

# SAGE: Spuriousness-Aware Guided Prompt Exploration for Mitigating Multimodal Bias

Wenqian Ye<sup>1</sup>, Di Wang<sup>1</sup>, Guangtao Zheng<sup>2</sup>, Bohan Liu<sup>1</sup>, Aidong Zhang<sup>1\*</sup>

<sup>1</sup>Department of Computer Science, University of Virginia, USA

<sup>2</sup>Accenture, USA

{wenqian, azm7tq, qzp4ta, aidong}@virginia.edu, zhguangt@gmail.com

## Abstract

Large vision-language models, such as CLIP, have shown strong zero-shot classification performance by aligning images and text in a shared embedding space. However, CLIP models often develop multimodal spurious biases, which is the undesirable tendency to rely on spurious features. For example, CLIP may infer object types in images based on frequently co-occurring backgrounds rather than the object’s core features. This bias significantly impairs the robustness of pre-trained CLIP models on out-of-distribution data, where such cross-modal associations no longer hold. Existing methods for mitigating multimodal spurious bias typically require fine-tuning on downstream data or prior knowledge of the bias, which undermines the out-of-the-box usability of CLIP. In this paper, we first theoretically analyze the impact of multimodal spurious bias in zero-shot classification. Based on this insight, we propose Spuriousness-Aware Guided Exploration (SAGE), a simple and effective method that mitigates spurious bias through guided prompt selection. SAGE requires no training, fine-tuning, or external annotations. It explores a space of prompt templates and selects the prompts that induce the largest semantic separation between classes, thereby improving worst-group robustness. Extensive experiments on four real-world benchmark datasets and five popular backbone models demonstrate that SAGE consistently improves zero-shot performance and generalization, outperforming previous zero-shot approaches without any external knowledge or model updates.

**Code** — [https://github.com/wenqian-ye/spurious\\_vlm](https://github.com/wenqian-ye/spurious_vlm)

## 1 Introduction

Pre-trained models hold promising potential for open-set classification without the need for additional data collection or training. Pre-trained vision-language models (VLMs) (Radford et al. 2021; Jia et al. 2021; Li et al. 2021; Wang et al. 2022; Li et al. 2022), such as contrastive language-image pre-training (CLIP) models (Radford et al. 2021), have demonstrated a strong zero-shot prediction capability across diverse downstream tasks. They typically consist of a pre-trained image encoder and a text encoder, from which vision and text representations are aligned in a shared joint

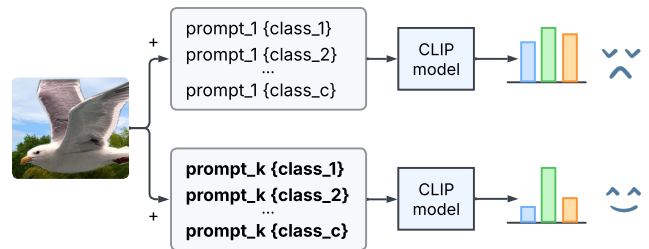


Figure 1: Prompts with greater separation between class similarity scores (e.g., prompt<sub>k</sub>) yield robust zero-shot performance under spurious correlations, whereas those with smaller score differences (e.g., prompt<sub>1</sub>) tend to yield poorer discrimination and worst-group performance.

embedding space. Thus, zero-shot classification of an image can be achieved simply by matching the image representation to a set of candidate text representations.

However, recent studies (You et al. 2024; Adila et al. 2024; Dehdashtian, Wang, and Boddeti 2024) have found that pre-trained CLIP models often develop undesirable tendencies to rely on spurious correlations between non-essential features and target labels across modalities when making predictions. For example, the class label `landbird` may become spuriously associated with `land background` due to frequent co-occurrence in the pre-training data (Zheng, Ye, and Zhang 2024b,a). As a result, a CLIP model may incorrectly predict a `waterbird` as a `landbird` simply because it appears in a `land background`. This kind of biased prediction behavior, referred to as multimodal spurious bias, severely impairs the zero-shot generalization ability of CLIP models on out-of-distribution data where such spurious correlations no longer hold (Ye et al. 2024). For instance, the correlation between `landbird` and `land background` may not exist in the downstream out-of-distribution evaluation setting.

Mitigating multimodal spurious bias is essential for ensuring robust generalization across downstream tasks. Existing methods vary significantly in their approaches. Some (Yang et al. 2023; You et al. 2024; Zhang et al. 2024; Dehdashtian, Wang, and Boddeti 2024) adopt fine-tuning strategies, focusing on task-specific biases and requiring additional data. Although these methods improve robustness to multimodal

\*Corresponding Author.

spurious bias over the vanilla zero-shot approach, they rely on labeled data and do not address the zero-shot setting. RO-BOSHOT (Adila et al. 2024) mitigates spurious bias within the language modality without training data, but typically requires specifying spurious attributes by prompting a large language model (LLM) for each downstream task. TIE\* (Lu, Chai, and Wang 2025), though not relying on LLMs, directly uses spurious attributes to obtain pseudo spurious labels for multimodal bias mitigation.

To mitigate multimodal spurious bias without relying on prior knowledge or external models, we propose a training-free framework, namely **Spuriousness-Aware Guided Exploration (SAGE)**. We begin by formally defining multimodal spurious bias and analyzing its impact on zero-shot classification. Our insight is that prompts with larger differences in inter-class similarity tend to better capture core class semantics, which helps reduce reliance on spurious correlations. As shown in Figure 1, the top part illustrates prompt<sub>1</sub>, where the green bar is only slightly higher than the blue bar, reflecting weak class separation and lower zero-shot performance. The bottom part shows prompt<sub>k</sub>, where the green bar is higher than the lowest blue bar, indicating stronger class discrimination and improved predictive accuracy. Our theoretical and empirical results suggest that higher *separation scores* are associated with greater focus on essential class features rather than spurious ones, thereby improving zero-shot robustness.

Based on this insight, SAGE utilizes a set of diverse candidate prompt templates commonly used with CLIP models or their variants. For each image, SAGE calculates the similarity scores between the image and the class labels under different prompt templates. The prompt template with top greatest difference between the highest and lowest class scores is selected for zero-shot inference. SAGE works entirely without fine-tuning or external supervision and can be applied to any zero-shot vision-language model. Extensive experiments on four benchmark datasets and five backbone models demonstrate that SAGE consistently enhances zero-shot accuracy while effectively mitigating multimodal spurious bias.

