Amend to Alignment: Decoupled Prompt Tuning for Mitigating Spurious Correlation in Vision-Language Models

Jie Zhang [*1] **Xiaosong Ma** [*1] **Song Guo** [2] **Peng Li** [3] **Wenchao Xu** [1] **Xueyang Tang** [1] **Zicong Hong** [1]

Abstract

Fine-tuning the learnable prompt for a pre-trained vision-language model (VLM), such as CLIP, has demonstrated exceptional efficiency in adapting to a broad range of downstream tasks. Existing prompt tuning methods for VLMs do not distinguish spurious features introduced by biased training data from invariant features, and employ a uniform alignment process when adapting to unseen target domains. This can impair the cross-modal feature alignment when the test data significantly deviate from the distribution of the training data, resulting in a poor out-of-distribution (OOD) generalization performance. In this paper, we reveal that the prompt tuning failure in such OOD scenarios can be attribute to the undesired alignment between the textual and the spurious feature. As a solution, we propose **CoOPood**, a fine-grained prompt tuning method that can discern the causal features and deliberately align the text modality with the invariant feature. Specifically, we design two independent contrastive phases using two lightweight projection layers during the alignment, each with different objectives: 1) pulling the text embedding closer to the invariant image embedding and 2) pushing the text embedding away from the spurious image embedding. We have illustrated that **CoOPood** can serve as a general framework for VLMs and can be seamlessly integrated with existing prompt tuning methods. Extensive experiments on various OOD datasets demonstrate the performance superiority over state-of-the-art methods.

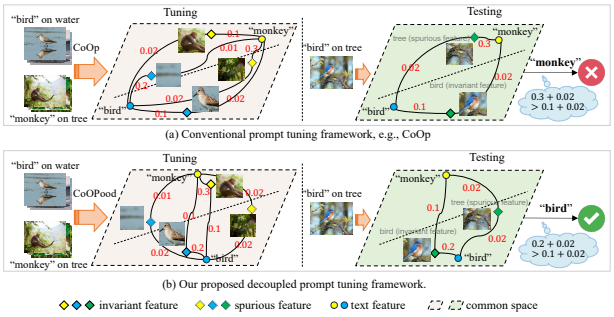


Figure 1. Compared to existing methods. Different from the conventional cross-modal alignment that may cause final prediction to rely on spurious correlation, our proposed method provides an invariant alignment to ensure more accurate prediction. The red numbers denote the similarity among different features.

1. Introduction

Recently, pre-trained vision-language models (VLMs), e.g., CLIP, have demonstrated impressive zero-shot learning performance in a wide range of downstream tasks (Radford et al., 2021), including image classification (Singh et al., 2022), object detection (Du et al., 2022; Gu et al., 2021; Li et al., 2023), and vision-language answering (VQA) (Garcia et al., 2020), etc. Different from traditional computer vision models and natural language models that interpret images or text in a unimodal manner, VLMs are capable of comprehending both visual information and textual context via a pair-wised cross-modal knowledge alignment in the semantic space (Zhou et al., 2022b). Expressly, for any new classification task, the VLM (e.g., CLIP) text-encoder first encodes the manually designed textual prompt (e.g., "`a photo of a` [`CLASS`]."), and then calculates the cosine similarity between textual features and image features for prediction. However, identifying appropriate manually designed prompts is more art than science because it requires both domain expertise and laborious prompt engineering.

To avoid the hand-crafted prompt design, some recent research (e.g., CoOp (Zhou et al., 2022b; Yao et al., 2021)) have proposed prompt tuning to directly learn prompts using training data from downstream tasks. By aligning images and texts in a common feature space using contrastive loss via learnable vectors, prompt tuning can adapt to unseen target domains in a parameter-efficient way. Despite its po-

[*]Equal contribution [1]Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China. [2]Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, China. [3]School of Cyber Science and Engineering, Xi'an Jiaotong University, Shaanxi, China.. Correspondence to: Song Guo <songguo@cse.ust.hk>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

tential, existing contrastive-based prompt tuning in VLMs has not yet considered different types of vision features (i.e., spurious and invariant features) when training data are biased, i.e., imbalanced amount of data for different categories, which further posts a question: *when spurious features exists, does the alignment process still work well?* It has been well recognized that machine learning models are prone to learning shortcuts that stem from spurious correlations (McCoy et al., 2019; Geirhos et al., 2020; Ming et al., 2022), and their out-of-distribution (OOD) generalization ability would be dramatically decreased when adapting to open-world unseen test domains. Thus, currently prompt tuning framework that heavily relies on uniform cross-modal alignment would inevitably fall into a undesired learning process, i.e., bringing textual features closer to spurious features.

A naive solution is to simply apply conventional spurious correlation mitigation methods (Bommasani et al., 2021; Nam et al., 2020; 2022; Creager et al., 2021), e,g., using two image encoders to decouple the spurious feature and invariant features. However, they cannot work in practical prompt tuning as parameters of both encoders are frozen. Also, even when the text/image encoder can be updated, the extensive computation overhead can be prohibitive in VLMs. Worse more, the existence of spurious correlation bring new challenges for the conventional contrastive learning-based alignment framework: *how to tackle the cross-modal interaction among identified spurious features and textual features requires further investigation.* To tackle these issues, in this paper, we propose **CoOPood**, a fine-grained prompt tuning method that can discern the causal features and properly align the two modalities during adaptation process. Specifically, we divide the original alignment process into two independent contrastive phases by two lightweight projection layers with the following properties: 1) pulling the text embedding closer to the invariant image embedding and 2) pushing the text embedding away from the spurious image embedding. Based on the above tuning framework, the learnt prompts are automatically able to focus only on the invariant features of OOD images, resulting in higher robustness in various downstream tasks (i.e., Figure 1(b)). We show that our method can significantly improve the model accuracy in an open-world setting when comparing with the state-of-the-art baselines over widely used models and downstream tasks (i.e., WaterBirds, CelebA, ImageNet-1K). The contributions of the paper are summarized as follows.

- To the best of our knowledge, we are the first to explore spurious correlation in prompt tuning, and explicitly elucidate that the underlying cause of the performance degradation reside in the cross-modal alignment phase.
- We design a brand-new decoupled prompt tuning framework, that can effectively align the text modality with the invariant feature with two independent contrastive learning processes.
- We conduct extensive experiments on three typical OOD image classification tasks. The empirical evaluation shows the superior performance of the proposed **CoOPood** over the state-of-the-art approaches.

2. Related Work

2.1. Pre-trained Vision-Language Models

Pre-trained vision-language models (VLMs) have emerged as a prominent trend to jointly learn text and image embeddings with large-scale image-text paired datasets (Jia et al., 2021; Radford et al., 2021; Zhang et al., 2022b). A representative work is CLIP (Garcia et al., 2020), which aggregates 400 million image-text pairs from websites, facilitating the vision-language representation learning using a contrastive objective. During inference, the names of categories to be recognized are filled into a properly designed prompt template, such as "`a photo of a` [`CLASS`]". Other VLMs, like ALIGN (Jia et al., 2021) and LiT (Zhai et al., 2022) are then proposed towards the same goal and can be extended to more challenging visual recognition tasks (Liang et al., 2022; Wang et al., 2022; Liang et al., 2023). To further boost the performance of CLIP to downstream tasks, several approaches focus on fine tuning the models in a parameter-efficient manner, including CLIP-adapter (Gao et al., 2023) and TIP-adapter (Zhang et al., 2021).

