

# PRISM: Reducing Spurious Implicit Biases in Vision-Language Models with LLM-Guided Embedding Projection

Mahdiyar Molahasani\*, Azadeh Motamed\*, Michael Greenspan, Il-Min Kim, Ali Etemad  
Queen's University, Canada

{m.molahasani, 19am43, michael.greenspan, ilmin.kim, ali.etemad}@queensu.ca

## Abstract

We introduce *Projection-based Reduction of Implicit Spurious bias in vision-language Models (PRISM)*, a new data-free and task-agnostic solution for bias mitigation in VLMs like CLIP. VLMs often inherit and amplify biases in their training data, leading to skewed predictions. PRISM is designed to debias VLMs without relying on predefined bias categories or additional external data. It operates in two stages: first, an LLM is prompted with simple class prompts to generate scene descriptions that contain spurious correlations. Next, PRISM uses our novel contrastive-style debiasing loss to learn a projection that maps the embeddings onto a latent space that minimizes spurious correlations while preserving the alignment between image and text embeddings. Extensive experiments demonstrate that PRISM outperforms current debiasing methods on the commonly used Waterbirds and CelebA datasets. We make our code public at: <https://github.com/MahdiyarMM/PRISM>.

## 1. Introduction

Vision-Language Models (VLMs) have emerged as a class of machine learning models capable of understanding and generating multi-modal content by jointly processing images and textual descriptions. Prominent VLMs, such as Contrastive Language-Image Pre-training (CLIP) [20], have demonstrated impressive performance in zero-shot classification and cross-modal retrieval. By aligning textual and visual representations in a shared embedding space, these models enable seamless interactions between vision and language tasks. Despite their unprecedented effectiveness, VLMs inherit and propagate systemic biases from their large-scale pretraining data, leading to inaccurate or skewed predictions.

Biases in VLMs manifest in various forms, including social [19, 27, 30, 33] and contextual biases [26, 31], of-

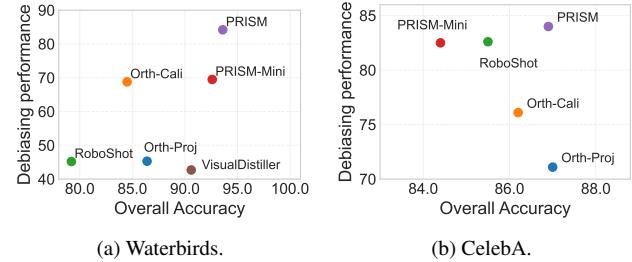


Figure 1. Performance comparison among data-free CLIP debiasing methods in terms of debiasing performance (i.e., the accuracy of the worst group) vs. the overall accuracy on (a) Waterbirds and (b) CelebA datasets. Our proposed method, PRISM, achieves state-of-the-art debiasing performance (higher worst-group accuracy) without compromising overall accuracy.

ten influenced by ‘spurious correlations’ in training data. Spurious correlations can stem from associating non-target attributes, such as image backgrounds or contextual features, with labels in unintended ways [5]. For instance, CLIP’s zero-shot classifiers frequently associate ‘camel’ with ‘desert’ and ‘panda’ with ‘bamboo’, reflecting underlying biases in training data. These biases can have real-world implications reducing model robustness in diverse settings. Therefore, mitigating implicit bias in VLMs has become a critical research direction.

Recent research has investigated various strategies to *identify*, *quantify*, and *reduce* spurious correlations in VLMs. These approaches can be classified into three categories: feature-space transformations [7, 29], prompt-based debiasing [1, 5, 6], and adapter design [17]. Methods with a primary focus on the feature space [7, 29] attempt to transform feature embeddings to eliminate spurious correlations. Prompt-based techniques [1, 5, 6] aim to redefine textual inputs to mitigate bias but often struggle with generalization. Meanwhile, adapter-based approaches such as [17] address implicit bias in VLMs by modifying representations at various stages of the model, including but not limited to the final layer.

\*These authors contributed equally.

Despite the effectiveness of these approaches in addressing spurious bias in VLMs, they fall short in many instances. These methods mostly require fixed predefined bias categories makes them difficult to scale to task-agnostic settings [5–7]. Furthermore, these debiasing solutions either compromise the overall performance of the model by removing essential semantic content [1, 5] or rely on external visual data for debiasing [7, 17, 29]. This underscores the need for a new debiasing method without the reliance on predefined bias categories or any external data that can preserve essential semantic content without compromising overall accuracy.

To address these limitations, we propose **P**rojection-based **R**eduction of **I**mplicit **S**purious **b**ias in vision-language **M**odels (PRISM). PRISM consists of two key stages. First, we prompt an LLM with labels to return a list of potential spurious connections corresponding to these labels and dynamically generate biased scene disruptions based on the extracted spurious attributes. Next, we find a projection of the embedding space that can effectively remove the impact of spurious attributes using the generated biased scene descriptions. To this end, we propose a novel loss named Latent space Debiasing loss (LD) to learn the debiasing projection in the embedding space effectively. Our proposed LD loss is a contrastive-style debiasing loss that aligns text and image embeddings based on semantic meaning rather than spurious correlations. It ensures that the text embeddings of the same class but different spurious attributes remain similar to each other, and the text embeddings of different classes but the same spurious attribute remain distinct.

By removing the impact of bias attributes from the embedding space, PRISM generates representations with no or reduced statistical dependence on the bias. To measure this dependency, previously proposed methods mostly employ a metric capable of capturing statistical dependencies, such as Mutual Information [22], Hilbert-Schmidt Independence Criterion (HSIC) [4], or bias scores [7]. In contrast, our proposed method eliminates the need for direct dependency measurement by utilizing an LLM to generate bias-aware descriptions. These descriptions capture common bias attributes in the dataset using only the labels. Our approach facilitates debiasing without the need for predefined bias categories, external text or image datasets, or lossy data conversions to identify bias features. This makes PRISM an effective, task-agnostic, and data-free solution for debiasing VLMs.