## 2 Related Work

**Spurious bias in single data modality.** Spurious bias refers to the reliance of models on spurious correlations between input features and targets, leading to poor generalization on out-of-distribution data (Beery, Van Horn, and Perona 2018; Geirhos et al. 2020; Zheng, Ye, and Zhang 2025b; Ye et al. 2025). Existing methods typically mitigate spurious bias by retraining models with labeled data, where spurious correlations are either explicitly annotated via group labels (Sagawa et al. 2019; Kirichenko, Izmailov, and Wilson 2023; Deng et al. 2024), implicitly identified through group inference (Nam et al. 2022; Ye, Zheng, and Zhang 2025; Zheng, Ye, and Zhang 2025a), or sample reweighting (Nam et al. 2020; Liu et al. 2021; Qiu et al. 2023; LaBonte, Muthukumar, and Kumar 2024). Our work addresses the relatively under-explored challenge of mitigating spurious bias in a zero-shot multimodal setting where no retraining data is available.

**Debiasing fine-tuned VLMs.** Multimodal spurious bias refers to the tendency of models to rely on spurious correlations in one modality (e.g., image background) to infer targets in another (e.g., object names). In VLMs, such bias may arise from misalignment between modalities during pre-training or fine-tuning (Tong et al. 2024; Sun et al. 2024). Existing approaches mitigate this bias in fine-tuned VLMs using contrastive learning with or without group labels (Yang et al. 2023; Zhang and Ré 2022; You et al. 2024), or by disentangling spurious and core features via prompt tuning (Zhang et al. 2024) or latent projection (Dehdashtian, Wang, and Boddeti 2024). In contrast, our method targets the zero-shot setting without any downstream data, which is applicable on a broader real-world scenarios.

**Zero-shot debiasing.** Debiasing in the zero-shot setting aims to mitigate multimodal spurious bias learned during pre-training, where target texts may align with spurious image features (Ge et al. 2023). Existing methods typically leverage text data: Chuang et al. (2023) use prompts with known spurious attributes to adjust CLIP classifier weights, while Adila et al. (2024) extract core and spurious attributes via LLMs to enhance image features. Moreover, Lu, Chai, and Wang (2025) rely on explicit spurious attribute information to generate pseudo-labels that help reduce bias in multimodal embeddings. In contrast, our method mitigates bias by selecting prompts based on a *separation score* computed from similarity differences between class labels, without relying on LLMs or prior knowledge.

## 3 Methodology

We first theoretically analyze the multimodal spurious bias in VLMs. Based on the insights gained in the analysis, we introduce spuriousness-aware guided prompt exploration (SAGE) that selects prompts based on a *separation score* to mitigate such bias in a zero-shot setting.

### 3.1 Preliminary

A CLIP (Radford et al. 2021) model is trained to align the representation of an image  $x$  from its vision encoder  $\phi$  and the representation of a text description  $t$  from its text encoder  $\psi$  in a joint embedding space when the text description  $t$  matches with the image  $x$ . Specifically, let  $\mathbf{v} = \phi(x) \in \mathbb{R}^D$  denote the vision representation for the image  $x$  and  $\mathbf{u} = \psi(t) \in \mathbb{R}^D$  be the text representation for the text description  $t$ , where  $D$  is the number of embedding dimensions. Then, the CLIP training objective (Radford et al. 2021) essentially aims to maximize the probability of  $\mathbf{v}$  given  $\mathbf{u}$  and the probability of  $\mathbf{u}$  given  $\mathbf{v}$  over all training image-text pairs, i.e.,

$$\phi, \psi = \arg \max_{\phi', \psi'} \mathbb{E}_{p(x,t)} \left( p(\mathbf{v}|\mathbf{u}) + p(\mathbf{u}|\mathbf{v}) \right), \quad (1)$$

where  $p(x, t)$  denotes the joint distribution of matching image-text pairs in the training set. During training, CLIP models learn to align the embeddings of matching image-text pairs—for example, bringing together the representation of a landbird image and the phrase “a photo of a landbird”. At the same time, it pushes apart the representations of mismatched pairs, such as a waterbird image with the same

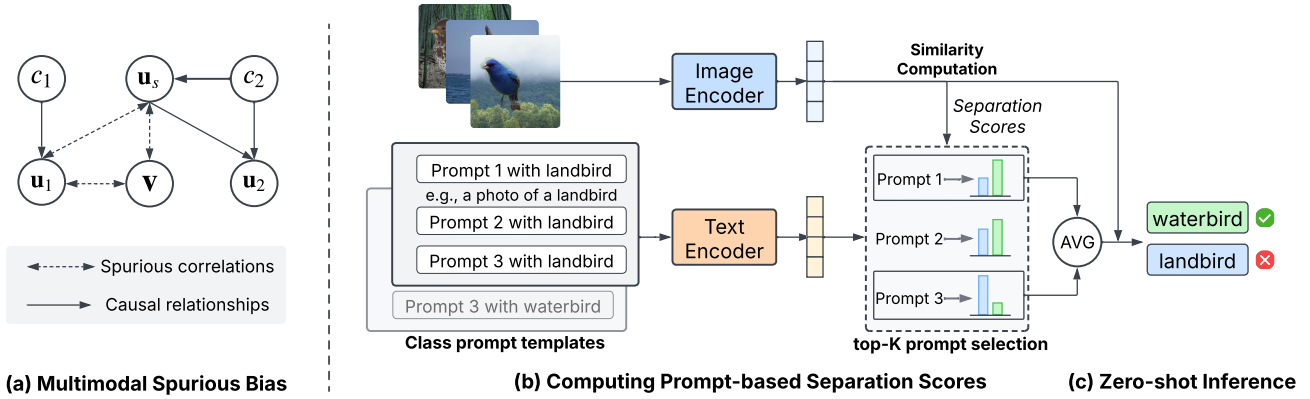


Figure 2: Method overview. (a) Illustration of multimodal spurious bias, where  $c_2$  denotes a class label,  $\mathbf{v}$  denotes an image representation,  $\mathbf{u}_s$  denotes a textual spurious feature,  $\mathbf{u}_1$  and  $\mathbf{u}_2$  denote text representations for the class  $c_1$  and  $c_2$  respectively. (b) For each test image, we evaluate  $M$  prompt templates and compute a *separation score* that measures how well each prompt distinguishes between classes in the joint image-text space. The top- $K$  templates with the highest scores are selected. (c) Zero-shot classification is then performed by ensembling predictions from the  $K$  class-discriminative prompts selected for that image.

phrase. Ideally, for a matching image-text pair  $(x, t)$ , we will obtain  $p(\mathbf{v}|\mathbf{u}) \approx p(\mathbf{u}|\mathbf{v})$  after training.

