FACET [Gustafson *et al.*, 2023], PATA [Seth *et al.*, 2023], and Flickr30K [Plummer *et al.*, 2015] to evaluate our debiasing method. Traditional face-centric datasets such as FairFace [Karkkainen and Joo, 2021], MS-COCO (MS) [Lin *et al.*, 2014], and Pascal-Sentence (PS) [Rashtchian *et al.*, 2010] are also used to show our method applies to various ranges of datasets and tasks.

6 Results and Discussions

We present analyses of our proposed approach to address research questions (RQ1 and RQ2) outlined in Sec. 5. We further include a parameter sensitivity analysis demonstrating how our choice of hyperparameters influences both bias mitigation and overall performance.

Evaluation of Skew-based Fairness (RQ1). Table 2 presents Skew-based metrics across diverse datasets and sensitive attributes. Our method consistently reduces MaxSkew and MaxSkew@k compared to the vanilla CLIP model (VA) and strong debiasing baselines (DeAR, SFID). Additionally, it shows comparable performance in image-to-text retrieval on Flickr30K while effectively lowering Skew values for sensitive attributes.

Figure 3 further illustrates these improvements through visual examples, where we use GradCAM [Selvaraju *et al.*, 2017] to reveal how our method shifts focus away from sensitive attributes such as faces to more contextually relevant features when we use the CLIP model to classify images with different sensitive attributes in different scenarios.

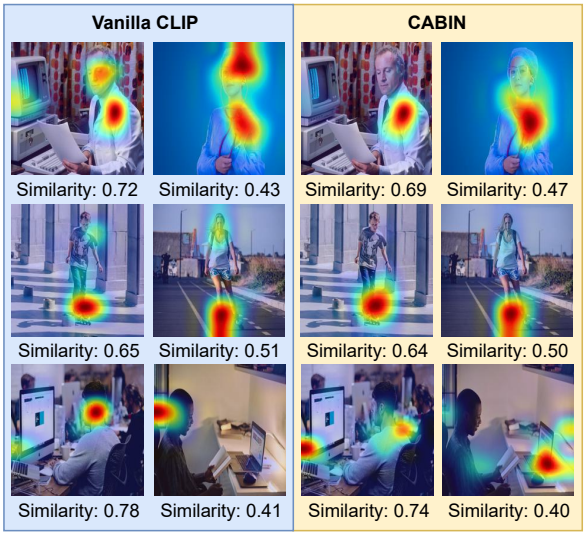

Figure 3: Comparison of visualizations between the Vanilla CLIP model representing $\phi(x)$ with ViT-B/32 backbone and CABIN representing $\tilde{\phi}(x)$. Each row corresponds to a different sensitive attribute bias using the prompts "A photo of a (doctor/person skating/person working)" respectively. In each scenario, one image depicts a socially dominant group member and the other a less-represented group. After applying CABIN, the model focuses less on features related to sensitive attributes and more on features related to prompts.

		FACET	ImageNet	Flickr30K		
		Accuracy(%)		R@1	R@5	R@10
ResNet50	VA	51.88±.56	57.92±.70	57.20±.55	81.67±.68	88.17±.59
	DR	51.36±.44	57.09±.52	57.15±.61	8.87±.63	87.20±.70
	SF	51.07±.60	56.92±.65	56.65±.66	8.63±.55	87.16±.66
	CB	51.65±.42	57.58±.45	57.02±.64	81.31±.51	87.89±.63
ViT-B/32	VA	52.19±.57	58.92±.64	58.90±.66	83.00±.72	89.20±.59
	DR	51.41±.67	57.86±.61	58.15±.58	82.03±.55	89.03±.50
	SF	51.55±.53	58.72±.61	57.93±.51	81.84±.63	88.54±.68
	CB	51.80±.48	58.56±.55	58.61±.53	82.70±.60	88.84±.64

Table 3: Comparison of performance across three datasets: FACET and ImageNet for image classification (Accuracy) and Flickr30K for text-to-image retrieval (R@1, R@5, R@10). We compare four methods, including Vanilla CLIP (VA), DeAR (DR), SFID (SF), and CABIN (CB), on two CLIP backbones (ResNet50, ViT-B/32). Results are shown as mean±std (%).

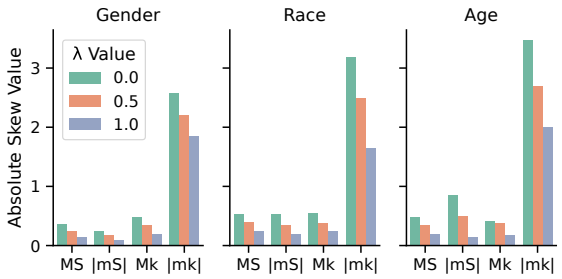

Figure 4: Effect of varying λ on Skew-based fairness metrics for the FACET dataset using ResNet50 backbone. Negative Skew values are normalized to absolute values for visual clarity.