2.2. Prompt Tuning for Vision-Language Models

Instead of fine-tuning the entire model for downstream tasks, prompt tuning serves as a promising paradigm in both natural language processing (Kenton & Toutanova, 2019), and computer vision (Chen et al., 2020) via training a learnable prompt with limited trainable parameters, e.g., "$[v_1, v_2, \cdots, v_L, \texttt{CLASS}]$". As a pioneer work, CoOp (Zhou et al., 2022b) applies prompt tuning to CLIP. By tuning the prompt on a collection of few-shot training samples, CoOp effectively improves CLIP's performance on the corresponding downstream tasks. Furthermore, CoCoOp (Zhou et al., 2022a) is proposed as an enhancement of CoOp. Specifically, CoCoOp introduces a lightweight network to generates additional image-conditional context for each image and combine it with the learnable prompt for improving generalization in unseen classes. MaPLe (Khattak et al., 2023) tunes both the vision prompt and text prompt via a vision-language coupling function to induce cross-modal synergy. Besides that, another line of work focuses on using multiple text prompts to achieve more fine-grained contrastive learning phases, i.e., PLOT (Chen et al., 2023), ProDA (Lu et al., 2022), MVLPT (Shen et al., 2024), ProD (Ma et al., 2023).

In this paper, instead of designing better prompt architecture, we aim to mitigate the negative impact of spurious correla-

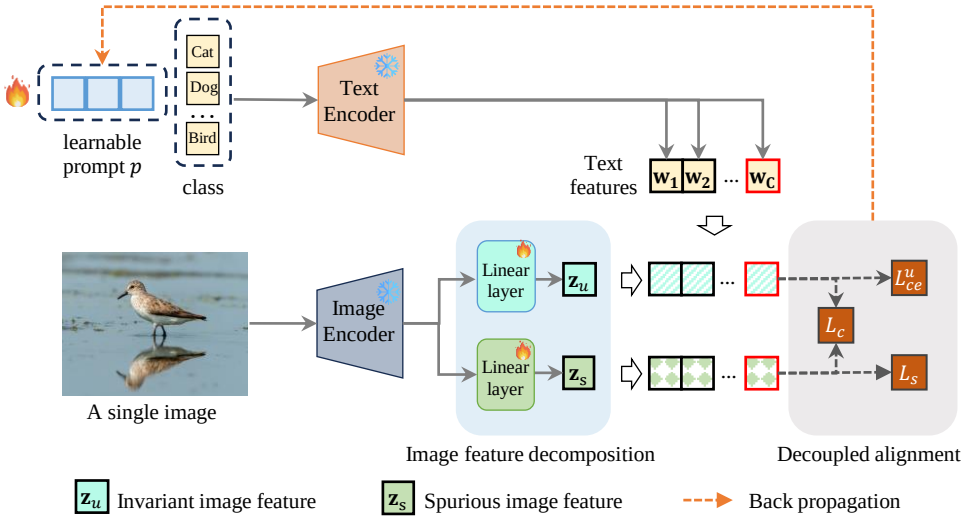


Figure 2. The framework of the decoupled prompt tuning for mitigating spurious correlation (**CoOPood**). We first use two simple projection layers (e.g., one or two linear layer(s)) to automatically decouple the spurious feature from the output of image encoder. Then, two independent contrastive learning process are designed to ensure unbiased alignment among image and text modality. $\mathcal{L}_{ce}^u$ is the standard cross-entropy loss on invariant feature, $\mathcal{L}_s$ is the constraint to prevent unnecessary correlation between spurious feature and corresponding text feature, $\mathcal{L}_c$ is a regularization term to guarantee the effectiveness of image feature decomposition.

tion derived from imbalanced training data on prompt tuning to improve OOD generalization performance. Note that our proposed CoOPood method is orthogonal with most prompt tuning mechanisms.

2.3. Spurious Correlation under OOD Generalization

In practical open-world and complex scenarios, machine learning models are shown to inevitably fall into the problem of OOD generalization, where the new data have different distributions with the training data (Liu et al., 2021). One typical research topic in solving OOD generalization problem is to explore approaches to mitigate the spurious correlation between class labels and spurious attributes (Sagawa et al., 2019). For example, in Figure 1(a), a pre-trained VLM (e.g., CoOp) has learned to use '*tree*' to identify '*bird* on tree' since the concept of monkey and tree are often appear together in training dataset, instead of actually learning the bird itself.

To mitigate the spurious correlation in learning process, some early works rely on the predefined or manually annotated spurious correlations (Sagawa et al., 2019; Li & Vasconcelos, 2019; Izmailov et al., 2022; Kirichenko et al., 2023) to recover state-of-the-art performance on benchmark spurious correlation problems by simply retraining the last layer of the model on a small held-out dataset where the spurious correlation does not hold, which may be expensive and sometimes impractical. Thus, many of the existing works assume that spurious correlation can be detected by a well-trained empirical risk minimization (ERM) model (Zhang et al., 2022a; Yang et al., 2023; Wei et al., 2023). Another clever method is to add one additional model/encoder to disentangle spurious features from imbalanced data (Luo et al., 2022; Hu et al., 2022; Anonymous, 2024). In terms of mitigation mechanism, synthesizing minority samples/features to balance the dataset is also widely utilized in removing the spurious correlation (Yao et al., 2022; Han et al., 2022; Liu et al., 2023; Kim et al., 2023; Wu et al., 2022).

2.4. Summary

When it comes to multi-modal models, spurious correlation under OOD generalization problem still cannot be ignored. Yang et al., (Yang et al., 2023; Shu et al., 2023) are the first to consider a fine-tuning approach for mitigating spurious correlations in multi-modal models, e.g., CLIP. However, these studies all focus on updating the parameters of model backbone, and mitigating spurious correlation in prompt tuning has not been exploited. Moreover, the existence of spurious correlation bring new challenges for the conventional contrastive learning-based alignment framework: *how to tackle the cross-modal interaction among identified spurious features and textual features to further enhance the prompt tuning process?* Inspired by the above observation, we are motivated to develop a novel prompt tuning framework to achieve unbiased cross-modal alignment.

3. Methodology

In this section, we first introduce the preliminary in Section 3.1. Then, in Section 3.2, we elaborate the proposed CoOPood framework, which contains the image feature decomposition phase and decoupled prompt tuning phase. The detailed workflow of CoOPood is shown in Figure 2.

3.1. Preliminary

Contrastive Language-Image Pre-training (CLIP). CLIP (Garcia et al., 2020) is a typical dual-encoder architecture consisting of an image encoder that maps the image input into a feature vector and a text encoder that does the same for the text input. With the goal of aligning the image feature space and text feature space by contrastive learning, CLIP can acquire the zero-shot transfer ability to downstream tasks. We denote a CLIP model by $\mathcal{M} = \{f, g\}$, with f and g being the image and text encoder, respectively.

In a downstream task, a hand-crafted prompt is fed into the text encoder to synthesize a zero-shot linear classifier by embedding the class names of the target dataset. We take a image classification as an example, the "[CLASS]" token can be first extended by a template, such as "a photo of a [CLASS]". Then, the sentence is treated as a prompt and is encoded by the text encoder g to derive a weight vector $\mathbf{w}_c$ for class c, where $c \in \{1, \cdots, C\}$, and C is the total number of categories. Given a single test image $x_{\texttt{test}}$, it is then fed into the image encoder f to generate image embedding $\mathbf{z}_{\texttt{test}}$. After that, the prediction probability is computed by calculating the cosine similarity between the image embedding and C text embeddings:

$$p(c|x_{\texttt{test}}) = \frac{\exp(\texttt{sim}(\mathbf{z}_{\texttt{test}}, \mathbf{w}_c)/\tau)}{\sum_{j=1}^{C} \exp(\texttt{sim}(\mathbf{z}_{\texttt{test}}, \mathbf{w}_j)/\tau)}, \quad (1)$$

where τ denotes the temperature parameter, $\texttt{sim}(\cdot)$ denotes the cosine similarity. Although Eq. (1) can be easily applied for zero-shot prediction, since CLIP employs a fixed hand-crafted prompt (e.g., "a photo of a []") to generate the textual embedding, the generability ability to the downstream tasks are potentially restricted. To address the above problem, Context Optimization (CoOp) (Zhou et al., 2022b) has been proposed to automatically learn a set of continuous context vectors for generating task-related textual embeddings.