**Zero-shot classification.** Given an image  $x$  belonging to one of  $C$  classes  $\{c_i\}_{i=1}^C$ , zero-shot classification first constructs  $C$  text descriptions by inserting each class name  $c_i$  into a predefined text template, such as “a photo of a [ $c_i$ ]” (e.g., “a photo of a landbird”). Each description is then encoded into a text representation  $\mathbf{u}_i$  for each class  $c_i$ . Then, the zero-shot prediction  $\hat{y}$  is:

$$\hat{y} = \arg \max_i p(\mathbf{u}_i|\mathbf{v}) = \arg \max_i \frac{\mathbf{v}^T \mathbf{u}_i}{\|\mathbf{v}\|_2 \|\mathbf{u}_i\|_2}, \quad (2)$$

where  $\mathbf{v}$  is the vision representation for the input image  $x$ ,  $\|\cdot\|_2$  is the Euclidean norm of a vector, and  $p(\mathbf{u}_i|\mathbf{v})$  is defined to be proportional to  $\mathbf{v}^T \mathbf{u}_i$ .

### 3.2 Multimodal Spurious Bias

In practice, a given text description  $t$  may not fully describe the content in  $x$ . For example,  $x$  could be an image depicting a landbird with a land background, and  $t$  could simply be “a photo of a landbird”, which only describes the primary object in the image. When a CLIP model learns to align many such image-text pairs where land backgrounds spuriously correlate with the target “landbird”, then the model may inadvertently learn to align the representation of “a photo of a landbird” with the representation of land backgrounds, instead of the defining features of landbirds. The misalignment causes a *multimodal spurious bias* in the model which tends to use land backgrounds in images to infer their descriptions. Due to the misalignment, an image of waterbird with a land background is incorrectly paired with the description “a photo of a landbird”.

To formally define multimodal spurious bias, we introduce  $\mathbf{u}_s \in \mathbb{R}^D$  to represent a latent textual spurious feature, such as the missing “land background” in the description “a photo of a landbird”. With  $\mathbf{u}_s$ , we can conveniently expand

$p(\mathbf{v}|\mathbf{u})$  and  $p(\mathbf{u}|\mathbf{v})$  in (1) as the marginalization over all possible textual spurious features, i.e.,

$$p(\mathbf{v}|\mathbf{u}) = \int_{\mathbf{u}_s} p(\mathbf{v}|\mathbf{u}, \mathbf{u}_s) p(\mathbf{u}_s|\mathbf{u}) d\mathbf{u}_s, \quad (3)$$

and

$$p(\mathbf{u}|\mathbf{v}) = \int_{\mathbf{u}_s} p(\mathbf{u}|\mathbf{v}, \mathbf{u}_s) p(\mathbf{u}_s|\mathbf{v}) d\mathbf{u}_s. \quad (4)$$

In the pre-training data, if the majority of images with their text representation  $\mathbf{u}$  have a spurious feature represented by  $\mathbf{u}_s$ , then a CLIP model may learn the strong correlations between the spurious feature  $\mathbf{u}_s$  and the image representation  $\mathbf{v}$  as well as the text representation  $\mathbf{u}$ . As a result, the model will develop multimodal spurious bias and we will have  $p(\mathbf{u}_s|\mathbf{u}) \approx 1$  and  $p(\mathbf{u}_s|\mathbf{v}) \approx 1$ . We formally define multimodal spurious bias as follows.

**Definition 1 (Multimodal spurious bias).** Consider a pre-trained CLIP model consisting of a vision encoder  $\phi$  and a text encoder  $\psi$ . Given an image-text pair  $(x, t)$  and a latent spurious feature  $\mathbf{u}_s$ , a multimodal spurious bias in the model relevant to  $\mathbf{u}_s$  satisfies the following conditions:

$$p(\mathbf{v}|\mathbf{u}) \approx p(\mathbf{v}|\mathbf{u}, \mathbf{u}_s), \quad (5)$$

and

$$p(\mathbf{u}|\mathbf{v}) \approx p(\mathbf{u}|\mathbf{v}, \mathbf{u}_s), \quad (6)$$

where  $\mathbf{v} = \phi(x)$  and  $\mathbf{u} = \psi(t)$ .

The above conditions indicate that  $p(\mathbf{u}_s|\mathbf{u}) \approx 1$  and  $p(\mathbf{u}_s|\mathbf{v}) \approx 1$  are based on Eq. (3) and Eq. (4), and the pre-trained model tends to align  $\mathbf{v}$  and  $\mathbf{u}$  with  $\mathbf{u}_s$ . This indicates a misalignment between the vision representation  $\mathbf{v}$  and the text representation  $\mathbf{u}$ . When the pre-trained model is tested on the data with  $p(\mathbf{u}_s|\mathbf{u}) \ll 1$  and  $p(\mathbf{u}_s|\mathbf{v}) \ll 1$ , i.e., the spurious features in the test data no longer have strong correlations with input images and the corresponding text descriptions compared to the training data, such as the waterbird image with a land background where a land background

is no longer associated with landbird, then the model may struggle on most of the test data, showing degraded zero-shot classification performance.

### 3.3 Theoretical Insights

We first theoretically analyze how multimodal spurious bias affects zero-shot classification. The insights derived from our analysis will guide the design of our multimodal spurious bias mitigation method in the following section.

Without loss of generality, we consider a zero-shot classification task with two classes  $c_1$  and  $c_2$ . Given a prompt template, we can obtain text representations for the two classes as  $\mathbf{u}_1$  and  $\mathbf{u}_2$ . Consider an image representation  $\mathbf{v}$  from class  $c_2$  with an unknown spurious feature described by the text representation  $\mathbf{u}_s$ . The zero-shot prediction  $\hat{y}$  can be obtained as follows,

$$\hat{y} = \arg \max_{i \in \{1,2\}} p(\mathbf{u}_i | \mathbf{v}). \quad (7)$$

We assume a multimodal spurious bias between  $\mathbf{u}_1$ ,  $\mathbf{v}$ , and  $\mathbf{u}_s$ , as indicated by the dashed arrows in Figure 2(a). Then, the zero-shot prediction may be biased towards the class label  $c_1$ , instead of the true class label  $c_2$ , as supported by the following theorem.