We assess the performance of PRISM on CLIP using different benchmarks, including Waterbirds and CelebA datasets, and compare its performance against various baselines. The results highlight the capability of the proposed PRISM in mitigating spurious biases in VLMs without compromising the overall performance. Figure 1 illustrates

that our proposed PRISM achieves state-of-the-art debiasing performance without compromising the total accuracy on both Waterbirds and CelebA benchmarks. Our key contributions in this paper can be summarized as follows:

- We leverage LLMs to generate biased textual descriptions that reflect common spurious correlations found in datasets. This provides a dynamic and adaptable way to identify biases without relying on predefined bias categories.
- We propose a novel task-agnostic data-free debiasing method for VLMs that learns a projection of the embedding space, removing the impact of spurious attributes using a novel contrastive style loss function.
- We demonstrate that our proposed method achieves state-of-the-art debiasing performance without compromising the model’s overall accuracy. To allow rapid reproducibility and contribute to the community, we make our code public at: <https://github.com/MahdiyarMM/PRISM>.

## 2. Related Works

**Bias identification in VLMs.** Recent research has placed considerable emphasis on identifying and assessing bias in foundation models [9], including VLMs [2, 25]. The initial step in debiasing involves recognizing and evaluating the sources of bias. Different frameworks proposed in [31] and [11, 21] comprehensively explored various types of biases in vision and text datasets, respectively. Previous works [2, 25, 28] investigated bias in CLIP’s representation space and proposed different methods to mitigate it. These biases cause unfair outcomes in zero-shot classification and reduce performance in certain underrepresented and overlooked categories [25].

**Debiasing VLMs.** The methods proposed in [3, 19, 31] apply mathematical transformations on data to isolate bias in datasets before feeding them into models. However, such transformations may result in information loss. In the field of feature space transformation, FairerCLIP [7] introduced a mathematically rigorous approach to reduce the bias in CLIP’s representations by leveraging kernel Hilbert spaces. SFID [15] proposed a selective feature imputation for debiasing, which combines feature pruning and low-confidence imputation to eliminate biases across tasks and modalities. These methods are grounded in rigorous mathematical principles for transformation but can be computationally expensive and risk removing crucial semantic information which compromises the overall accuracy of the model.

In the field of prompt engineering, VisualDistiller [6] focused on refining visual representations to counter dataset biases. The solutions proposed in [5], namely Orth-cali and Orth-proj, efficiently remove the bias by projecting out the biased components from CLIP’s text embeddings using predefined biased prompts. However, because these bi-

ased prompts are fixed, the method may struggle to generalize to biases that weren't anticipated in the prompts. In the field of adapter designing, DFR [17] suggested debiasing by retraining only the model's final layer. On the other hand, contrastive adapter [32] proposed to insert lightweight adapter modules into the pre-trained model architecture. These adapters are trained using a contrastive loss designed to pull representations from similar groups closer while pushing those from different groups apart.

In a recent study, RoboShot [1] and BendVLM [10] also leveraged language models to extract insights from task descriptions. The primary focus of RoboShot was on improving robustness in underrepresented and adversarial data distributions rather than bias mitigation. BendVLM was proposed to tackle the bias in VLMs by learning task-specific projections via labeled images, with part of the projections computed dynamically during inference. In contrast, our PRISM explicitly targets bias mitigation in CLIP by leveraging LLMs for adaptive bias identification and embedding transformation. PRISM eliminates the need for LLM calls during inference and removes spurious correlations through learning a debiasing projection of the embedding space, ensuring effective bias removal without relying on predefined bias categories or external visual data. Furthermore, unlike prompt-based debiasing techniques that modify only text embeddings, our approach preserves the alignment between text and image representations, ensuring debiased zero-shot classification without compromising the overall accuracy.

### 3. Method

#### 3.1. Problem formulation

Let us consider a multimodal foundation model, such as CLIP [20], which maps images and natural language onto a shared  $d$ -dimensional space using two encoders: an image encoder  $\phi_I(\cdot)$  and a text encoder  $\phi_T(\cdot)$ . In the zero-shot setting, given image  $x$  and a set of text prompts  $\{t_y\}_{y=1}^K$  corresponding to class labels  $y \in \{1, \dots, K\}$ , the prediction is made via

$$\hat{y} = \arg \max_y \langle \phi_I(x), \phi_T(t_y) \rangle, \quad (1)$$

where  $\langle \cdot, \cdot \rangle$  represents the vector inner product. Here, the text prompts are descriptions of the object in each class as:  $t_y = "A \text{ photo of a } <\text{class } y>"$ .

This formulation presumes that  $\phi_I(x)$  predominantly captures the core, task-relevant features of  $x$ . However, real-world data often contain spurious correlations that can cause the encoder to also capture irrelevant cues. The image embedding can be modeled as a (normally, unknown) function  $f$  of a core component and a spurious component:

$$\phi_I(x) = f(h_{\text{core}}(x), h_{\text{spu}}(x)), \quad (2)$$

where  $h_{\text{core}}(x)$  represents the features essential for correctly predicting the label  $y$ , and  $h_{\text{spu}}(x)$  encodes extraneous information (e.g., background elements) that is spuriously correlated with  $y$ .