Context Optimization (CoOp). Denote the learnable context vector as $\mathbf{v} = [v_1, v_2, \cdots, v_L]$, with each v_l, $l \in \{1, \cdots, L\}$ being a vector with the same dimension as word embeddings (i.e., 512 for CLIP), and L is a hyperparameter specifying the number of context tokens. Then the soft prompt in CoOp can be represented as:

$$\mathbf{t} = [v_1, v_2, \cdots, v_L, \texttt{CLASS}]. \quad (2)$$

By forwarding the prompt for the c-th class, i.e., $\mathbf{t}_c = [\mathbf{v}, c]$, into the text encoder g, we can obtain the textual class embedding $\widetilde{\mathbf{w}}_c = g(t_c)$. Let $\mathcal{D}_{\texttt{train}} = \{(x_i, y_i)\}_{i=1}^{N}$ be the training dataset from the downstream task, where x_i is the i-th input data sample in space X, y_i is the corresponding label in $Y = \{1, 2, \ldots, C\}$, N is the size of the dataset. For all training data, CoOp calculates the probabilities of all classes and minimizes the cross-entropy loss $\mathcal{L}_{ce}$ to tune the prompt. The problem can be formulated as:

$$\mathbf{t}^* = \arg\min_{\mathbf{t}} \mathcal{L}_{ce}, \quad (3)$$

$$\text{where } \mathcal{L}_{ce} = -\sum_{i=1}^{N} y_i \log p(y_i|x_i), \quad (4)$$

and $p(y_i|x_i) = \frac{\exp(\texttt{sim}(\mathbf{z}_i, \widetilde{\mathbf{w}}_{y_i})/\tau)}{\sum_{j=1}^{C} \exp(\texttt{sim}(\mathbf{z}_i, \widetilde{\mathbf{w}}_j)/\tau)}$, $\mathbf{z}_i = f(x_i)$, $\widetilde{\mathbf{w}}_{y_i} = g(\mathbf{t}_{y_i}) = g([\mathbf{v}, y_i])$. Note that the parameters of image encoder and the text encoder are frozen during the tuning phase, and only the prompts are optimized. For large-scale pre-trained models, prompt tuning is often more effective and efficient than traditional finetuning methods such as linear probing and full fine-tuning of all layers.

However, CoOp faces a challenge that the transfer performance drops when spurious correlation exists in the tuning phase, even under-performs the zero-shot CLIP, i.e., from 96.7% (ID) to 83.1% (OOD). We find that the reason behind this is due to the current alignment pattern among image and text modality, whose effectiveness lies in an implicit assumption: *the test data is in-distribution (ID) with respect to its training data, and the image feature embeddings are unbiased (i.e., no spurious correlation exits).* Unfortunately, above assumption may not always hold, especially for some practical scenarios that data attributes have strong correlation with their categories, e.g., misclassifying boats when there is no water in the background. To address this issue, we propose a novel and efficient prompt tuning framework, called CoOPood, to provide an unbiased alignment strategy between image and text modality.

3.2. Overview of CoOPood

Conventional prompt tuning framework aligns images and texts in a common feature space by directly using the features extracted from frozen encoders, imposing inaccurate and even biased learning process. Instead, we propose to decouple the features extracted by the vision encoder into two parts: invariant features and spurious features. Then, two independent contrastive learning processes are designed to achieve unbiased cross-modal alignment.

Image Feature Decomposition. First, instead of roughly adding an additional vision encoder with a large number of parameters, we add two projection layers to decouple the spurious features and invariant features from the output of image encoder. In practice, these projection layers can just be some linear layers. The two projection layers are denoted by ϕ and ψ, respectively. Then, for a single image x_i, the output of ϕ and ψ can be regarded as invariant image embedding and spurious image embedding as follows:

$$\mathbf{z}_{i,u} = \phi(\mathbf{z}_i), \quad \mathbf{z}_{i,s} = \psi(\mathbf{z}_i). \quad (5)$$

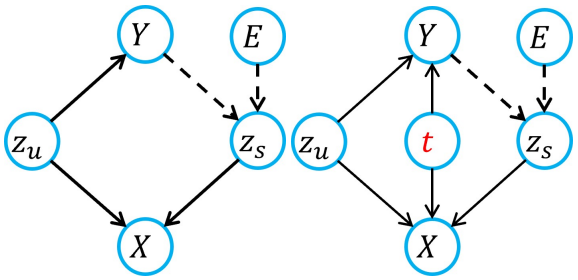


Figure 3. The evolution of SCM in different scenarios. X and Y denote the input and target space respectively. E is the indicator of spurious correlation, which is usually the environment variable. $\mathbf{z}_u$ and $\mathbf{z}_s$ denote the invariant and spurious features respectively. Dotted arrows indicate unstable causal relations that can vary in different environments. (a) SCM in traditional uni-modal learning; (b) New SCM in our work, where $\mathbf{t}$ denotes the additional textual feature that does not exit in traditional uni-modal learning.

To ensure ϕ and ψ can efficiently capture the corresponding features, we theoretically analyze the image feature decomposition problem by formulating structured causal models (SCM) (Ahuja et al., 2021) to simulate the data generating process in the downstream task. A valid SCM is depicted by a directed acyclic graph where each node represents a random variable and each edge describe a directed functional relationship between the corresponding variables (Ahuja et al., 2021).

When we study the prompt tuning in OOD setting, the additional textual modality need to be considered. The detailed SCMs are show in Figure 3. According to the causal Markov condition (Theorem 1.4.1) proved in (Pearl, 2009), we can obtain following Lemma,

Lemma 1 (Conditional Independence)**.** *If the data generating mechanism of each VLM obeys the causal graph in Figure 3, we have:*

- $\boldsymbol{z}_u \perp \boldsymbol{z}_s \mid Y$*, which means that the invariant features* $\boldsymbol{z}_s$ *are conditionally independent of the spurious features* $\boldsymbol{z}_u$ *given variable* Y*.*
- $\boldsymbol{z}_s \perp \boldsymbol{t} \mid Y$*, which means that the spurious features* $\boldsymbol{z}_s$ *are conditionally independent of the text features* $\boldsymbol{t}$ *given variable* Y*.*

The above Lemma can be easily proved using the d-separation criterion (Pearl, 2009), we omit it here and provide the detailed proof in the appendix A. Then, we can utilize the first signature of conditional independence to extract the spurious features. Specifically, we formulate $(\mathbf{z}_u \perp \mathbf{z}_s \mid Y)$ as a regularization term $\mathcal{L}_c$ to minimize the conditional mutual information (CMI) between the invariant features and the spurious features,

$$\mathcal{L}_c = I(\mathbf{z}_u; \mathbf{z}_s \mid Y), \tag{6}$$

where $I(\cdot)$ denotes the Shannon mutual information. In the practical implementation, we adopt the metric used in (Jiang & Veitch, 2022) to estimate the above conditional mutual information, i.e.,

$$I(\mathbf{z}_u; \mathbf{z}_s \mid Y) := \left\| \frac{1}{N} \sum_{i=1}^{N} \mathbf{z}_{i,u} \left(\mathbf{z}_{i,s} - \sum_{j=1}^{N} \frac{q_j^i}{\sum_{j=1}^{N} q_j^i} \mathbf{z}_{j,s} \right) \right\|_1,$$

where $q_j^i = 1$ if and only if $y_j = y_i$; otherwise $q_j^i = 0$.

Decoupled Prompt Tuning. After decoupling the spurious image features and invariant image features, we design a two-fold contrastive learning process to align two modalities in an unbiased manner. The main idea is to pull the text embedding closer to invariant image embedding while pushing text embedding away from spurious image embedding.