**Theorem 1.** *Consider a pre-trained CLIP model from which we obtain two text representations  $\mathbf{u}_1$ ,  $\mathbf{u}_2$  for the classes  $c_1$  and  $c_2$  respectively, an image representation  $\mathbf{v}$  with the class label  $c_2$ , and a textual spurious feature  $\mathbf{u}_s$  related to  $\mathbf{v}$ . Assume  $\mathbf{u}_1$ ,  $\mathbf{v}$ , and  $\mathbf{u}_s$  formulate a multimodal spurious bias. Then, the model is biased towards predicting  $\mathbf{v}$  as  $c_1$  instead of its true class label  $c_2$ .*

*Proof.* We first follow Eq. (4) to expand  $p(\mathbf{u}_1 | \mathbf{v})$ , i.e.,

$$p(\mathbf{u}_1 | \mathbf{v}) = \int_{\mathbf{u}_s} p(\mathbf{u}_1 | \mathbf{v}, \mathbf{u}_s) p(\mathbf{u}_s | \mathbf{v}) d\mathbf{u}_s \quad (8)$$

$$\approx p(\mathbf{u}_1 | \mathbf{v}, \mathbf{u}_s) p(\mathbf{u}_s | \mathbf{v}) \quad (9)$$

$$= p(\mathbf{u}_s | \mathbf{u}_1) p(\mathbf{u}_1), \quad (10)$$

where the approximation in (9) uses the definition of multimodal spurious bias in Definition 1, and Eq. (10) can be derived via Bayes' theorem, i.e.,

$$p(\mathbf{u}_1 | \mathbf{v}, \mathbf{u}_s) = \frac{p(\mathbf{u}_1, \mathbf{u}_s | \mathbf{v})}{p(\mathbf{u}_s | \mathbf{v})} = \frac{p(\mathbf{u}_s | \mathbf{u}_1) p(\mathbf{u}_1)}{p(\mathbf{u}_s | \mathbf{v})}, \quad (11)$$

where the last equality follows the fact that  $p(\mathbf{u}_1, \mathbf{u}_s | \mathbf{v}) = p(\mathbf{u}_1, \mathbf{u}_s)$ , i.e.,  $\mathbf{u}_s$  and  $\mathbf{u}_1$  do not depend on  $\mathbf{v}$ , as depicted in Figure 2(a). Therefore, we have the following inequality:

$$\frac{p(\mathbf{u}_1 | \mathbf{v})}{p(\mathbf{u}_2 | \mathbf{v})} \approx \frac{p(\mathbf{u}_s | \mathbf{u}_1) p(\mathbf{u}_1)}{p(\mathbf{u}_2 | \mathbf{v})} > 1, \quad (12)$$

where the inequality follows from the condition that  $\mathbf{u}_1$ ,  $\mathbf{v}$ , and  $\mathbf{u}_s$  formulate a multimodal spurious bias, i.e.,  $p(\mathbf{u}_s | \mathbf{u}_1) \approx 1$ ,  $p(\mathbf{u}_2 | \mathbf{v}) \approx 0$  given that  $p(\mathbf{u}_s | \mathbf{v}) \approx 1$ , and  $p(\mathbf{u}_1) > 0$  is a constant. Therefore, the model's prediction on  $\mathbf{v}$  is biased towards the incorrect label  $c_1$ .  $\square$

Based on the above analysis, SAGE is motivated to choose such prompt template  $\mathbf{u}$  that directly controls the multimodal spurious bias term  $p(\mathbf{u}_s | \mathbf{u}_1)$ . When a prompt induces spurious biases (i.e.,  $p(\mathbf{u}_s | \mathbf{u}_1) \approx 1$ ), the model's predictive probability on two classes  $p(\mathbf{u}_1 | \mathbf{v})$  and  $p(\mathbf{u}_2 | \mathbf{v})$  becomes arbitrarily close. This indicates a low class separation. Conversely, when another prompt template  $\mathbf{u}'$  is less affected by spurious biases (i.e.,  $p(\mathbf{u}_s | \mathbf{u}'_1) \ll 1$ ), the predictive probability for the incorrect class  $p(\mathbf{u}'_1 | \mathbf{v})$  decreases significantly, while the predictive probability for the correct class  $p(\mathbf{u}'_2 | \mathbf{v})$  remains high. This creates a large margin between  $p(\mathbf{u}'_2 | \mathbf{v})$  and  $p(\mathbf{u}'_1 | \mathbf{v})$ . Therefore, maximizing class separation margin can serve as a robust and practical proxy that effectively mitigates spurious biases without prior knowledge of the spurious attributes in advance.

### 3.4 Spuriousness-Aware Guided Exploration

**Prompt-based class separation.** Given a fixed class label  $c_i$  and an input image embedding  $\mathbf{v}$ , we construct a set of  $M$  prompt templates  $\mathcal{T} = \{T_j\}_{j=1}^M$  and generate class-specific textual descriptions  $\mathcal{D}_i = \{T_j(c_i)\}_{j=1}^M$ , where  $T_j(c_i)$  denotes the  $j$ -th prompt filled with class  $c_i$ . For example, as illustrated in Figure 2(b) with  $C = 2$  and  $M = 3$ , when  $c_1$  is "landbird",  $T_1(c_1)$  could be "a photo of a landbird". These prompts are encoded via the text encoder  $\psi$ , producing corresponding representations  $\{\mathbf{u}_i^j\}_{j=1}^M$ , where  $\mathbf{u}_i^j = \psi(T_j(c_i))$ . Given  $N$  test images denoted as  $\{x_n\}_{n=1}^N$ , we obtain their visual embeddings using the vision encoder  $\phi$ , resulting in  $\mathbf{v}_n = \phi(x_n)$  for each  $n = 1, 2, \dots, N$ .

We evaluate each prompt template based on its ability to separate different classes in the joint image-text embedding space. Concretely, for a given prompt template  $T_j$  and image  $x_n$ , we compute its cosine similarity with the image embedding  $\mathbf{v}_n$  across all classes  $c_1, \dots, c_C$ . We then define the *separation score* of  $T_j$  for  $x_n$  as:

$$\sigma_j^n = \max_{i \in \{1, \dots, C\}} \frac{\mathbf{v}_n^T \mathbf{u}_i^j}{\|\mathbf{v}_n\|_2 \|\mathbf{u}_i^j\|_2} - \min_{i \in \{1, \dots, C\}} \frac{\mathbf{v}_n^T \mathbf{u}_i^j}{\|\mathbf{v}_n\|_2 \|\mathbf{u}_i^j\|_2}. \quad (13)$$

A higher  $\sigma_j^n$  indicates that the prompt better distinguishes between classes in terms of alignment with the image embedding, suggesting it is less biased and more informative.