Evaluation of Accuracy and Recall-based Performance (RQ2). Although our primary goal is bias mitigation, preserving task performance is also essential. Table 3 reports zero-shot classification accuracy and image retrieval Recall@k before and after debiasing, comparing our method with unmodified CLIP models and two baselines (DeAR, SFID). Our method maintains competitive performance while achieving improved fairness.

Parameter Analysis. Intuitively, a loss function containing only $\mathcal{L}_{\text{diff}}$ is unsuitable, as a mapper trained solely on mismatched image–text pairs cannot accurately transform text embeddings into image embeddings. The weighting factor λ balances the alignment loss $\mathcal{L}_{\text{align}}$ and the contrastive difference loss $\mathcal{L}_{\text{diff}}$, influencing how the mapper trades off embedding fidelity against discriminative power. While PCGrad [Yu *et al.*, 2020] is used to mitigate gradient conflicts between these two objectives, λ itself remains a fixed hyperparameter. To study its impact, we evaluate three settings ($\lambda = 0$, $\lambda = 0.5$, and $\lambda = 1$). As shown in Figure 4, different λ values yield distinct trade-offs between bias mitigation and task performance. When λ is near zero, the model prioritizes alignment. The mapper lacks the ability to distinguish between matched and mismatched pairs, resulting in weak attribute direction signals extracted from attribute pairs. Conversely, at $\lambda = 1$, the contrastive term dominates: attribute subspaces become more sharply defined, yielding stronger bias mitigation but at the cost of embedding alignment and occasional drops in performance. An intermediate setting ($\lambda = 0.5$) achieves moderate bias reduction while retaining reasonable task accuracy. However, as shown in Table 2 and Table 3, even $\lambda = 0.5$ falls short of the stable bias–performance balance achieved when using PCGrad to dynamically reconcile both losses; manually tuning λ across fixed values is also computationally expensive for only marginal gains.

7 Conclusion

We present CABIN, a novel debiasing framework that systematically mitigates stereotypical biases in Vision-Language Models (VLMs) through causal inference while preserving the models' original capabilities. By using a lightweight mapper to identify and neutralize sensitive attribute directions in the joint embedding space and applying backdoor adjustment to remove confounding influences, CABIN substantially reduces Skew-based bias across demographic groups on downstream tasks (*e.g.*, image classification and text-to-image retrieval) without requiring retraining. CABIN focuses on debiasing the model rather than specific downstream tasks, such that it can be easily extended to other applications. Experiments and ablation studies on multiple benchmarks demonstrate that CABIN achieves a favourable balance between fairness enhancement and performance retention, underscoring its adaptability to diverse real-world scenarios.

Acknowledgements

This research was supported by the Our Voices study (https://ourvoices.auckland.ac.nz), funded by the Ministry of Business, Innovation and Employment (2019–2025) under Endeavour grant UOAX1912, and a University of Auckland Masters scholarship awarded to the first author. The authors also gratefully acknowledge the Centre for Machine Learning for Social Good at the University of Auckland for supporting the first author's travel to present this work. The authors are especially grateful to the rangatahi / youth who participated in the Our Voices study, designed by Susan M.B. Morton and the Our Voices study team.

References

[Berg *et al.*, 2022] 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 Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang, editors, *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, Online only, November 2022. Association for Computational Linguistics.

[Birhane *et al.*, 2021] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.

[Choi *et al.*, 2020] Kristy Choi, Aditya Grover, Trisha Singh, Rui Shu, and Stefano Ermon. Fair generative modeling via weak supervision. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

[Chuang *et al.*, 2023] 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.

[Friedrich *et al.*, 2023] Felix Friedrich, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Patrick Schramowski, Sasha Luccioni, and Kristian Kersting. Fair diffusion: Instructing text-to-image generation models on fairness. *arXiv preprint arXiv:2302.10893*, 2023.

[Gustafson *et al.*, 2023] Laura Gustafson, Chloe Rolland, Nikhila Ravi, Quentin Duval, Aaron Adcock, Cheng-Yang Fu, Melissa Hall, and Candace Ross. FACET: Fairness in Computer Vision Evaluation Benchmark. In *2023 IEEE/CVF International Conference on Computer Vision*, pages 20313–20325, Paris, France, October 2023. IEEE.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2016.

[Hong *et al.*, 2024] Rachel Hong, William Agnew, Tadayoshi Kohno, and Jamie Morgenstern. Who's in and who's out? a case study of multimodal clip-filtering in datacomp. In *Proceedings of the 4th Association for Computing Machinery Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '24, New York, NY, USA, 2024. Association for Computing Machinery.

[Jung *et al.*, 2024] Hoin Jung, Taeuk Jang, and Xiaoqian Wang. A unified debiasing approach for vision-language models across modalities and tasks. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[Karkkainen and Joo, 2021] Kimmo Karkkainen and Jungseock Joo. 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*, pages 1548–1558, 2021.