Specifically, for the constrast process between invariant image feature and text features, we continue to use the cross-entropy loss as that in CoOP, which can be re-presented as,

$$\mathcal{L}_{ce}^{u} = -\sum_{i=1}^{N} y_i \log p_u(y_i|x_i), \tag{7}$$

where $p_u(y_i|x_i) = \frac{\exp(\texttt{sim}(\mathbf{z}_{i,u}, \widetilde{\mathbf{w}}_{y_i})/\tau)}{\sum_{j=1}^{C} \exp(\texttt{sim}(\mathbf{z}_{i,u}, \widetilde{\mathbf{w}}_j)/\tau)}$. In terms of a spurious image feature, the second signature of conditional independence indicates that the one text embedding should be tuned away far from its spurious image feature. Therefore, we regularize the predictive distribution to be close to a uniform distribution to prevent the model from classifying the image into one of classes and thus destroy the its discrimination ability on spurious feature, i.e.,

$$\mathcal{L}_s = \sum_{i=1}^{N} \ell_{KL}(p_s(y_i|x_i) \mid p_0), \tag{8}$$

where $p_s(y_i|x_i) = \frac{\exp(\texttt{sim}(\mathbf{z}_{i,s}, \widetilde{\mathbf{w}}_{y_i})/\tau)}{\sum_{j=1}^{C} \exp(\texttt{sim}(\mathbf{z}_{i,s}, \widetilde{\mathbf{w}}_j)/\tau)}$, $p_0 = \frac{1}{C}$ denotes a uniform distribution, and ℓ_{KL} is the Kullback-Leibler (KL) divergence loss.

By combining all loss functions, the final objective function is:

$$\mathcal{L} = \mathcal{L}_{ce}^{u} + \alpha \mathcal{L}_s + \beta \mathcal{L}_c \tag{9}$$

where α and β are used to balance the effect of different terms.

4. Evaluation

4.1. Experimental Setup

Datasets. We evaluate the proposed CoOPood over three datasets: Waterbirds (Sagawa et al., 2019), CelebA (Liu et al., 2015), and ImageNet-1K (Russakovsky et al., 2015). Waterbirds is a commonly used benchmark dataset for studying spurious correlations. The task is to classify whether

Table 1. Average and worst-group classification accuracy of compared methods on two backbones over three benchmark datasets. Worst groups: waterbird on land for Waterbirds dataset; blond males for CelebA dataset; baby pacifier without baby for ImageNet-1K dataset. Asterisk ($*$) in *Group DRO and *Group CoOPood denotes that using group information during training.

#Method	ResNet-50 (%)						ViT-B/32 (%)					
	Waterbirds		CelebA		ImageNet		Waterbirds		CelebA		ImageNet	
	Avg.	Worst	Avg.	Worst	Avg.	Worst	Avg.	Worst	Avg.	Worst	Avg.	Worst
Pre-trained CLIP	68.35	42.21	**83.32**	**67.78**	67.15	36.08	64.53	40.34	**85.13**	**69.44**	75.25	50.51
CoOp	78.98	48.91	76.73	26.11	88.55	78.87	77.10	43.93	76.26	25.56	93.45	88.66
ERM	81.50	57.17	77.33	27.78	89.55	80.93	78.18	47.98	76.05	23.89	94.35	88.66
CoOPood	**82.38**	**60.28**	78.10	31.11	**93.05**	**86.08**	**79.85**	**53.74**	76.98	27.22	**95.35**	**90.72**
*Group DRO	88.93	83.33	87.98	71.11	95.75	93.30	86.98	79.91	89.85	79.44	96.90	93.81
*Group CoOPood	**89.71**	**85.64**	**88.82**	**72.24**	**97.68**	**95.36**	**88.05**	**80.53**	**90.18**	**81.33**	**97.17**	**94.33**

an image shows a landbird or a waterbird. The background (land and water) from Places dataset (Zhou et al., 2017) can be used as a spurious attribute for bird classification. The groups correspond to images of landbirds on land background (G_1), landbirds on water background (G_2), waterbirds on land background (G_3) and waterbirds on water background (G_4) with proportions 73.0%, 3.8%, 1.2%, and 22.0% of the data, respectively; the group G_3 is the minority group. In the training set, landbirds appeared more often on land backgrounds, while waterbirds appeared more often on water backgrounds, so models fine-tuned on this dataset tended to rely on backgrounds rather than birds. However, in the testing set, both landbirds and waterbirds have the same probability of appearing on a land background as on a water background, which leads to a degradation of the model's performance.

Similar to Watebirds, CelebA is a hair color prediction dataset, which also has 4 groups: non-blond females (G_1), non-blond males (G_2), blond females (G_3) and blond males (G_4) with proportions 3.9%, 73.9%, 21.1%, and 1.1% of the data, respectively; the group G_4 is the minority group, and the gender serves as a spurious feature.

In ImageNet-1K dataset, there are some features spuriously correlated with some categories (Singla et al., 2021). For example, for Baby pacifier class, the spurious attribute is baby face. Samples without babies in the image are susceptible to being classified as water bottles rather than baby pacifier. CLIP using ResNet-50 has a 98.2% classification accuracy for samples with babies in the image, but only 36.1% for samples without babies. We use the water bottle class and the baby pacifier class in ImageNet-1K as the training set, which has three groups: water bottles (G_1), baby pacifier without baby (G_2), baby pacifier with baby (G_3) with proportions 73.9%, 5.2%, and 20.9% of the data, respectively; the group G_2 is the minority group. Note that since the validation set for ImagenNet contains only 50 images per class, we transferred a portion of the data from the original training set to the test set.

Table 2. Performance comparison over three more complicated OOD datasets: PACS, ImageNet-A and ImageNet-R, and two regular datasets: Food101, Flower102.

Dataset \ #Method		Pret-trained CLIP	CoOp	ERM	CoOPood
PACS	Avg.	90.75	91.28	91.33	**91.98**
	Worst	79.42	80.92	79.37	**81.15**
ImageNet-A	Avg.	21.69	37.50	56.07	**56.45**
ImageNet-R	Avg.	55.98	63.72	64.16	**64.58**
Food101	Avg.	75.21	78.93	79.57	**79.66**
Flower102	Avg.	60.98	89.93	90.52	**90.91**

Baselines. We compare the performance of CoOPood with the state-of-the-art methods. In addition to zero-shot CLIP (Radford et al., 2021), we also include CoOp (Zhou et al., 2022b), a widely adopted prompt tuning method, which only minimize the contrastive loss $\mathcal{L}_{ce}^{u}$; Empirical Risk Mimimization (ERM), the standard technique for minimizing classification loss which also only minimize the $\mathcal{L}_{ce}^{u}$. Different from CoOp, under our model framework, the ERM method will also use the invariant projection layer, which is discussed in detail in the experimental Section 4.3; Group DRO (Sagawa et al., 2019), which uses group information on training set and adaptively increases the weight of the worst-group examples during training. It should be noted that Group DRO needs the group label of each training sample, which is not necessary in other methods.

Implementation Details. In all experiments, we use the publicly available CLIP model with the ResNet-50 (He et al., 2016) and ViT-B/32 (Dosovitskiy et al., 2020) as the backbone model. The prompt used in all methods has 4 learnable tokens and initialized as the default one "`a photo of a`". When comparing the performance with baselines, we optimize the prompts for 50 epochs with SGD optimizer and a cosine decay learning rate scheduler, the initial learning rate is 0.002. The batch size of images is 32 on all datasets. For CoOPood, unless otherwise specified, the value of hyper-parameters α and β are 1.0 and 2.0 for CelebA and ImageNet-1K; 1.0 and 10.0 for Waterbirds.