Assume that the spurious attribute is denoted by  $a$ . Accordingly, we define the *group identity* of each sample as

$$g = (a, y), \quad \text{with } g \in \mathcal{G} = A \times Y, \quad (3)$$

where  $A$  is the set of possible spurious attributes  $a$ , and  $Y = \{1, \dots, K\}$  corresponds to the set of true class labels. For example, in a bird classification task,  $x$  is an image of a bird, and  $y \in \{0, 1\}$  represents the true class where 0 denotes Landbird and 1 denotes Waterbird. The spurious attribute  $a$  belongs to the set  $A = \{\text{land background, water background}\}$ , which indicates the type of background present in the image. Each combination  $(a, y)$  defines a distinct group within  $\mathcal{G}$ , resulting in a total of four groups in this example.

When the model undesirably relies on the spurious component  $h_{\text{spu}}(x)$ , it may assign  $x$  to the incorrect group, leading to misclassification. To mitigate this bias, the objective is to minimize the worst-case error over all groups. Formally, we aim to solve

$$\min_{\theta \in \Theta} \max_{g \in \mathcal{G}} \mathbb{E} \left[ \ell \left( \langle \phi_I(x), \phi_T(t_y) \rangle, y \right) \mid g \right], \quad (4)$$

where  $\ell(\cdot, \cdot)$  denotes the classification loss.

There are several challenges associated with the optimization of Eq. 4. First, optimizing over the parameter space of  $\theta$  is costly due to the number of parameters of the model, CLIP. This is also in contrast to the original purpose of CLIP which is to serve as a universal and general zero-shot classifier (Eq. 1). Second, the identification of spurious attributes  $A$  for an arbitrary set of text prompts is challenging and requires prior knowledge about the task, or an external dataset, which is not always accessible. Moreover, we hypothesize that due to the nature of spurious correlations, similar biases should also be present in the text domain. Hence, spurious correlations identified in the text modality can also be leveraged to debias the image modality. We will empirically verify this hypothesis in Section 5.

#### 3.2. PRISM

To address the challenges mentioned in Section 3.1, we propose **P**rojection-based **R**eduction of **I**mplicit **S**purious bias in vision-language **M**odels (PRISM), which consists of two main steps: (i) PRISM detects potential spurious correlations  $A$  by prompting an LLM  $\mathcal{F}$  with a set of inputs  $\{t_y\}_{y=1}^K$  and instructs the model to generate a set of scene descriptions  $\mathcal{T}$ , corresponding to different groups  $\mathcal{G}$ . (ii) To preserve CLIP's generality and to avoid the need to fine-tune the network, we then introduce a novel method for

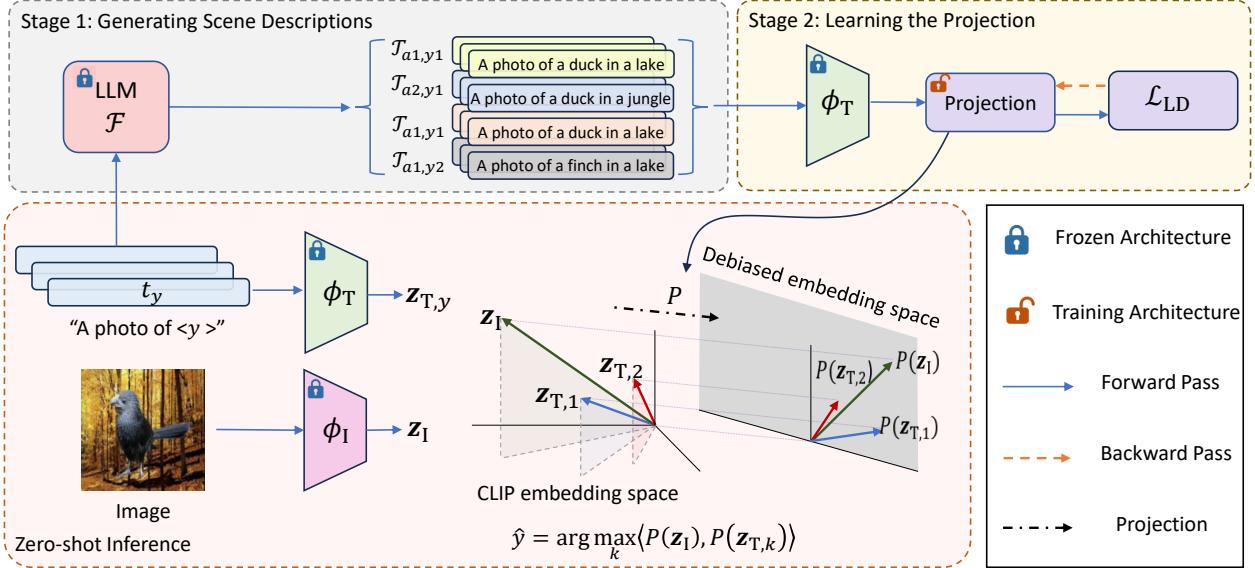


Figure 2. An overview of the proposed PRISM. In Stage 1, by passing text prompts as ‘A photo of <  $y$  >’ to an LLM, a set of scene descriptions with different labels and spurious attributes is generated. Then, in Stage 2, these scene descriptions are used for learning a debiasing projection of the embedding space,  $P$ , using LD loss. Finally, zero-shot classification is performed in the projected debiased embedding space of the VLM.

finding an embedding projector  $P$  that maps the embedding space onto another subspace where the influence of spurious attributes  $A$  is minimized. Figure 2 presents a high-level illustration of our proposed method.