[Lee *et al.*, 2023] Nayeon Lee, Yejin Bang, Holy Lovenia, Samuel Cahyawijaya, Wenliang Dai, and Pascale Fung. Survey of Social Bias in Vision-Language Models, September 2023.

[Lim *et al.*, 2023] Jongin Lim, Youngdong Kim, Byungjai Kim, Chanho Ahn, Jinwoo Shin, Eunho Yang, and Seungju Han. BiasAdv: Bias-Adversarial Augmentation for Model Debiasing. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3832–3841, Vancouver, BC, Canada, June 2023. IEEE.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing.

[Locatello *et al.*, 2019] Francesco Locatello, Stefan Bauer, Mario Lučić, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Frederic Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, 2019. Best Paper Award.

[Lülf *et al.*, 2024] Christian Lülf, Denis Mayr Lima Martins, Marcos Antonio Vaz Salles, Yongluan Zhou, and Fabian Gieseke. Clip-branches: Interactive fine-tuning for text-image retrieval. In *Proceedings of the 47th International Association for Computing Machinery SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 2719–2723, New York, NY, USA, 2024. Association for Computing Machinery.

[Parraga *et al.*, 2023] Otavio Parraga, Martin D. More, Christian M. Oliveira, Nathan S. Gavenski, Lucas S. Kupssinskü, Adilson Medronha, Luis V. Moura, Gabriel S. Simões, and Rodrigo C. Barros. Fairness in Deep Learning: A Survey on Vision and Language Research. *Association for Computing Machinery Computing Surveys*, page 3637549, December 2023.

[Patil *et al.*, 2023] Vaidehi Patil, Adyasha Maharana, and Mohit Bansal. Debiasing multimodal models via causal information minimization. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: Empirical Methods in Natural Language Processing 2023*, pages 4108–4123, Singapore, December 2023. Association for Computational Linguistics.

[Plummer *et al.*, 2015] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

[Qiao *et al.*, 2025] Tingrui Qiao, Caroline Walker, Chris Cunningham, and Yun Sing Koh. Thematic-lm: A llm-based multi-agent system for large-scale thematic analysis. In *Proceedings of the Association for Computing Machinery on Web Conference 2025*, WWW '25, page 649–658, New York, NY, USA, 2025. Association for Computing Machinery.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.

[Rashtchian *et al.*, 2010] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. Collecting image annotations using Amazon's Mechanical Turk. In Chris Callison-Burch and Mark Dredze, editors, *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147, Los Angeles, June 2010. Association for Computational Linguistics.

[Ratzlaff *et al.*, 2024] Neale Ratzlaff, Matthew Lyle Olson, Musashi Hinck, Shao-Yen Tseng, Vasudev Lal, and Phillip Howard. Debiasing large vision-language models by ablating protected attribute representations. In *Neurips Safe Generative AI Workshop 2024*, 2024.

[Schuhmann *et al.*, 2021] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[Selvaraju *et al.*, 2017] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 Institute of Electrical and Electronics Engineers International Conference on Computer Vision*, pages 618–626, 2017.

[Seth *et al.*, 2023] Ashish Seth, Mayur Hemani, and Chirag Agarwal. Dear: Debiasing vision-language models with additive residuals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6820–6829, June 2023.

[Wang *et al.*, 2021] Jialu Wang, Yang Liu, and Xin Wang. Are gender-neutral queries really gender-neutral? mitigating gender bias in image search. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1995–2008, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[Weng *et al.*, 2024] Zhaotian Weng, Zijun Gao, Jerone Andrews, and Jieyu Zhao. Images speak louder than words: Understanding and mitigating bias in vision-language model from a causal mediation perspective. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15669–15680, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[Yu *et al.*, 2020] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[Zhang *et al.*, 2018] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 Association for the Advancement of Artificial Intelligence/Association for Computing Machinery Conference on AI, Ethics, and Society*, AIES '18, page 335–340, New York, NY, USA, 2018. Association for Computing Machinery.

[Zhang *et al.*, 2022] Yi Zhang, Junyang Wang, and Jitao Sang. Counterfactually measuring and eliminating social bias in vision-language pre-training models. In *Proceedings of the 30th Association for Computing Machinery International Conference on Multimedia*, MM '22, page 4996–5004, New York, NY, USA, 2022. Association for Computing Machinery.

[Zhang *et al.*, 2023] Yi Zhang, Jitao Sang, Junyang Wang, Dongmei Jiang, and Yaowei Wang. Benign shortcut for debiasing: Fair visual recognition via intervention with shortcut features. In *Proceedings of the 31st Association for Computing Machinery International Conference on Multimedia*, MM '23, page 8860–8868, New York, NY, USA, 2023. Association for Computing Machinery.

[Zhao *et al.*, 2024] Di Zhao, Yun Sing Koh, Gillian Dobbie, Hongsheng Hu, and Philippe Fournier-Viger. Symmetric self-paced learning for domain generalization. *Proceedings of the Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, 38(15):16961–16969, Mar. 2024.