**Template selection and zero-shot inference.** For each image, we rank all prompt templates based on their *separation scores*  $\sigma_j^n$  and select the top- $K$  templates with the highest scores. Here,  $K$  is a hyperparameter that determines how many top-scoring templates are chosen for zero-shot inference. These templates are then used to construct  $K$  zero-shot classifiers. As illustrated in Figure 2(c), we use  $K = 2$  as an example, where the top two prompt templates are selected for zero-shot inference. For each selected template  $T_k$ , we compute the text embeddings  $\mathbf{u}_i^k = \psi(T_k(c_i))$  for all class labels  $i = 1, \dots, C$ . The final prediction for the  $n$ -th image is obtained by averaging the similarity scores across the  $K$  classifiers:

$$\hat{y}_n = \arg \max_i \frac{1}{K} \sum_{k=1}^K \frac{\mathbf{v}_n^T \mathbf{u}_i^k}{\|\mathbf{v}_n\|_2 \|\mathbf{u}_i^k\|_2}. \quad (14)$$

Method	Model	Waterbirds			CelebA			PACS			VLCS		
		AVG( $\uparrow$ )	WGA( $\uparrow$ )	HM( $\uparrow$ )	AVG( $\uparrow$ )	WGA( $\uparrow$ )	HM( $\uparrow$ )	AVG( $\uparrow$ )	WGA( $\uparrow$ )	HM( $\uparrow$ )	AVG( $\uparrow$ )	WGA( $\uparrow$ )	HM( $\uparrow$ )
ZS	CLIP-RN-50	88.7	41.0	56.1	81.6	75.2	78.3	91.8	63.3	74.9	75.5	34.1	47.0
	CLIP-ViT-B/32	80.4	27.5	41.0	78.3	68.9	73.3	96.6	82.1	88.8	75.4	20.5	32.2
	CLIP-ViT-L/14	88.6	27.6	42.1	80.5	74.0	77.1	98.1	79.8	88.0	72.4	4.1	7.8
	ALIGN	72.3	50.0	59.1	82.4	78.2	80.2	95.8	69.6	80.6	78.5	34.1	47.5
	AltCLIP	90.3	37.2	52.7	82.9	80.2	81.5	98.5	82.5	89.8	78.8	22.0	34.4
	<b>Average</b>	<b>84.1</b>	<b>36.7</b>	<b>51.1</b>	<b>81.1</b>	<b>75.3</b>	<b>78.1</b>	<b>96.2</b>	<b>75.5</b>	<b>84.6</b>	<b>76.1</b>	<b>23.0</b>	<b>35.3</b>
ROBOSHOT	CLIP-RN-50	72.1	27.6	39.9	81.6	74.9	78.1	92.3	72.4	81.1	77.6	37.6	50.7
	CLIP-ViT-B/32	74.2	39.3	51.4	82.1	75.2	78.5	96.6	83.5	89.6	77.1	35.2	48.3
	CLIP-ViT-L/14	79.8	48.1	60.0	85.3	82.2	83.7	98.0	81.3	88.9	70.9	12.2	20.8
	ALIGN	52.6	38.3	44.3	87.0	84.8	85.9	94.7	63.2	75.8	77.4	39.8	52.6
	AltCLIP	78.5	54.2	64.1	86.1	80.6	83.3	98.8	89.4	93.9	78.3	25.7	38.7
	<b>Average</b>	<b>71.4</b>	<b>41.5</b>	<b>52.5</b>	<b>84.4</b>	<b>79.5</b>	<b>81.9</b>	<b>96.1</b>	<b>78.0</b>	<b>86.1</b>	<b>76.3</b>	<b>30.1</b>	<b>43.2</b>
TIE*	CLIP-RN-50	83.8	34.1	48.5	72.2	65.8	68.9	88.6	54.7	67.6	79.3	40.5	53.6
	CLIP-ViT-B/32	86.3	55.8	67.8	85.9	69.1	76.6	97.3	83.1	89.6	79.7	32.4	46.1
	CLIP-ViT-L/14	87.6	39.1	54.1	88.2	83.5	85.8	97.7	82.5	89.5	80.3	34.1	47.9
	ALIGN	80.9	42.2	55.5	86.6	82.2	84.3	96.4	76.4	85.2	81.3	26.2	39.6
	AltCLIP	82.9	21.0	33.5	50.6	48.6	49.6	98.6	89.2	93.7	81.9	25.3	38.7
	<b>Average</b>	<b>84.3</b>	<b>38.4</b>	<b>52.8</b>	<b>76.7</b>	<b>69.8</b>	<b>73.1</b>	<b>95.7</b>	<b>77.2</b>	<b>85.5</b>	<b>80.5</b>	<b>31.7</b>	<b>45.5</b>
Ours (SAGE)	CLIP-RN-50	91.5	41.3	56.9	82.2	77.5	79.8	91.9	63.6	75.2	74.8	36.8	49.3
	CLIP-ViT-B/32	92.3	46.0	61.4	79.6	76.0	77.8	97.0	85.0	90.6	76.9	38.1	51.0
	CLIP-ViT-L/14	90.2	47.8	62.5	85.7	83.9	84.8	98.3	84.6	90.9	74.6	23.9	36.2
	ALIGN	81.6	47.0	59.6	84.3	82.3	83.3	97.6	87.0	92.0	72.9	37.5	49.5
	AltCLIP	89.1	42.6	57.6	85.2	83.3	84.2	98.4	89.5	93.7	79.9	32.5	46.2
	<b>Average</b>	<b>88.9</b>	<b>44.9</b>	<b>59.7</b>	<b>83.4</b>	<b>80.6</b>	<b>82.0</b>	<b>96.6</b>	<b>81.9</b>	<b>88.7</b>	<b>75.8</b>	<b>33.8</b>	<b>46.7</b>

Table 1: Performance on fine-grained (Waterbirds, CelebA) and coarse-grained (PACS, VLCS) spurious correlation benchmarks. The best worst-group accuracy (WGA) and harmonic mean (HM) are shown in **boldface**, while the second best WGA and HM are shown in underline. Unlike ROBOSHOT and TIE\*, which assume prior knowledge of spurious attributes, our method requires no such information.

This procedure enables a robust ensemble of diverse, class-discriminative templates without requiring any external knowledge of spurious attributes.

## 4 Experiments

### 4.1 Datasets

We experiment on two datasets with *fine-grained spurious correlations*, where each class is correlated with certain spurious features, such as backgrounds and gender.

- **Waterbirds** (Sagawa et al. 2019) is an image dataset for recognizing waterbirds and landbirds. It is generated synthetically by combining images of two kinds of birds from the CUB dataset (Welinder et al. 2010) and the backgrounds, water and land, from the Places dataset (Zhou et al. 2017).
- **CelebA** (Liu et al. 2015) is a large-scale image dataset of celebrity faces. The task is to identify hair color, non-blond or blond, with the gender as the spurious attributes.

We also experiment on two datasets with *coarse-grained spurious correlations* where classes are associated with domain-specific features.

- **PACS** (Zhou et al. 2020) is a domain generalization dataset that includes four visually different styles: Photo,

Art Painting, Cartoon, and Sketch. The task is to identify object categories (dog, elephant, giraffe, guitar, horse, house, person).

- **VLCS** (Fang, Xu, and Rockmore 2013) is a domain generalization benchmark composed of four datasets: PASCAL VOC 2007 (Everingham et al. 2010) (V), LabelMe (Russell et al. 2008) (L), Caltech101 (Bansal et al. 2023) (C), and SUN09 (Choi et al. 2010) (S). It contains five overlapping classes (bird, car, chair, dog, and person) drawn from each dataset. The main challenge is to learn invariant features that generalize across these domains.