**Stage 1: Leveraging LLMs to identify spurious correlations.** To propose a fully task-agnostic debiasing method without relying on any external data or predefined bias categories, we propose using pre-trained LLM for bias category discovery. For this purpose, an LLM is prompted with the labels and instructed to return potential spurious correlations. The LLMs’ effectiveness in the discovery of spurious correlations stems from their training scheme. These foundation models are trained over a massive corpus  $\mathcal{D}$  tokenized to  $w_i$  by minimizing the following:

$$\frac{1}{N} \sum_{(w_1, \dots, w_n) \in \mathcal{D}} \sum_{i=1}^n -\log P_\theta(w_i | w_{<i}), \quad (5)$$

thereby learning the joint distribution  $P_\theta(w_1, \dots, w_n)$ . Whenever a concept  $\alpha$  (e.g., ‘camel’) frequently co-occurs with another concept  $\beta$  (e.g., ‘desert’) in  $\mathcal{D}$ , the conditional probability  $P_\theta(\beta | \alpha)$  becomes large, even if  $\alpha$  and  $\beta$  are merely spuriously correlated rather than causally connected. Hence, for any set of text prompts  $\{t_y\}_{y=1}^K$ , the LLM’s learned parameters encode high-probability transitions from  $\{t_y\}_{y=1}^K$  to  $A$  solely due to their co-occurrences in text [24]. This statistical encoding of spurious correlations arises because the model optimizes the likelihood over  $\mathcal{D}$  without discriminating between causal and non-causal

patterns. Consequently, prompting the LLM with queries about  $\{t_y\}_{y=1}^K$  produces  $A$ , as follows:

$$A = \mathcal{F}(\{t_y\}_{y=1}^K). \quad (6)$$

Then, we use the returned spurious attributes to generate a set of controlled scene descriptions  $\mathcal{T}_{a,y}$  correspond to each group  $g = (a, y)$  in  $\mathcal{G}$ . The full prompts used in this stage of PRISM are provided in Appendix B.

**Stage 2: Learning a debiasing projection of the embedding space.** Our method finds a projection  $P$  to minimize the impact of  $A$  in the embedding space. This is then followed by zero-shot classification in the debiased projected embedding space. To this end, we propose a novel loss, which we call Latent space Debiasing Loss (LD), formulated as:

$$\begin{aligned} \mathcal{L}_{\text{LD}} = & \frac{1}{K \cdot \binom{|A|}{2}} \sum_{\substack{(a,y) \in \mathcal{G} \\ (a',y') \in \mathcal{G} \\ a \neq a' \\ y = y'}} \underbrace{\left( 1 - \langle \phi_T(\mathcal{T}_{a,y}), \phi_T(\mathcal{T}_{a',y'}) \rangle \right)}_{\text{(a) Intraclass, inter-attribute term}} + \\ & \frac{1}{|A| \cdot \binom{K}{2}} \sum_{\substack{(a,y) \in \mathcal{G} \\ (a',y') \in \mathcal{G} \\ a = a' \\ y \neq y'}} \underbrace{\max \left\{ 0, \langle \phi_T(\mathcal{T}_{a,y}), \phi_T(\mathcal{T}_{a',y'}) \rangle - m \right\}}_{\text{(b) Interclass, intra-attribute term}}, \end{aligned} \quad (7)$$

where  $m$  is the margin hyperparameter. We propose  $\mathcal{L}_{LD}$  to learn a projection that robustly mitigates spurious correlations in the model’s joint embedding space. For each iteration of optimization over  $P$ , we use a set of scene descriptions  $\mathcal{T}$ , generated by the LLM for each group,  $g = (a, y)$ . The first component of the loss ((a) Intra-class, inter-attribute term) encourages invariance by minimizing the difference between embeddings of scene descriptions from the same class but differing spurious attributes. Specifically, for each pair  $(a, a')$  with  $a \neq a'$ , the term  $1 - \langle \phi_T((a, y)), \phi_T((a', y')) \rangle$  is minimized. This term ensures that the core class-specific features are consistently represented, independent of the confounding spurious cues. Meanwhile, the second component of the loss ((b) Inter-class, intra-attribute term) employs a margin-based penalty that enforces a minimum separation between embeddings from the same spurious attribute but different classes. For any  $y' \neq y$ , the penalty  $\max\{0, \langle \phi_T((a, y)), \phi_T((a', y')) \rangle - m\}$  is applied, ensuring that even if the spurious context is shared, class-discriminative features remain distinct.

The design of  $\mathcal{L}_{LD}$  is inspired by contrastive loss in self-supervised learning, where the objective is to draw semantically similar pairs closer and push dissimilar ones apart. By jointly optimizing for intra-class invariance and inter-class separation, our loss effectively penalizes deviations from the desired embedding structure. Importantly, this approach removes the need to fine-tune the entire CLIP model, as the learnable projection  $P$  is optimized independently. Empirical evaluations confirm that minimizing  $\mathcal{L}_{LD}$  not only enhances the robustness of the zero-shot classifier by reducing reliance on spurious cues but also improves worst-group performance and leads to disentanglement of the embedding space.

Finally, leveraging the learned projection,  $P$ , by minimizing Eq. 7, we perform zero-shot classification on the projected embedding space as follows:

$$\hat{y} = \arg \max_k \langle P(\phi_I(x)), P(\phi_T(t_k)) \rangle. \quad (8)$$

**PRISM-mini.** Here, we propose a variation of our method called PRISM-mini, which is mostly different from PRISM in stage 2. Unlike PRISM, which uses scene descriptions for finding  $P$  by optimizing over LD loss, PRISM-mini relies on orthogonalizing the embedding space again spurious attributes. Hence, PRISM-mini is less computationally complex and is more suitable for use-cases where computation resources are extremely limited.