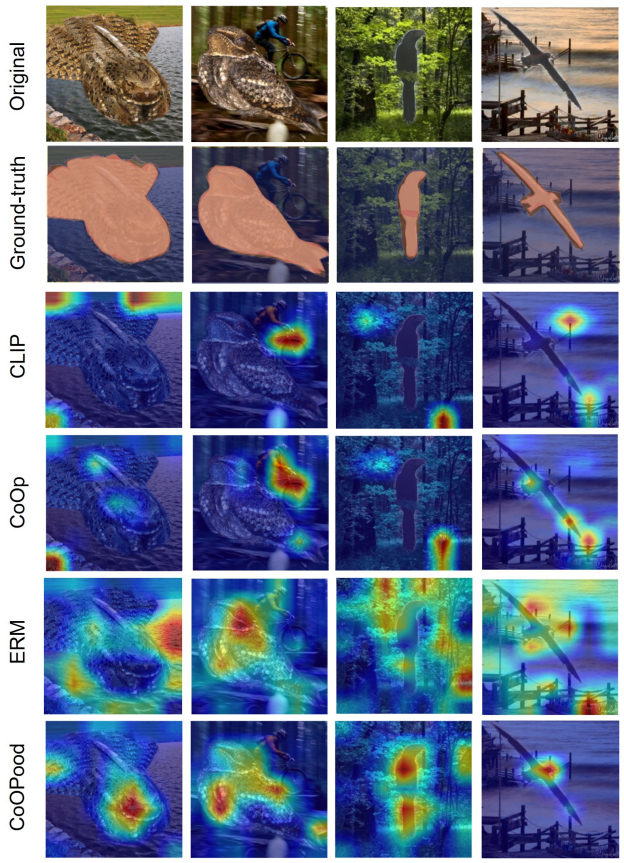


Figure 4. GradCAM explanations for different approaches based on CLIP ResNet-50 for the Waterbirds dataset.

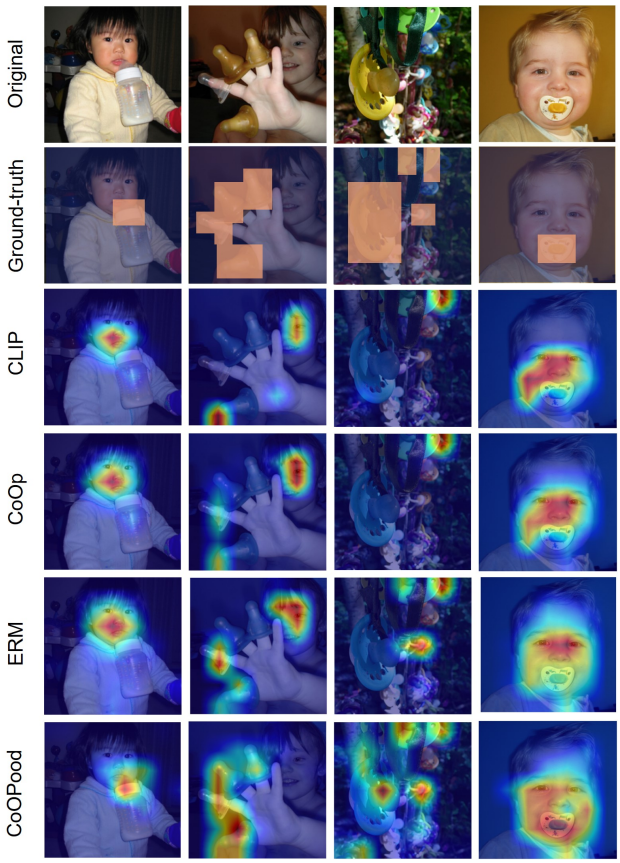


Figure 5. GradCAM explanations for different approaches based on CLIP ResNet-50 for the ImageNet-1K dataset.

We do all experiments on a workstation with an RTX 4090 GPU, a 3.0-GHZ Intel Core i9-13900K CPU and 64GB of RAM.

4.2. Performance Comparison

OOD Datasets with 2 Classes. First, we compare our CoOPood with the baseline methods on 2 backbones over 3 benchmark datasets. The classification accuracy is listed in Table 1. It should be noted that *Group DRO and *Group CoOPood are trained with group information (i.e., the group label for each training image). From the results in Table 1, when there is no group information, it can be observed that proposed CoOPood provides superior OOD generalization performance than baselines on Waterbirds and baby pacifier of ImageNet-1K. Although the average classification accuracy does not improve much, there is a significant performance improvement on the worst-group classification accuracy. For the CelebA, the worst-group accuracy does not increase with our proposed decoupled prompt tuning, while the pre-trained CLIP has the best performance. This phenomenon is also confirmed by previous work (Mao et al., 2022; Sagawa et al., 2019; Yang et al., 2023). They suggest that this is partly due to the properties of the CelebA dataset itself and partly related to the division of the training set. Specifically, the number of training samples in the worst group (i.e., G_4, blond males) is too low, resulting in G_4 having a low representation in the training data. Thus, when the group information is introduced into the training (i.e., *Group DRO and *Group CoOPood) and the loss weight of the worst-group is increased, the representation of G_4 will also increase. There is a detailed discussion about the division of CelebA Dataset's training set in Appendix C. When group information is available during training, our *Group CoOPood not only outperforms the pre-trained CLIP on CelebA, but also surpasses all baselines on all datasets, including the *Group DRO.

OOD Datasets with More Classes. Considering that the above comparisons over Waterbirds, CelebA and ImageNet-1K dataset only involve two classes, we further conduct experiments on other kinds of OOD datasets with more classes, i.e., PCAS (Li et al., 2017) dataset with 7 class and ImageNet variants like ImageNet-A (Hendrycks et al., 2021b) and ImageNet-R (Hendrycks et al., 2021a) with 200 classes. We use the regular setting of those datasets. For PACS, we adopt the "leave-one-domain-out" strategy and the target domain with the lowest accuracy is the worst

Table 3. Performance Comparison over more baselines.

#Method	ViT-L-14, Waterbirds (%)	
	Average acc.	Worst-group acc.
TextSpan	79.90	45.20
RobotShot	84.13	72.90
CoOPood	90.65	75.24
*Group DRO	91.49	86.15
*Group DFR	93.32	85.71
*Group PDE	88.66	76.24
*Group SDM	88.16	80.14
*Group CoOPood	93.49	88.81

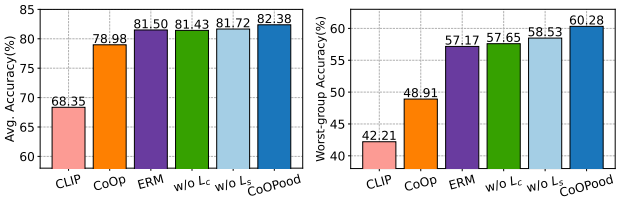
Table 4. The ablated components of compared methods.

#Method	Ablated components
CoOPood	None
CoOPood w/o $\mathcal{L}_c$	w/o $\mathcal{L}_c$
CoOPood w/o $\mathcal{L}_s$	w/o $\mathcal{L}_s$
ERM	w/o $\mathcal{L}_c$, $\mathcal{L}_s$; ψ
CoOp	w/o $\mathcal{L}_c$, $\mathcal{L}_s$; ψ, ϕ
CLIP	w/o $\mathcal{L}_c$, $\mathcal{L}_s$; ψ, ϕ; soft prompt t

group. For ImageNet-A and ImageNet-R, because data from different domains are mixed together, we can only calculate average accuracy. The value of hyper-parameters α and β are 1.0 and 2.0, other experimental settings are the same as in Table 1. As shown in upper half of Table 2, we can see that our proposed CoOPood method still has a good performance than other baselines on PACS, ImageNet-A, ImageNet-R dataset.

In-distribution Datasets. To verify that our proposed CoOPood can also perform well on other regular datasets, we then conduct new experiments on two regular datasets, i.e., Food-101 (Bossard et al., 2014), Flower-102 (Nilsback & Zisserman, 2008). Detailed results are shown in the bottom half of Table 2. We can see that our method still has a good performance on the normal data without spurious features, demonstrating using the learned "invariant image feature" will not hurt the model's performance on in-distribution data.

