

CROSS-MODAL MITIGATION OF SPURIOUS CORRELATION FOR PROMPT-TUNING IN VLMS WITH CAUSALLY MOTIVATED LOGIC ALIGNMENT

Anonymous authors

Paper under double-blind review

ABSTRACT

Recent studies have shown that pre-trained vision-language models can effectively adapt to diverse downstream tasks through parameter-efficient prompt tuning. Unfortunately, the tuned models can exploit spurious correlations during prediction, resulting in a failure to generalize to out-of-distribution test data, especially when the tuning dataset exhibits bias. How to achieve cross-modal mitigation of spurious correlations during prompt tuning of vision-language models remains an open question. In this paper, the challenging problem is tackled by leveraging the stable relationship between necessary and sufficient causal features and the corresponding label. On the one hand, we constrain the learning process of prompt by reinforcing the necessary and sufficient connection between the textual labels and textual features. On the other hand, the probability of necessity and sufficiency between the textual features and the filtered visual features is measured and maximized to enhance cross-modal feature alignment. By iteratively optimizing these two objectives, we can achieve cross-modal mitigation of spurious correlations because the logic equivalence between textual labels and visual features is bolstered. The theoretical analysis on generalization error indicates that our method can achieve a tighter generalization error bound than existing approaches. We evaluate the proposed method on several commonly adopted out-of-distribution datasets, and the empirical results demonstrate the superiority of our method over the state-of-the-art competitors.

1 INTRODUCTION

Vision-language models (VLMs), which integrate visual and textual data processing for complex real-world tasks (Zhou et al., 2020; Radford et al., 2021; Zhao et al., 2024; Zhang et al., 2024c), have become a cornerstone of multi-modal learning. Recent advancements have demonstrated the powerful zero-shot generalization capabilities of pre-trained vision-language models (VLMs), enabling them highly adaptable to a wide range of downstream tasks, especially image classification (Radford et al., 2021). To harness the flexible adaptability of pre-trained VLMs, prompt tuning emerges as a parameter-efficient tuning technique and has achieved significant success (Zhou et al., 2022b;a; Chen et al., 2023). Rather than fine-tuning all model parameters, prompt tuning focuses on modifying the text prompts while keeping the model’s pre-trained parameters largely intact. Optimizing the learnable prompts can enhance the alignment between textual and visual representations, thereby improving the performance of vision-language models.

It has been found that modern machine learning and data-driven models can easily rely on spurious correlations to make prediction (Geirhos et al., 2020; Ye et al., 2024). Referring to statistical associations between variables, spurious correlations arise from statistical bias and confounding factors rather than representing a true causal relationship. Consequently, spurious correlations are unstable and can vary across different data distributions. Thus, the performance of models utilizing spurious correlations can degrade dramatically on test data when a distribution shift occurs between the training/tuning data and test data, even though they demonstrate perfect performance on training/tuning data. In other words, models that employ spurious correlations exhibit poor out-of-distribution (OOD) generalization performance. A further complication is that this issue is especially preva-

054
055
056
057
058
059
060
061
062
063
064
065
066
067
068
069
070
071
072
073
074
075
076
077
078
079
080
081
082
083
084
085
086
087
088
089
090
091
092
093
094
095
096
097
098
099
100
101
102
103
104
105
106
107
lent in complex datasets where high-dimensional inputs, including image data and text data, may contain hidden biases.

057
058
059
060
061
062
063
064
065
066
067
068
069
070
071
072
073
074
075
076
077
078
079
080
081
082
083
084
085
086
087
088
089
090
091
092
093
094
095
096
097
098
099
0100
0101
0102
0103
0104
0105
0106
0107
Although considerable efforts have been made to mitigate spurious correlations in both visual modality (Arjovsky et al., 2019; Creager et al., 2021; Yang et al., 2023b; Qiu et al., 2024) and textual modality (Peyrard et al., 2022; Zhou et al., 2023), these methods are primarily designed for single-modal learning and are not applicable to multi-modal learning. In contrast to single-modal learning, the critical challenge of cross-modal mitigation of spurious correlations lies in ***how to organically integrate mitigation in visual modality, mitigation in textual modality and cross-modal alignment of representations***.