In the first stage, we pass the class prompts to the LLM and obtain a set of spurious attributes rather than full scene descriptions. Then, all these spurious attributes are encoded using CLIP’s text encoder as follows:

$$\mathcal{A} = \phi_T(\mathcal{F}(\{t_k\}_{k=1}^K)), \quad (9)$$

---

### Algorithm 1 PRISM

---

**Require:**

image  $x$   
 $\{t_y\}_{y=1}^K$ : Text prompts for  $K$  classes  
 $\phi_I, \phi_T$ : CLIP’s image and text encoders  
 $\mathcal{F}$ : LLM

- 1: **Stage 1: Identify Spurious Directions**
  - 2:  $A = \mathcal{F}(\{t_y\}_{y=1}^K)$   $\triangleright$  Extract spurious attributes
  - 3: **Stage 2 (PRISM): Learned Projection**
  - 4: Generate scene descriptions  $\mathcal{T}_{a,y}$  via  $\mathcal{F}$  and  $A$
  - 5: Initialize  $P$
  - 6: **repeat**
  - 7:     Compute  $\mathcal{L}_{LD}$  (Eq. 7) on  $\{(a, y)\}$
  - 8:     Update  $P$  via gradient descent
  - 9: **until** converged
  - 10: **Stage 2 (PRISM-mini): Simple Orthogonal Projection**
  - 11:  $\mathcal{A} = \phi_T(A)$
  - 12:  $P \leftarrow I - \mathcal{A}(\mathcal{A}^\top \mathcal{A})^{-1} \mathcal{A}^\top$
  - 13: **Stage 3: Zero-shot Inference**
  - 14:  $\mathbf{z}_I \leftarrow P(\phi_I(x)), \mathbf{z}_{T,k} \leftarrow P(\phi_T(t_k))$
  - 15:  $\hat{y} \leftarrow \arg \max_k \langle \mathbf{z}_I, \mathbf{z}_{T,k} \rangle$
- 

where  $\mathcal{A}$  is a matrix whose columns are the embeddings of bias attributes in  $A$ . Inspired by [5], here we seek to reduce the impact of these spurious attributes on the embedding space by orthogonalizing it against  $\mathcal{A}$ . To this end, we calculate the projection  $P$  as follows:

$$P = I - \mathcal{A}(\mathcal{A}^\top \mathcal{A})^{-1} \mathcal{A}^\top. \quad (10)$$

The rest of the method is essentially the same as PRISM. PRISM-mini can effectively reduce the impact of spurious correlations in CLIP without the need for any optimization for finding  $P$ . As a result, this method is very cost-effective. However, compared to PRISM, it might not be capable of fully mitigating the spurious biases in the embedding space due to its simpler projection mechanism. Algorithm 1 summarizes both variations of the proposed method.

## 4. Experiment setup

**Datasets.** To evaluate the effectiveness of our method, we conduct various experiments on Waterbirds [23] and CelebA [18] datasets, following prior works such as [7]. The Waterbirds dataset consists of images where bird species are paired with specific backgrounds, deliberately inducing spurious correlations between the foreground object and the environmental context. The CelebA dataset includes 200,000 facial images of celebrities annotated with 40 attributes, such as gender, hair color, face shape, etc.

**Evaluation.** For performance evaluation, we use Five metrics: Accuracy (Acc), Worst-Group accuracy (WG), which is the lowest accuracy of all subgroups,  $\Delta WG$ , which indicates the worst group accuracy improvement compared to

Model	Waterbirds					CelebA				
	WG $\uparrow$	Acc $\uparrow$	$\Delta$ WG $\uparrow$	$\Delta$ Acc $\uparrow$	Gap $\downarrow$	WG $\uparrow$	Acc $\uparrow$	$\Delta$ WG $\uparrow$	$\Delta$ Acc $\uparrow$	Gap $\downarrow$
<b>Baseline</b>										
Zero-shot	36.4%	89.3%	—	—	52.9%	72.8%	87.6%	—	—	14.8%
<b>Methods using images for debiasing</b>										
WiSE-FT [29]	65.9%	97.6%	29.5%	8.3%	31.7%	80.0%	87.4%	7.2%	-0.2%	7.4%
DFR (Sub) [17]	51.9%	95.7%	15.5%	6.4%	43.8%	76.3%	92.1%	3.5%	4.5%	15.8%
DFR (Up) [17]	65.9%	96.1%	29.5%	6.8%	30.2%	83.7%	91.2%	10.9%	3.6%	7.5%
FairerCLIP [7]	78.1%	85.1%	41.7%	-4.2%	<b>7.0%</b>	86.1%	88.0%	13.3%	0.4%	<u>1.9%</u>
<b>Data-free methods</b>										
VisualDistiller [6]	42.7%	90.6%	6.3%	1.3%	47.9%	—	—	—	—	—
Orth-Proj [5]	45.3%	86.4%	8.9%	-2.9%	41.1%	71.1%	<b>87.0%</b>	-1.7%	<b>-0.6%</b>	15.9%
Orth-Cali [5]	68.8%	84.5%	32.4%	-4.3%	15.7%	76.1%	86.2%	3.3%	-1.4%	10.1%
RoboShot [1]	45.2%	79.2%	8.8%	-10.1%	34.0%	<u>82.6%</u>	85.5%	<u>9.8%</u>	-2.1%	2.9%
PRISM-mini (ours)	<u>69.5%</u>	<u>92.6%</u>	<u>33.1%</u>	<u>3.3%</u>	23.1%	<u>82.6%</u>	84.4%	<u>9.8%</u>	-3.2%	<b>1.8%</b>
PRISM (ours)	<b>84.2%</b>	<b>93.6%</b>	<b>47.8%</b>	<b>4.3%</b>	<u>9.4%</u>	<b>84.0%</b>	<u>86.9%</u>	<b>11.2%</b>	<u>-0.7%</u>	2.9%

Table 1. The performance of PRISM compared to baselines on CLIP-ViT-L/14 on Waterbirds and CelebA datasets. The top performance in each column is in **bold**, and the second-best is underlined.

the zero-shot clip,  $\Delta$ Acc, which indicates the overall accuracy improvement compared to the zero-shot performance of CLIP, and the Gap between Acc and WG.