### 4.2 Experimental Setup

**Evaluated methods.** For zero-shot classification performance comparison, we evaluate the standard baseline ZS (Zero-Shot CLIP), as well as two recent state-of-the-art debiasing methods: ROBOSHOT (Adila et al. 2024), which utilizes large language models (LLMs) to identify spurious attributes and mitigate multimodal spurious bias, and TIE\* (Lu, Chai, and Wang 2025), which introduces spurious prompts to infer pseudo-spurious labels and remove their influence from the image embeddings. Our proposed method, SAGE, by default selects the single prompt with the highest *separation score*, i.e., setting  $K = 1$  in Equation 13. All ex-

Dataset	Model	Ensemble (K=80)			Random			Ours (SAGE)		
		AVG( $\uparrow$ )	WGA( $\uparrow$ )	HM( $\uparrow$ )	AVG( $\uparrow$ )	WGA( $\uparrow$ )	HM( $\uparrow$ )	AVG( $\uparrow$ )	WGA( $\uparrow$ )	HM( $\uparrow$ )
Waterbirds	CLIP-RN-50	92.7	48.4	63.6	87.3	45.8	60.1	91.5	41.3	56.9
	CLIP-ViT-B/32	92.7	23.5	37.5	91.4	34.2	49.8	92.3	46.0	61.4
	CLIP-ViT-L/14	93.2	33.3	49.1	92.1	39.6	55.4	90.2	47.8	62.5
	ALIGN	81.7	46.1	58.9	80.0	46.6	58.9	81.6	47.0	59.6
	AltCLIP	88.3	29.6	44.3	88.0	34.1	49.2	89.1	42.6	57.6
	<b>Average</b>		89.7	36.2	51.6	87.8	40.1	55.1	88.9	<b>44.9</b>
CelebA	CLIP-RN-50	79.1	70.5	74.6	75.3	68.1	71.5	82.2	77.5	79.8
	CLIP-ViT-B/32	77.8	67.6	72.3	76.0	69.1	72.4	79.6	76.0	77.8
	CLIP-ViT-L/14	82.0	75.5	78.6	82.8	80.4	81.6	85.7	83.9	84.8
	ALIGN	80.2	74.4	77.2	81.0	76.7	78.8	84.3	82.3	83.3
	AltCLIP	81.8	78.1	79.9	83.7	80.0	81.8	85.2	83.3	84.2
	<b>Average</b>		80.2	73.2	76.5	79.8	74.9	77.3	83.4	<b>80.6</b>

Table 2: Ablation results comparing our method (SAGE) with random prompt selection and prompt ensembling on the Waterbirds and CelebA datasets. **Bold** numbers indicate the best performance among the three. Our method consistently achieves higher worst-group accuracy (WGA) and harmonic mean (HM) of WGA and average accuracy (AVG), demonstrating its effectiveness in mitigating spurious correlations.

periments were conducted on NVIDIA Quadro RTX 8000 GPUs (48GB).

**Models.** We evaluate CLIP and its variant models with different sizes and architectures: CLIP-RN-50, CLIP-ViT-B/32, CLIP-ViT-L/14, ALIGN (Jia et al. 2021), and AltCLIP (Chen et al. 2023). While the four ViT-based models are aligned with the setups in ROBOSHOT (Adila et al. 2024), we additionally include CLIP-RN-50 to diversify the backbone architectures beyond Transformers.

**Evaluation metrics.** We report the zero-shot classification performance of a model using three metrics: average accuracy (AVG), worst-group accuracy (WGA), and the harmonic mean (HM) of AVG and WGA. WGA is the *primary* robustness metric that measures the model’s worst-group performance in the test set (Sagawa et al. 2020), which can reflect the overall robustness to spurious correlations under distribution shifts. While AVG reflects overall performance, it can be dominated by the majority of the test set. To better demonstrate the overall performance for both prediction and robustness, we propose to use the harmonic mean (HM) of AVG and WGA as another main evaluation metric. The HM is defined as:

$$HM = \frac{2 \cdot AVG \cdot WGA}{AVG + WGA}, \quad (15)$$

which is sensitive to low values and penalizes imbalanced performance. In zero-shot multimodal classification, a model may achieve high AVG by performing well on most samples but still have low WGA on the worst group. HM emphasizes the importance of maintaining both high AVG and WGA. Therefore, a robust model should exhibit high WGA and HM to indicate good performance on both prediction and robustness (Hasna and Alouini 2004).

### 4.3 Main Results

We evaluate the effectiveness of our method, SAGE, in mitigating multimodal spurious biases at both the *fine-grained*

level where each class is correlated with specific spurious features (e.g., Waterbirds and CelebA) and the *coarse-grained* level where classes are associated with broader domain-specific features (e.g., PACS and VLCS). As shown in Table 1, SAGE consistently ranks among the best or second-best performers in both HM and WGA across different models and datasets, demonstrating strong overall effectiveness and robustness.

To better capture performance consistency across varying model architectures, we report results averaged over five different backbones. This shows that SAGE consistently achieves the best WGA and HM, demonstrating strong balance between accuracy and fairness, and robust performance across subgroups.

These results highlight SAGE’s ability to reduce multimodal spurious bias without sacrificing predictive accuracy. Unlike ROBOSHOT and TIE\*, which assume prior knowledge of spurious attributes, SAGE uses a fixed and general set of 80 prompts for controlled experiments, making it a practical, out-of-the-box debiasing solution for CLIP-style models. Note that SAGE is not limited to this specific setting and can be generalized to other sets of prompt templates.

### 4.4 Ablation Studies

#### Correlation Analysis of Separation Score and Accuracy.

To better understand the effectiveness of our template selection strategy, we analyze the relationship between the *separation score* (used to rank prompt templates) and the zero-shot classification performance. Specifically, we compute the Pearson correlation coefficient (PCC) between the *separation score* and WGA of each candidate template.