More Baselines on Vit-L-14. Apart from the already included baselines on spurious correlation mitigation, i.e., ERM, *Group DRO, we select other five referenced methods as additional baselines to further verify the effectiveness of our proposed CoOPood from two perspectives: 1) traditional spurious correlation mitigation methods, DFR (Kirichenko et al., 2023), PDE (Deng et al., 2024); 2) CLIP-based spurious correlation mitigation methods, i.e., SDM (Yang et al., 2023), TextSpan (Gandelsman et al., 2024), and Robothot (Adila et al., 2024). It is worth noting that we do not include these baselines in Table 1 due to the non-



(a) Average accuracy (b) Worst-group accuracy

Figure 6. Average and worst-group classification accuracy of compared methods on ResNet-50 and Waterbirds dataset. w/o $\mathcal{L}_c$ and w/o $\mathcal{L}_s$ means CoOPood does not apply the object function $\mathcal{L}_c$ and $\mathcal{L}_s$ respectively.

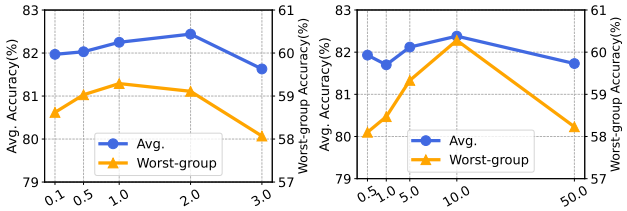
repeatability of some methods, i.e., TextSpan, Robothot. Among these five baselines, both DRF, PDE and SDM rely on a group-balanced subset of training data to re-train the last linear layer or prevent the learning of spurious features, which assumes the group information is known. Therefore, to guarantee a fair comparison, we extended DRF, PDE and SDM to VLM and named them *Group DFR, *Group PDE and *Group SDM, so as to compare it with *Group CoOPood. In addition, since TextSpan and RoboShot are training-free methods with the help of ChatGPT, we directly use the results of these two papers on the premise of applying the same experimental settings, i.e., dataset, backbone model. Then, we record the average model accuracy and worst-group accuracy across four groups of Waterbirds using ViT-L-14 as the backbone. As shown in Table 3, We can see that our method still has a good performance compared to the state-of-the-art methods.

Visualization. Figure 4 and Figure 5 show the visual explanation maps. We chose the waterbirds, baby pacifier class from ImageNet-1K and ResNet-50 to observe the model's attention across different methods. The results show that the pre-trained CLIP, CoOp, and ERM all have varying degrees of spurious correlation, i.e., a significant portion of the model's attention is not focused on the ground truth (the bird and the pacifier). In contrast, our CoOPood significantly mitigates the spurious correlation, and most of the model's attention is contained within the ground truth region.

4.3. Ablation Study

In this subsection, detailed analyses are shown to help understand the superiority of our proposed CoOPood method, including the analysis on the objective function, analysis on the sensitivity of hyper-parameters, and analysis on the computational overhead. We also analyse the context length and initialization of prompt, the division of training set of CelebA dataset and the few-shot training set in Appendix.

Analysis on the Objective Function. To demonstrate the effectiveness of our proposed CoOPood, we systematically evaluate the performance when adopting different objective functions on ResNet-50 and Waterbirds dataset. The ablated components of compared methods are shown in Table 4.



(a) Results for different α (b) Results for different β

Figure 7. Analysis on the sensitivity of hyper-parameters α and β.

Figure 6 shows the quantitative analysis on different components of CoOPood. First, we use pre-trained CLIP as the basic baseline, which only use the fixed prompt "`a photo of a`" to classify images without tuning. Secondly, CoOp tunes the prompt only with the contrastive loss $\mathcal{L}_{ce}^{u}$. Thirdly, compare with CoOp, ERM adds an projection layer ϕ after the image encoder, thus the contrastive loss $\mathcal{L}_{ce}^{u}$ applies on the $\mathbf{z}_u$ and text embedding. Note that there is no spurious projection layer ψ and no Shannon mutual information loss $\mathcal{L}_c$ in ERM, which means that the $\mathbf{z}_u$ is just the output of ϕ, does not represent the decoupled invariant features. It can be observed that the accuracy has improved compared to CoOp, which demonstrates the benefits of the projection layer. This result is also mentioned in another work about vision-language models (Gao et al., 2023). Then, Compared to CoOPood, w/o $\mathcal{L}_c$ means does not apply the Shannon mutual information loss $\mathcal{L}_c$, and the performance is worse than CoOPood. Finally, the case of using all objective functions, i.e., the complete CoOPood, has the best performance, which demonstrates that both the loss function $\mathcal{L}_c$ for decoupling the spurious feature and invariant feature, and $\mathcal{L}_s$ for regulating the discrimination ability of spurious feature can further improve the prompt tuning process.

Analysis on the Sensitivity of Hyper-parameters. To explore the sensitivity of hyper-parameters α and β for CoOPood, we conduct experiments with different values of α and β on ResNet-50 and Waterbirds dataset. The default values of α and β are 1.0 and 2.0, respectively. Other experimental settings are the same as Table 1. The results are shown in Figure 7. For α, the larger value means the greater effect of $\mathcal{L}_s$. When $\alpha = 1.0$, our proposed method achieves the best worst-group performance on all five accuracies and the performance does not degrade severely with α changes. For β, the larger value means the greater effect of $\mathcal{L}_c$, namely, the invariant image embedding will be pushed further away from spurious image embedding. The results show that the proposed method achieves the best worst-group accuracy when $\beta = 10.0$. It can be seen that CoOPood is not sensitive to the choice of hyper-parameters α and β in most cases. Results of CoOPood with different hyper-parameters settings in Figure 7 still outperform baselines in Table 1.

Analysis on the Computational Overhead. Table 5 shows the computational overhead of CoOPood in comparison with CoOp and ERM. Although CoOPood utilizes two additional projection layers, its overall params and Floating Point Operations (FLOPS) are only 2.10% and 0.02% higher than those of CLIP and CoOp, respectively. Compared with the performance improvement of OOD generalization, CoOPood has an acceptable computational overhead in terms of the number of parameters and FLOPS.

Table 5. Analysis in computational overhead among different methods. Params+ %CLIP and FLOPS+ %CoOp mean the percentage of increased params and FLOPS to CLIP and CoOp, respectively.

#Method	Params	Params+ %CLIP	FLOPS	FLOPS+ %CoOp
CoOp	2048	0.004%	354.50G	-
ERM	0.514M	1.05%	354.53G	0.01%
CoOPood	1.026M	2.10%	354.56G	0.02%

Table 6. Analysis on Compatibility and Plug-and-Play Functionality of CoOPood.

#Method	ResNet-50 (%)			
	Waterbirds		CelebA	
	Avg.	Worst	Avg.	Worst
CoCoOp	79.13	52.26	85.86	61.11
CoCoOp + CoOPood	**82.16**	**60.97**	**87.98**	**70.64**

Analysis on Compatibility. Our proposed CoOPood breaks the traditional cross-modal alignment pattern and achieves an unbiased vision-language contrastive phase. Thus, it is orthogonal to most prompt tuning mechanisms. We further conduct experiments to demonstrate the compatibility and plug-and-play functionality of CoOPood. Specifically, we integrate the concept of CoOPood into CoCoOP to measure the performance gain on different OOD datasets. The results in the Table 6 show that the unbiased alignment phase in CoOPood can provide a performance improvement of 2.1% $\sim$ 9.5%.