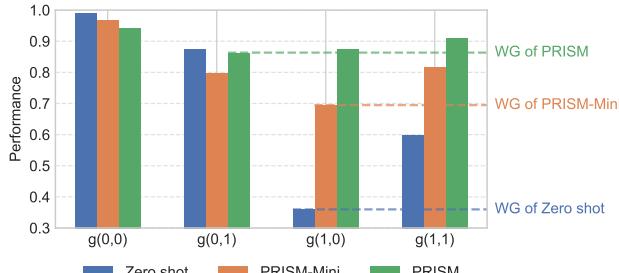
**Baselines.** We compare our method against two categories of baselines: the methods that use auxiliary images for debiasing and data-free methods. The first category employs a set of labeled or unlabeled images in the training process to reduce bias in CLIP. These images are used to keep the alignment of text and image embeddings intact throughout the debiasing process. Three state-of-the-art methods from this category are compared to ours, including WiSE-FT [29], DFR [17], and FairerCLIP [7]. The second category consists of data-free models which do not require any images for debiasing. These methods are inherently task-agnostic as they do not rely on any task-specific visual data to achieve effective debiasing. VisualDistiller [6], Orth [5], and RoboShot [1] are used in the following experiments as state-of-the-art methods in this category. Note that both variations of our method fall into this category.

**Implementation details.** All experiments were conducted using PyTorch on an NVIDIA GeForce RTX 3090 GPU with 24 GB VRAM and 64 GB RAM. Each experiment was run three times with different initialization seeds, and the average performance of the models was reported. Adam optimizer [16] with the learning rate of 0.1 (Waterbirds) and 0.01 (CelebA) is used for performing optimization over  $P$ . The batch size is set to 64, and the training for the projection step is only performed for a single epoch.

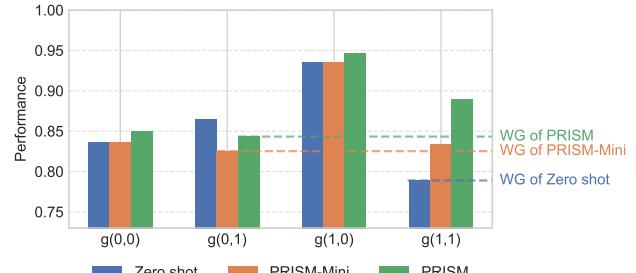
## 5. Results

**Performance.** We evaluate the performance of our method and compare it against prior baselines on two datasets (Waterbirds and CelebA) using the CLIP-ViT-L/14 backbone in Table 1. The results demonstrate that PRISM outperforms prior works, setting a new state-of-the-art over data-free debiasing methods. Specifically, we observe that our proposed PRISM achieves the highest WG in both datasets, meaning it effectively reduces bias without needing additional image data. The effectiveness of our method is further demonstrated as PRISM outperforms the models using auxiliary data in the Waterbirds benchmark and achieve competitive results in the CelebA benchmark. FairerCLIP is the best among image-based methods, improving WG at the cost of some accuracy loss. WiSE-FT and DFR achieve high accuracies but have limited bias mitigation, indicating that fine-tuning might not be a good option to mitigate bias. Orth and RoboShot perform moderately well but struggle in WG gains, demonstrating that although prompt tuning is a viable method for guiding CLIP’s outputs, it is not a robust solution for bias removal. Moreover, the efficient variation of our method, PRISM-mini, also achieves the second-best performance in debiasing across all data-free methods while avoiding any optimization, making it a strong candidate for scenarios where resources are very limited.

**Per-group accuracy.** In Figure 3, each subfigure shows bar plots of per-group accuracy for three different methods, i.e., Zero-shot, PRISM-mini, and PRISM, on Waterbirds (Figure 3a) and CelebA (Figure 3b) datasets. The groups on the  $x$ -axis ( $g(0, 0)$ ,  $g(0, 1)$ ,  $g(1, 0)$ ,  $g(1, 1)$ ) represent different



(a) Waterbirds.



(b) CelebA.

Figure 3. Per-group accuracy of our method on (a) Waterbirds and (b) CelebA datasets.

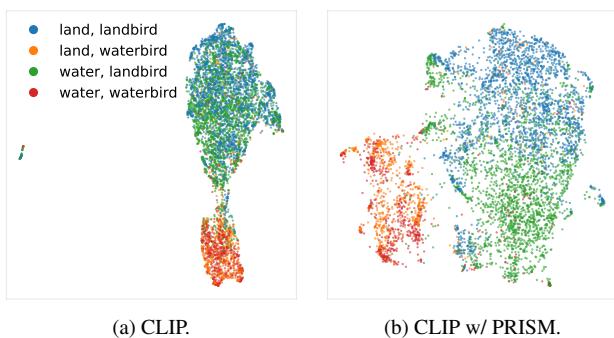


Figure 4. The representation map at the output of CLIP (a) without and (b) with PRISM .

subpopulations within each dataset, and the y-axis shows the accuracy achieved by each method on those subpopulations. The dashed horizontal lines indicate the WG accuracy achieved by each method, i.e., the lowest accuracy among the four subgroups. By looking at where each bar stands relative to the dashed lines, we can see how each method handles the most challenging subgroup. Across both datasets, PRISM improves accuracy over the zero-shot baseline in each subgroup and tends to raise the WG performance. PRISM-mini also provides gains over the zero-shot baseline, but it may not always reach the same level of WG performance as PRISM. In essence, the figure illustrates that both PRISM variants can help reduce performance disparities across different groups compared to a purely zero-shot approach, with PRISM typically offering the strongest and the most balanced improvements.