We perform this analysis on the CelebA dataset, a challenging benchmark characterized by subtle attribute differences and imbalanced group distributions. These properties make it well-suited for evaluating the robustness of prompt selection across diverse model backbones. For each of the

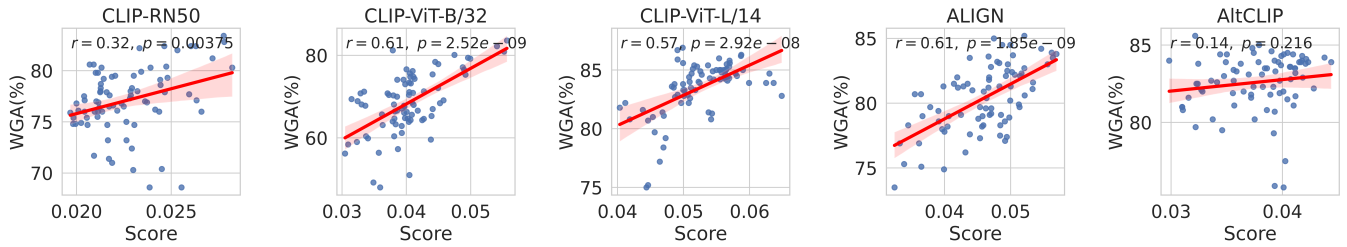


Figure 3: Pearson correlation analysis of *separation scores* and WGA on CelebA across five backbone models. Each scatter plot shows the relationship between the score assigned to a candidate template and its corresponding WGA in zero-shot inference. The consistent positive correlation observed across all models indicates that templates with higher *separation scores* tend to yield better worst-group performance, validating the effectiveness of our scoring method for robust template selection.

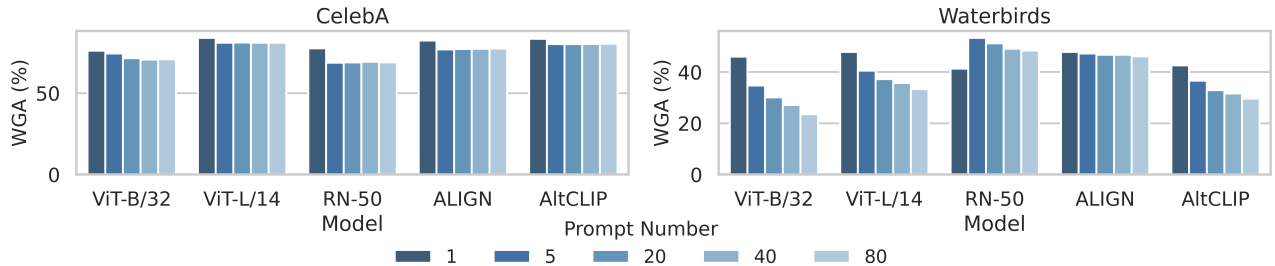


Figure 4: Ablation study on the effect of varying prompt numbers in different Models with our proposed method.

five model backbones, we plot the *separation score* against the corresponding WGA and report the PCC. As shown in Figure 3, the results consistently show a positive correlation.

These findings validate that templates with higher *separation scores* tend to yield better worst-group performance, demonstrating that the *separation score* is a reliable indicator of prompt robustness.

**Validating the Prompt Selection Strategy in SAGE.** We evaluate the effectiveness of using the *separation score* to select prompt templates by comparing three strategies: using all prompts (Ensemble), randomly selecting one prompt (Random, averaged over 3 runs), and selecting the prompt with SAGE. As shown in Table 2, our score-based method consistently achieves the best performance across nearly all settings on both Waterbirds and CelebA, significantly improving worst-group accuracy (WGA) and overall accuracy. In contrast, ensemble and random strategies often include suboptimal prompts that degrade performance. These results confirm that the *separation score* is a reliable and effective criterion for prompt selection in zero-shot inference.

**Impact of Number of Selected Prompts on Performance.** Even though templates can be ranked by their scores, the optimal number of prompts to use at inference time is not obvious. To explore this, we conduct an ablation study by varying the number of top-ranked templates  $K$  used for zero-shot inference. We evaluate  $K = 1, 5, 20, 40, 80$ , ranging from using a single best prompt to all 80 templates.

Figure 4 reports WGA across different  $K$  values on Waterbirds and CelebA using five backbones. On CelebA (left), using only the top-1 prompt consistently achieves the best

WGA, especially for CLIP-RN-50 and ALIGN. This aligns with CelebA’s fine-grained attributes, where precise prompt-image alignment is crucial and additional prompts may introduce noise.

On Waterbirds (right), the top-1 prompt yields the best performance for ViT-B/32, ViT-L/14, ALIGN, and AltCLIP, suggesting that prompts selected by SAGE remain effective even for large-scale pretrained models. In contrast, CLIP-RN-50 achieves its best results at  $K = 5$ , indicating that moderate prompt diversity may help smaller models capture more robust and complementary visual cues.

Given these results, we choose  $K = 1$  as the default setting for SAGE, as it consistently provides strong performance across datasets and model sizes. While SAGE selects prompt templates for each image, we also observe interesting trends in the most frequently selected templates.

## 5 Conclusion

In this paper, we addressed the challenge of mitigating multimodal spurious biases in pre-trained CLIP models for zero-shot classification. We first provided a theoretical definition of multimodal spurious bias and analyzed its impact on zero-shot classification. Based on these insights, we introduced SAGE, which is inspired by our theoretical insights. Our approach operates out-of-the-box with CLIP models, requiring no additional training data or prior knowledge of biases. It is broadly effective across various model sizes, architectures, and types of spurious correlations. Moreover, it achieves a strong balance between average and worst-group zero-shot classification accuracy, highlighting its practical utility in robust zero-shot predictions.

## Acknowledgements

This work is supported in part by the US National Science Foundation under grants CCF-2217071, CNS-2213700, IIS-2106913. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