064
065
066
067
068
069
070
071
072
073
074
075
076
077
078
079
080
081
082
083
084
085
086
087
088
089
090
091
092
093
094
095
096
097
098
099
0100
0101
0102
0103
0104
0105
0106
0107
Among recent studies, the cross-modal contrastive learning framework presented in (Yang et al., 2023c) addresses the mitigation of spurious correlations in both textual and visual modalities while requiring access to text descriptions of spurious features/objects. In general scenarios, spurious features are typically latent and unobservable. Moreover, the method proposed in (Yang et al., 2023c), which is designed for fine-tuning of VLMs and alters all model parameters, cannot be applied to prompt tuning of VLMs. Besides, CoOPood (Zhang et al., 2024b) focuses on mitigating spurious correlations in visual modality during prompt-tuning of VLMs. It overlooks the spurious correlations in the textual modality. Furthermore, CoOPood relies on the assumption that the spurious correlations between spurious features and the target label are approximately subject to uniform probability distributions. Therefore, how to organically integrate mitigation in visual modality, mitigation in textual modality and cross-modal alignment of representations, without invoking unnatural assumptions, remains an open problem.

075
076
077
078
079
080
081
082
083
084
085
086
087
088
089
090
091
092
093
094
095
096
097
098
099
0100
0101
0102
0103
0104
0105
0106
0107
Inspired by the causal intervention-based calculation of the probability of necessity and sufficiency (PNS) between two variables (Tian & Pearl, 2000; Wang & Jordan, 2021; Yang et al., 2023b), we introduce the concept *logic alignment* (i.e., alignment with necessity and sufficiency) to integrate mitigation of spurious correlations and cross-modal alignment of representations organically for prompt tuning of VLMs. The key insight is that logic equivalence (i.e., necessary and sufficient) not only facilitates mitigation of spurious correlations (Wang & Jordan, 2021; Yang et al., 2023b), but also enhances dimensionality-agnostic alignment between two variables. In the context of vision-language models, the overall objective is to achieve the logic equivalence between visual causal representations (denoted by Φ_v) and textual label (denoted by Y), i.e., $Y \Leftrightarrow \Phi_v$. Considering spurious correlations can exist in both visual and textual modalities, the equivalence $Y \Leftrightarrow \Phi_v$ alone cannot guarantee that the aligned textual representations exclude spurious features. Therefore, establishing a stricter equivalence chain $Y \Leftrightarrow \Phi_t \Leftrightarrow \Phi_v$ (where Φ_t represents textual causal representations) is our final objective. Specifically, our framework can be divided into two components: 1) $Y \Leftrightarrow \Phi_t$ eliminates the spurious correlations in textual modality; 2) $\Phi_t \Leftrightarrow \Phi_v$ integrates mitigation of spurious correlations in visual modality and cross-modal alignment of representations organically when $Y \Leftrightarrow \Phi_t$ excludes spurious features in Φ_t . In practical implementation, the logic equivalence between two variables is achieved by maximizing the probability of necessity and sufficiency (PNS) between them. The main contributions of this work are summarized as follows:

- We introduce the concept *logic alignment* to address cross-modal mitigation of spurious correlations for prompt-tuning in vision-language models. Capable of integrating mitigation of spurious correlations and cross-modal alignment of representations organically, *Logic alignment* can serve as a promising technique for handling spurious correlations in various multi-modal learning scenarios.
- We design a practical framework to calculate the PNS between the textual label and textual representations, as well as the PNS between textual representations and visual representations. By maximizing these two PNS terms, the proposed objective can effectively achieve cross-modal mitigation of spurious correlations for prompt-tuning in VLMs.
- The theoretical analysis proves that our method can yield a tighter generalization error bound compared to existing approaches. Moreover, the detailed components of the derived generalization error bound verify the importance of maximizing the two proposed PNS terms from a theoretical perspective.
- The experimental results across diverse datasets demonstrate the superiority of the proposed framework in out-of-distribution generalization performance, compared with the state-of-the-art competitors.