**Bias text.** We previously hypothesized that since CLIP aligns image and text embeddings, the inherent nature of spurious bias should cause similar biases to emerge in the text domain alone [24]. Consequently, spurious correlations extracted from the text domain can also be applicable to images. To empirically evaluate this hypothesis, we generate a set of scene descriptions featuring different bird species in various locations. We then perform zero-shot classifica-

Zero shot	WG ↑	Acc ↑	$g(0,0)$	$g(0,1)$	$g(1,0)$	$g(1,1)$
Image	38.3%	92.8%	99.0%	87.6%	38.3%	59.8%
Scene description	33.1%	96.0%	100%	33.1%	41.9%	96.8%

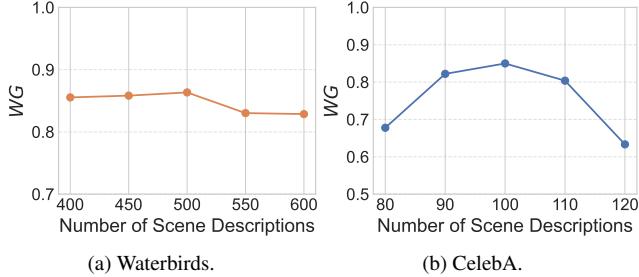
Table 2. Bias similarity in text and image modalities.

tion on the embeddings of these descriptions and present the results in Table 2, demonstrating that similar biases can indeed be observed in the text domain.

**Trade-off analysis.** Figures 1a and 1b compare different data-free CLIP debiasing methods in terms of accuracy and debiasing performance, measured through WG, on Waterbirds and CelebA datasets. Although some methods exhibit high debiasing performances, their overall performance is drastically degraded. PRISM is the most effective among all compared methods, achieving the best trade-off between these two metrics.

**Feature disentanglement.** Figure 4 compares the representation distributions of zero-shot CLIP with PRISM on the Waterbirds dataset, where each color represents a group  $(a, y) \in \mathcal{G}$ . In CLIP’s representation space (Figure 4a), clusters are highly mixed, indicating that the model heavily relies on spurious background correlations. In contrast, PRISM’s representation space (Figure 4b) shows more structured and well-disentangled clusters, demonstrating that it effectively reduces the influence of background information. This analysis confirms that PRISM successfully mitigates bias rather than being influenced by spurious background correlations.

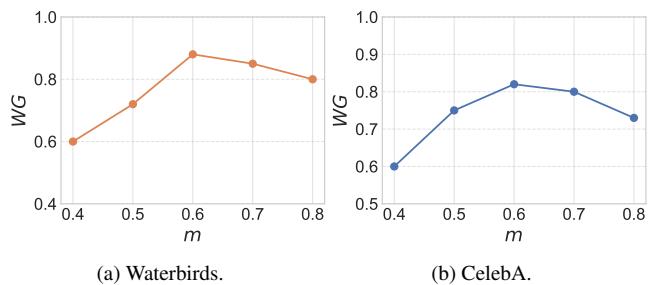
**Number of scene descriptions.** Figure 5 shows the effect of varying the number of scene descriptions used during training the projection  $P$ . In these experiments, we evaluate how increasing or decreasing the volume of scene descriptions, which serve as bias-aware prompts generated by the LLM, affects the model’s ability to mitigate spurious correlations. For both the Waterbirds and CelebA datasets, the results illustrate that there is an optimal range for the number of descriptions: too few may provide insufficient guid-



(a) Waterbirds.

(b) CelebA.

Figure 5. Worst-group performance of PRISM for different numbers of scene descriptions used in training for (a) Waterbirds and (b) CelebA datasets.



(a) Waterbirds.

(b) CelebA.

Figure 6. Worst-group performance of PRISM for different values of margin hyperparameter  $m$  for (a) Waterbirds and (b) CelebA datasets.

ance for debiasing, while too many might undermine the discriminative signal, ultimately impacting the worst-group accuracy.

**m in LD loss.** Figure 6 shows the sensitivity of PRISM’s performance to the hyperparameter  $m$  in the LD loss function, which determines the margin in the contrastive component of the loss. This parameter influences how aggressively the model penalizes similarities between embeddings of different classes with the same spurious attribute. The results, demonstrate that for both datasets, an  $m$  of 0.6 maximizes WG performance by separating the representations across different groups, achieving an optimal balance between preserving essential semantic information and removing spurious correlations.

**Role of LLMs.** Table 3 shows that the effectiveness of PRISM in reducing bias, measured by WG and Acc, can be significantly influenced by the choice of LLM. Selecting a more capable LLM for bias identification can improve debiasing without severely compromising the overall classification performance. The table compares several LLMs that vary in size and capability, where we see that GPT-4o achieves the best WG performance with score of 84.0% and an overall accuracy of 86.9%.

**Limitations.** While PRISM demonstrates strong performance in mitigating spurious biases in VLMs, it has some

LLM	WG ↑	Acc ↑
Llama 3.2 (70B) [8]	58.9%	90.1%
DeepSeek r1 (8B) [12]	68.3%	89.2%
GPT 4o mini (~8B) [13]	81.6%	86.6%
GPT o1 (~300B) [14]	77.9%	83.9%
GPT 4o (~1.7T) [13]	84.0%	86.9%

Table 3. The performance of PRISM on CelebA dataset using different LLMs.

limitations. First, PRISM relies on LLMs to generate spurious attribute descriptions, and the quality of these descriptions can vary depending on the specific LLM used. While our results show that more capable LLMs yield better debiasing performance, less powerful models may struggle to extract meaningful spurious correlations, potentially reducing effectiveness. Additionally, our approach focuses on linear projections to remove bias in the embedding space, which may not always capture highly complex or non-linear biases present in real-world data.