## References

- Adila, D.; Shin, C.; Cai, L.; and Sala, F. 2024. Zero-Shot Robustification of Zero-Shot Models. In *International Conference on Learning Representations*.
- Bansal, M.; Kumar, M.; Sachdeva, M.; and Mittal, A. 2023. Transfer learning for image classification using VGG19: Caltech-101 image data set. *Journal of ambient intelligence and humanized computing*, 1–12.
- Beery, S.; Van Horn, G.; and Perona, P. 2018. Recognition in terra incognita. In *ECCV*, 456–473.
- Chen, Z.; Liu, G.; Zhang, B.-W.; Yang, Q.; and Wu, L. 2023. AltCLIP: Altering the Language Encoder in CLIP for Extended Language Capabilities. In *Findings of the Association for Computational Linguistics: ACL 2023*, 8666–8682.
- Choi, M. J.; Lim, J. J.; Torralba, A.; and Willsky, A. S. 2010. Exploiting hierarchical context on a large database of object categories. In *2010 IEEE computer society conference on computer vision and pattern recognition*, 129–136. IEEE.
- Chuang, C.-Y.; Jampani, V.; Li, Y.; Torralba, A.; and Jegelka, S. 2023. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*.
- Dehdashtian, S.; Wang, L.; and Boddeti, V. 2024. FairerCLIP: Debiasing CLIP’s Zero-Shot Predictions using Functions in RKHSs. In *International Conference on Learning Representations*.
- Deng, Y.; Yang, Y.; Mirzasoleiman, B.; and Gu, Q. 2024. Robust learning with progressive data expansion against spurious correlation. *Advances in Neural Information Processing Systems*, 36.
- Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88: 303–338.
- Fang, C.; Xu, Y.; and Rockmore, D. N. 2013. Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In *Proceedings of the IEEE international conference on computer vision*, 1657–1664.
- Ge, Y.; Ren, J.; Gallagher, A.; Wang, Y.; Yang, M.-H.; Adam, H.; Itti, L.; Lakshminarayanan, B.; and Zhao, J. 2023. Improving zero-shot generalization and robustness of multimodal models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11093–11101.
- Geirhos, R.; Jacobsen, J.-H.; Michaelis, C.; Zemel, R.; Brendel, W.; Bethge, M.; and Wichmann, F. A. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11): 665–673.
- Hasna, M. O.; and Alouini, M.-S. 2004. Harmonic mean and end-to-end performance of transmission systems with relays. *IEEE Transactions on communications*, 52(1): 130–135.
- Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, 4904–4916. PMLR.
- Kirichenko, P.; Izmailov, P.; and Wilson, A. G. 2023. Last Layer Re-Training is Sufficient for Robustness to Spurious Correlations. In *ICLR*.
- LaBonte, T.; Muthukumar, V.; and Kumar, A. 2024. Towards last-layer retraining for group robustness with fewer annotations. *Advances in Neural Information Processing Systems*, 36.
- Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34: 9694–9705.
- Li, L. H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N.; et al. 2022. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10965–10975.
- Liu, E. Z.; Haghighi, B.; Chen, A. S.; Raghunathan, A.; Koh, P. W.; Sagawa, S.; Liang, P.; and Finn, C. 2021. Just train twice: Improving group robustness without training group information. In *ICML*, 6781–6792. PMLR.
- Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *ICCV*, 3730–3738.
- Lu, S.; Chai, J.; and Wang, X. 2025. Mitigating Spurious Correlations in Zero-Shot Multimodal Models. In *International Conference on Representation Learning*, volume 2025, 14727–14753.
- Nam, J.; Cha, H.; Ahn, S.; Lee, J.; and Shin, J. 2020. Learning from failure: De-biasing classifier from biased classifier. *NeurIPS*, 33: 20673–20684.
- Nam, J.; Kim, J.; Lee, J.; and Shin, J. 2022. Spread Spurious Attribute: Improving Worst-group Accuracy with Spurious Attribute Estimation. In *ICLR*.
- Qiu, S.; Potapczynski, A.; Izmailov, P.; and Wilson, A. G. 2023. Simple and fast group robustness by automatic feature reweighting. In *International Conference on Machine Learning*, 28448–28467. PMLR.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.
- Russell, B. C.; Torralba, A.; Murphy, K. P.; and Freeman, W. T. 2008. LabelMe: a database and web-based tool for image annotation. *International journal of computer vision*, 77: 157–173.

- Sagawa, S.; Koh, P. W.; Hashimoto, T. B.; and Liang, P. 2019. Distributionally Robust Neural Networks. In *ICLR*.
- Sagawa, S.; Koh, P. W.; Hashimoto, T. B.; and Liang, P. 2020. Distributionally Robust Neural Networks. In *International Conference on Learning Representations*.
- Sun, Z.; Shen, S.; Cao, S.; Liu, H.; Li, C.; Shen, Y.; Gan, C.; Gui, L.; Wang, Y.-X.; Yang, Y.; et al. 2024. Aligning Large Multimodal Models with Factually Augmented RLHF. In *Findings of the Association for Computational Linguistics ACL 2024*, 13088–13110.
- Tong, S.; Liu, Z.; Zhai, Y.; Ma, Y.; LeCun, Y.; and Xie, S. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9568–9578.
- Wang, Z.; Yu, J.; Yu, A. W.; Dai, Z.; Tsvetkov, Y.; and Cao, Y. 2022. SimVLM: Simple Visual Language Model Pre-training with Weak Supervision. In *International Conference on Learning Representations*.
- Welinder, P.; Branson, S.; Mita, T.; Wah, C.; Schroff, F.; Belongie, S.; and Perona, P. 2010. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology.
- Yang, Y.; Nushi, B.; Palangi, H.; and Mirzasoleiman, B. 2023. Mitigating spurious correlations in multi-modal models during fine-tuning. In *International Conference on Machine Learning*, 39365–39379. PMLR.
- Ye, W.; Jiang, L.; Xie, E.; Zheng, G.; Ma, Y.; Cao, X.; Guo, D.; Qi, D.; He, Z.; Tian, Y.; Coffee, M.; Zeng, Z.; Li, S.; Ting-hao; Huang; Wang, Z.; Rehg, J. M.; Kautz, H.; and Zhang, A. 2025. The Clever Hans Mirage: A Comprehensive Survey on Spurious Correlations in Machine Learning. arXiv:2402.12715.
- Ye, W.; Zheng, G.; Ma, Y.; Cao, X.; Lai, B.; Rehg, J. M.; and Zhang, A. 2024. MM-SpuBench: Towards better understanding of spurious biases in multimodal llms. *arXiv preprint arXiv:2406.17126*.
- Ye, W.; Zheng, G.; and Zhang, A. 2025. Improving Group Robustness on Spurious Correlation via Evidential Alignment. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, 3610–3621.
- You, C.; Mint, Y.; Dai, W.; Sekhon, J. S.; Staib, L.; and Duncan, J. S. 2024. Calibrating multi-modal representations: A pursuit of group robustness without annotations. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 26140–26150. IEEE.
- Zhang, J.; Ma, X.; Guo, S.; Li, P.; Xu, W.; Tang, X.; and Hong, Z. 2024. Amend to Alignment: Decoupled Prompt Tuning for Mitigating Spurious Correlation in Vision-Language Models. In *Forty-first International Conference on Machine Learning*.
- Zhang, M.; and Ré, C. 2022. Contrastive adapters for foundation model group robustness. *Advances in Neural Information Processing Systems*, 35: 21682–21697.
- Zheng, G.; Ye, W.; and Zhang, A. 2024a. Learning robust classifiers with self-guided spurious correlation mitigation. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 5599–5607.
- Zheng, G.; Ye, W.; and Zhang, A. 2024b. Spuriousness-Aware Meta-Learning for Learning Robust Classifiers. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4524–4535.
- Zheng, G.; Ye, W.; and Zhang, A. 2025a. NeuronTune: Towards Self-Guided Spurious Bias Mitigation. In *Forty-second International Conference on Machine Learning*.
- Zheng, G.; Ye, W.; and Zhang, A. 2025b. ShortcutProbe: Probing Prediction Shortcuts for Learning Robust Models. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*. International Joint Conferences on Artificial Intelligence Organization. Main Track.
- Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6): 1452–1464.
- Zhou, K.; Yang, Y.; Hospedales, T.; and Xiang, T. 2020. Deep domain-adversarial image generation for domain generalisation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 13025–13032.