5. Conclusion

In this paper, we have investigated a novel decoupled prompt tuning framework under the existence of spurious features, **CoOPood**, to provide unbiased tuning phase among vision and text modality with high OOD generalization ability. Specifically, we divided the original alignment process into two independent contrastive phases by introducing two lightweight projection layers with different objectives: 1) pulling the text embedding closer to the invariant image embedding and 2) pushing the text embedding away from the spurious image embedding. Extensive experiments have been conducted over various models and datasets to verify the effectiveness and superior performance of **CoOPood**.

Impact Statement

Prompt tuning has been emerged as a promising paradigm in both computer vision and natural language processing by training a limited learnable vector with model parameter being fixed. Due to the complexity and dynamics of current open-world (e.g., spurious correlation and out-of-distribution data etc.), it is essential to investigate the effect of spurious correlation and avoid its possible negative impact on prompt tuning for future development. The proposed **CoOPood** breaks the barriers of balanced data constraint, improving the robustness of vision-language models in real-world applications. This research has the potential to enable multi-modal machine learning models to efficiently adapt to various downstream real-world tasks.

Acknowledgements

This research was supported by fundings from the Key-Area Research and Development Program of Guangdong Province (No. 2021B0101400003), Hong Kong RGC Research Impact Fund (No. R5060-19, No. R5034-18), Areas of Excellence Scheme (AoE/E-601/22-R), General Research Fund (No. 152203/20E, 152244/21E, 152169/22E, 152228/23E).

References

Adila, D., Shin, C., Cai, L., and Sala, F. Zero-shot robustification of zero-shot models. In *The Twelfth International Conference on Learning Representations*, 2024.

Ahuja, K., Caballero, E., Zhang, D., Gagnon-Audet, J.-C., Bengio, Y., Mitliagkas, I., and Rish, I. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34, 2021.

Anonymous. Learning personalized causally invariant representations for heterogeneous federated clients. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=8FHWkY0SwF.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Bossard, L., Guillaumin, M., and Van Gool, L. Food-101–mining discriminative components with random forests. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pp. 446–461. Springer, 2014.

Chen, G., Yao, W., Song, X., Li, X., Rao, Y., and Zhang, K. PLOT: Prompt learning with optimal transport for vision-language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=zqwryBoXYnh.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Creager, E., Jacobsen, J.-H., and Zemel, R. Environment inference for invariant learning. In *International Conference on Machine Learning*, pp. 2189–2200. PMLR, 2021.

Deng, Y., Yang, Y., Mirzasoleiman, B., and Gu, Q. Robust learning with progressive data expansion against spurious correlation. *Advances in Neural Information Processing Systems*, 36, 2024.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Du, Y., Wei, F., Zhang, Z., Shi, M., Gao, Y., and Li, G. Learning to prompt for open-vocabulary object detection with vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14084–14093, 2022.

Gandelsman, Y., Efros, A. A., and Steinhardt, J. Interpreting CLIP's image representation via text-based decomposition. In *The Twelfth International Conference on Learning Representations*, 2024.

Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., and Qiao, Y. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, pp. 1–15, 2023.

Garcia, N., Otani, M., Chu, C., and Nakashima, Y. Knowit vqa: Answering knowledge-based questions about videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 10826–10834, 2020.

Geirhos, R., Jacobsen, J., Michaelis, C., Zemel, R. S., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nat. Mach. Intell.*, 2(11):665–673, 2020. doi: 10.1038/S42256-020-00257-Z. URL https://doi.org/10.1038/s42256-020-00257-z.

Gu, X., Lin, T.-Y., Kuo, W., and Cui, Y. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.

Han, Z., Liang, Z., Yang, F., Liu, L., Li, L., Bian, Y., Zhao, P., Wu, B., Zhang, C., and Yao, J. Umix: Improving importance weighting for subpopulation shift via uncertainty-aware mixup. *Advances in Neural Information Processing Systems*, 35:37704–37718, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8340–8349, 2021a.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15262–15271, 2021b.

Hu, Z., Zhao, Z., Yi, X., Yao, T., Hong, L., Sun, Y., and Chi, E. Improving multi-task generalization via regularizing spurious correlation. *Advances in Neural Information Processing Systems*, 35:11450–11466, 2022.

Izmailov, P., Kirichenko, P., Gruver, N., and Wilson, A. G. On feature learning in the presence of spurious correlations. *Advances in Neural Information Processing Systems*, 35:38516–38532, 2022.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Jiang, Y. and Veitch, V. Invariant and transportable representations for anti-causal domain shifts. *Advances in Neural Information Processing Systems*, 35:20782–20794, 2022.

Kenton, J. D. M.-W. C. and Toutanova, L. K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, pp. 2, 2019.

Khattak, M. U., Rasheed, H., Maaz, M., Khan, S., and Khan, F. S. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19113–19122, 2023.

Kim, J. M., Koepke, A., Schmid, C., and Akata, Z. Exposing and mitigating spurious correlations for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2584–2594, 2023.

Kirichenko, P., Izmailov, P., and Wilson, A. G. Last layer re-training is sufficient for robustness to spurious correlations. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Zb6c8A-Fghk.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. M. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 5542–5550, 2017.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

Li, Y. and Vasconcelos, N. Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9572–9581, 2019.

Liang, C., Wang, W., Zhou, T., and Yang, Y. Visual abductive reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15565–15575, 2022.

Liang, C., Wang, W., Zhou, T., Miao, J., Luo, Y., and Yang, Y. Local-global context aware transformer for language-guided video segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Liu, J., Shen, Z., He, Y., Zhang, X., Xu, R., Yu, H., and Cui, P. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021.

Liu, S., Zhang, X., Sekhar, N., Wu, Y., Singhal, P., and Fernandez-Granda, C. Avoiding spurious correlations via logit correction. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=5BaqCFVh5qL.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Lu, Y., Liu, J., Zhang, Y., Liu, Y., and Tian, X. Prompt distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5206–5215, 2022.

Luo, Z., Wang, Y., Wang, Z., Sun, Z., and Tan, T. Disentangled federated learning for tackling attributes skew via

invariant aggregation and diversity transferring. In *International Conference on Machine Learning*, pp. 14527–14541. PMLR, 2022.

Ma, T., Sun, Y., Yang, Z., and Yang, Y. Prod: Prompting-to-disentangle domain knowledge for cross-domain few-shot image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19754–19763, 2023.

Mao, C., Xia, K., Wang, J., Wang, H., Yang, J., Bareinboim, E., and Vondrick, C. Causal transportability for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7521–7531, 2022.

McCoy, T., Pavlick, E., and Linzen, T. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL*, pp. 3428–3448, 2019. doi: 10.18653/V1/P19-1334. URL https://doi.org/10.18653/v1/p19-1334.

Ming, Y., Yin, H., and Li, Y. On the impact of spurious correlation for out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 10051–10059, 2022.

Nam, J., Cha, H., Ahn, S., Lee, J., and Shin, J. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33: 20673–20684, 2020.

Nam, J., Kim, J., Lee, J., and Shin, J. Spread spurious attribute: Improving worst-group accuracy with spurious attribute estimation. *arXiv preprint arXiv:2204.02070*, 2022.

Nilsback, M.-E. and Zisserman, A. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pp. 722–729. IEEE, 2008.

Pearl, J. *Causality*. Cambridge university press, 2009.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252, 2015.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Shen, S., Yang, S., Zhang, T., Zhai, B., Gonzalez, J. E., Keutzer, K., and Darrell, T. Multitask vision-language prompt tuning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5656–5667, 2024.

Shu, Y., Guo, X., Wu, J., Wang, X., Wang, J., and Long, M. Clipood: Generalizing clip to out-of-distributions. *arXiv preprint arXiv:2302.00864*, 2023.

Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., and Kiela, D. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15638–15650, 2022.

Singla, S., Nushi, B., Shah, S., Kamar, E., and Horvitz, E. Understanding failures of deep networks via robust feature extraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12853–12862, 2021.

Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., and Yang, H. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pp. 23318–23340. PMLR, 2022.