## 6. Conclusion

We propose PRISM, a data-free debiasing framework for vision-language models that mitigates spurious correlations in CLIP’s embedding space without requiring any weight updates. PRISM has two stages. In the first stage, an LLM is leveraged with simple class prompts to generate scene descriptions that unravel common spurious correlations in the training data. In the second stage, these generated descriptions guide the learning of an optimized projection mapping, which is carried out via our novel LD loss to systematically reduce spurious biases while preserving the semantic alignment between image and text embeddings. Through extensive evaluations on benchmark datasets (Waterbirds and CelebA), we demonstrated that PRISM significantly improves worst-group accuracy while maintaining overall classification performance. Beyond its empirical effectiveness, PRISM offers a data-free and task-agnostic solution to debiasing multimodal representations. Future research directions include exploring non-linear projections for embedding space debiasing, which could potentially better capture complex spurious correlations beyond linear transformations.

## Acknowledgment

We would like to thank Geotab Inc., the City of Kingston, and NSERC for their invaluable and continued support of this work.

## References

- [1] Dyah Adila, Changho Shin, Linrong Cai, and Frederic Sala. Zero-shot robustification of zero-shot models. In *International Conference on Learning Representations*, 2024. 1, 2, 3, 6
- [2] Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. Evaluating clip: towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*, 2021. 2
- [3] Ibrahim Alabdulmohsin, Xiao Wang, Andreas Peter Steiner, Priya Goyal, Alexander D'Amour, and Xiaohua Zhai. CLIP the bias: How useful is balancing data in multimodal learning? In *International Conference on Learning Representations*, 2024. 2
- [4] Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In *International conference on machine learning*, pages 528–539. PMLR, 2020. 2
- [5] Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*, 2023. 1, 2, 5, 6
- [6] Haocheng Dai and Sarang Joshi. Refining skewed perceptions in vision-language models through visual representations. *arXiv preprint arXiv:2405.14030*, 2024. 1, 2, 6
- [7] Sepehr Dehdashian, Lan Wang, and Vishnu Naresh Boddeti. Fairerclip: Debiasing clip's zero-shot predictions using functions in rkhss. *arXiv preprint arXiv:2403.15593*, 2024. 1, 2, 5, 6
- [8] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 8
- [9] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, pages 1–79, 2024. 2
- [10] Walter Gerych, Haoran Zhang, Kimia Hamidieh, Eileen Pan, Maanas K Sharma, Tom Hartvigsen, and Marzyeh Ghassemi. Bendylm: Test-time debiasing of vision-language embeddings. *NeurIPS*, 37:62480–62502, 2024. 3
- [11] Priyanshu Govil, Hemang Jain, Vamshi Krishna Bonagiri, Aman Chadha, Ponnurangam Kumaraguru, Manas Gaur, and Sanorita Dey. Cobias: Contextual reliability in bias assessment. *arXiv preprint arXiv:2402.14889*, 2024. 2
- [12] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 8
- [13] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 8
- [14] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024. 8
- [15] Hoin Jung, Taeuk Jang, and Xiaoqian Wang. A unified debiasing approach for vision-language models across modalities and tasks. *Advances in Neural Information Processing Systems*, 37:21034–21058, 2025. 2
- [16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6
- [17] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *International Conference on Learning Representations*, 2023. 1, 2, 3, 6
- [18] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 5
- [19] Nicole Meister, Dora Zhao, Angelina Wang, Vikram V Ramaswamy, Ruth Fong, and Olga Russakovsky. Gender artifacts in visual datasets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4837–4848, 2023. 1, 2
- [20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 3
- [21] Shaina Raza, Mizanur Rahman, and Michael R Zhang. Beads: Bias evaluation across domains. *arXiv preprint arXiv:2406.04220*, 2024. 2
- [22] Christian Reimers, Paul Bodesheim, Jakob Runge, and Joachim Denzler. Towards learning an unbiased classifier from biased data via conditional adversarial debiasing. *arXiv preprint arXiv:2103.06179*, 2021. 2
- [23] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *International Conference on Learning Representations*, 2020. 5
- [24] Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Ö Arić, and Tomas Pfister. Data-augmented phrase-level alignment for mitigating object hallucination. *International Conference on Learning Representations*, 2025. 4, 7
- [25] Jie-Jing Shao, Jiang-Xin Shi, Xiao-Wen Yang, Lan-Zhe Guo, and Yu-Feng Li. Investigating the limitation of clip models: The worst-performing categories. *CoRR*, 2023. 2
- [26] Angelina Wang, Alexander Liu, Ryan Zhang, Anat Kleiman, Leslie Kim, Dora Zhao, Iroha Shirai, Arvind Narayanan, and Olga Russakovsky. Revise: A tool for measuring and mitigating bias in visual datasets. *International Journal of Computer Vision*, 130(7):1790–1810, 2022. 1
- [27] Jialu Wang, Yang Liu, and Xin Wang. Are gender-neutral queries really gender-neutral? mitigating gender bias in im-

- age search. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021. 1
- [28] Junyang Wang, Yi Zhang, and Jitao Sang. Fairclip: Social bias elimination based on attribute prototype learning and representation neutralization. *arXiv preprint arXiv:2210.14562*, 2022. 2
- [29] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7959–7971, 2022. 1, 2, 6
- [30] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 547–558, 2020. 1
- [31] Boya Zeng, Yida Yin, and Zhuang Liu. Understanding bias in large-scale visual datasets. *arXiv preprint arXiv:2412.01876*, 2024. 1, 2
- [32] Michael Zhang and Christopher Ré. Contrastive adapters for foundation model group robustness. *Advances in Neural Information Processing Systems*, 35:21682–21697, 2022. 3
- [33] Kankan Zhou, Yibin LAI, and Jing Jiang. Vlseteroset: A study of stereotypical bias in pre-trained vision-language models. In *Proceedings of the Association for Computational Linguistics: International Joint Conference on Natural Language Processing*, 2022. 1