Wei, J., Narasimhan, H., Amid, E., Chu, W.-S., Liu, Y., and Kumar, A. Distributionally robust post-hoc classifiers under prior shifts. *arXiv preprint arXiv:2309.08825*, 2023.

Wu, Y., Gardner, M., Stenetorp, P., and Dasigi, P. Generating data to mitigate spurious correlations in natural language inference datasets. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2660–2676, 2022.

Yang, Y., Nushi, B., Palangi, H., and Mirzasoleiman, B. Mitigating spurious correlations in multi-modal models during fine-tuning. *arXiv preprint arXiv:2304.03916*, 2023.

Yao, H., Wang, Y., Li, S., Zhang, L., Liang, W., Zou, J., and Finn, C. Improving out-of-distribution robustness via selective augmentation. In *International Conference on Machine Learning*, pp. 25407–25437. PMLR, 2022.

Yao, Y., Zhang, A., Zhang, Z., Liu, Z., Chua, T.-S., and Sun, M. Cpt: Colorful prompt tuning for pre-trained vision-language models. *arXiv preprint arXiv:2109.11797*, 2021.

Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., and Beyer, L. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18123–18133, 2022.

Zhang, M., Sohoni, N. S., Zhang, H. R., Finn, C., and Ré, C. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. *arXiv preprint arXiv:2203.01517*, 2022a.

Zhang, R., Fang, R., Zhang, W., Gao, P., Li, K., Dai, J., Qiao, Y., and Li, H. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021.

Zhang, Y., Jiang, H., Miura, Y., Manning, C. D., and Langlotz, C. P. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pp. 2–25. PMLR, 2022b.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16816–16825, 2022a.

Zhou, K., Yang, J., Loy, C. C., and Liu, Z. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022b.

A. Proof of Lemma 1

In this section, we will provide the complete proofs of the Lemma 1 stated in the main content.

Lemma 1 (Conditional Independence)**.** *If the data generating mechanism of each VLM obeys the causal graph in Figure 3, we have:*

- $\boldsymbol{z}_u \perp \boldsymbol{z}_s \mid Y$*, which means that the invariant features* $\boldsymbol{z}_s$ *are conditionally independent of the spurious features* $\boldsymbol{z}_u$ *given variable* Y*.*
- $\boldsymbol{z}_s \perp \boldsymbol{t} \mid Y$*, which means that the spurious features* $\boldsymbol{z}_s$ *are conditionally independent of the text features* $\boldsymbol{t}$ *given variable* Y*.*

Proof. According to the causal Markov condition (Theorem 1.4.1) proved in (Pearl, 2009), we know that the variable $\mathbf{z}_s$ is independent of all its nondescendants, given its parents in the (Markov) causal graph. Since Y and E are the parent variables of $\mathbf{z}_s$ and $\mathbf{z}_u$ is a nondescendant of $\mathbf{z}_s$, the first causal signature in Lemma 1 is guaranteed. Moreover, based on the $d-$separation criterion in (Pearl, 2009), we can find that the variable Y $d-$separates $\mathbf{z}_s$ from $\mathbf{t}$ in the right SCM of Figure 3. Thus, we get the second causal signature in Lemma 1. □

B. Analysis on the Context Length and Initialization of Prompt

To explore whether our CoOPood works equally well on prompts with different context lengths and initialization, we repeat experiments on CLIP ResNet-50 and Waterbirds dataset by varying the context length from 4 to 16, and initializing randomly. Other experimental settings are the same as in Table 1. The results are shown in Table 7, which indicates that having more context tokens leads to a slight increase in accuracy on the worst-group. For the initialization, We find that random initialization has little effect on the final accuracy. CoOpood still maintains advanced performance on different context lengths and initialization.

Table 7. Analysis on the context length and initialization of prompt. Manual means that the prompt is initialized as "`a photo of a`"; Random-4, 8, 16 means that the prompt is initialized randomly with length of 4, 8, 16.

#Method	Manual		Random-4		Random-8		Random-16	
	Avg.	Worst	Avg.	Worst	Avg.	Worst	Avg.	Worst
CoOPood	82.38	60.28	82.56	58.93	**83.57**	62.54	83.17	**63.48**

To demonstrate the effectiveness of our proposed CoOPood with different lengths of prompt, we also conduct experiments among different methods on a randomly initialized prompt with 16 context lengths. The results are shown in Table 8. It can be seen that our CoOPood and *Group CoOPood still have the best average and worst-group accuracy.

Table 8. Analysis on the Context Length of Prompt among Different Methods. The prompt is initialized randomly with length of 16. Asterisk ($*$) in *Group DRO and *Group CoOPood denotes that using group information during training.

#Dataset	CoOp		ERM		CoOPood		*Group DRO		*Group CoOPood	
	Avg.	Worst	Avg.	Worst	Avg.	Worst	Avg.	Worst	Avg.	Worst
Waterbirds	79.72	50.92	83.16	62.04	**83.17**	**63.48**	88.75	85.91	**89.73**	**87.27**

C. Analysis on the Division of Training Set of CelebA Dataset

In section 4.2, we discussed the reasons why other methods do not perform as well as pre-trained CLIP on the CelebA dataset. We mentioned that the partitioning strategy of the training set was an important reason. Therefore, we re-divide the training set of CelebA to give a higher percentage to the worst group. The new group proportions are 19.4%, 58.3%, 16.7% and 5.6% for G_1 to G_4. Other experimental settings are the same as in Table 1. The experimental results are shown in the Table 9. We find that when the proportion of the worst group is increased, the accuracies of CoOp, ERM and CoOPood are

improved, which confirms our analysis. Meanwhile, our method outperforms ERM and CoOp regardless of the training set division, which again proves the superiority of CoOPood.

Table 9. Analysis on the division of training set of CelebA dataset. Old division means the results from Table 1. In the New division, the proportion of the worst group is increased from 1.1% to 5.6%.

#CelebA	CLIP		CoOp		ERM		CoOPood	
	Avg.	Worst	Avg.	Worst	Avg.	Worst	Avg.	Worst
New division	83.32	67.78	86.26	63.34	88.03	73.82	**88.94**	**76.17**
Old division	**83.32**	**67.78**	76.73	26.11	77.33	27.78	78.10	31.11

D. Analysis on the Few-Shot Training Set

In this section, we explore the performance of different methods using a few-shot training set. Few-shot performance verifies whether a method can work properly with extremely sparse training samples. In existing work on prompt tuning, "x-shot" usually represents there are "x" data samples per class in the training set. However, in the OOD environment, the existing division strategy may not be appropriate. Taking the Waterbirds dataset as an example, if a 16-shot training set only represents 16 waterbirds and 16 landbirds, but we do not specify their backgrounds, then we will not know the specific proportion of each group in the training set. Thus, in this section, we define "x-shot" to be the number of samples of the worst group in the training set. We conduct experiments with 16-shot Waterbird dataset, and the proportion of each group is consistent with Table 1. The value of hyper-parameters α and β are 1.0 and 2.0, other experimental settings are the same as in Table 1. The results are shown in the Table 10. Compared to Table 1, the accuracy of all methods decreases, which is consistent with the intuition that model performance will decrease under few-shot training settings. The performance of our CoOPood still outperforms all baselines, which proves that our method is well adapted to the extremely sparse training set.

Table 10. Analysis on the few-shot training set. Full means the results from Table 1, which uses all training samples.

#Waterbirds	CLIP		CoOp		ERM		CoOPood		*Group DRO		*Group CoOPood	
	Avg.	Worst	Avg.	Worst	Avg.	Worst	Avg.	Worst	Avg.	Worst	Avg.	Worst
16-shot	68.35	42.21	75.92	42.33	79.43	53.51	**79.78**	**55.44**	87.77	81.84	**88.39**	**82.53**
Full	68.35	42.21	78.98	48.91	81.50	57.17	**82.38**	**60.28**	88.93	83.33	**89.71**	**85.64**