108 **2 RELATED WORK**

110 **Causal Representation Learning** Attaining causally invariant predictors over varied data distributions
 111 is proposed in the field of causal inference Peters et al. (2016), and introduced into machine
 112 learning to tackle the OOD generalization problem by IRM Arjovsky et al. (2019). Then, many ef-
 113 forts are dedicated to facilitating the application of invariant representation learning to more general
 114 scenarios. Some works focus on achieving invariant learning when the environment label is unavail-
 115 able, e.g., EIIL Creager et al. (2021), HRM Liu et al. (2021a), KerHRM Liu et al. (2021b), ED-
 116 NIL Huang et al. (2022) and ZIN Lin et al. (2022). IFM Chen et al. (2022b) lowers the requirement
 117 on the number of available environments. Another branch Ahuja et al. (2021); Chen et al. (2022a);
 118 Huh & Baidya (2022) completes the constraints that IRM misses. Besides, iCaRL Lu et al. (2022)
 119 extends causal representation learning to non-linear causal representations while ACTIR Jiang &
 120 Veitch (2022) extends causal representation learning to anti-causal scenarios. Causal representation
 121 learning is also applied to graph representation learning Li et al. (2022); Chen et al. (2022c) and
 122 natural language modeling Peyrard et al. (2022). These methods are devised for handling spurious
 123 correlations in single-modal learning scenarios.

124 **Prompt Tuning of Vision-Language Models** The typical vision-language model, CLIP (Radford
 125 et al., 2021) is trained using a contrastive learning framework where textual and visual represen-
 126 tations are aligned by maximizing the cosine similarity between the image and text embeddings of cor-
 127 rect pairs. To fully exploit the powerful adaptation capability, prompt tuning is proposed to improve
 128 the performance of pre-trained vision-language models (e.g., CLIP) on downstream task (Zhou et al.,
 129 2022b;a). Among these attempts, CoOp (Zhou et al., 2022b) designs learnable prompts to adjust the
 130 mapping from textual label to textual representations and greatly improves the performance of pre-
 131 trained CLIP on downstream visual tasks. Furthermore, CoCoOp (Zhou et al., 2022a) introduce a
 132 image-conditional context generator to improve the zero-shot generalization performance of CoOp.
 133 Subsequently, MaPLE (Khattak et al., 2023a) adopts both textual and visual learnable prompts to
 134 enhance the alignment of textual and visual representations in downstream tasks. Another prevalent
 135 line of works utilize fine-grained learnable textual prompt to tackle the imbalance between textual
 136 and visual modalities (Chen et al., 2023; Shen et al., 2024; Li et al., 2024). All above prompt tun-
 137 ing methods do not consider the mitigation of spurious correlations in vision-language models. In
 138 particular, CoOPood (Zhang et al., 2024b) is proposed as a pioneering work focusing on mitigating
 139 spurious correlations in visual modality during prompt-tuning of VLMs. However, it overlooks the
 140 spurious correlations in the textual modality. Moreover, CoOPood relies on the assumption that the
 141 spurious correlations between spurious features and the target label are approximately subject to
 142 uniform probability distributions, which limits the applicability of CoOPood to general scenarios.

143 **3 PRELIMINARY**

145 We introduce the background knowledge about prompt tuning of VLMs and causally motivated
 146 calculation for probability of necessity and sufficiency (i.e., PNS) in this section.

147 **3.1 PROMPT TUNING OF CLIP**

149 Contrastive Language-Image Pre-training (CLIP) (Radford et al., 2021) maintains two separate en-
 150 coder: text encoder extracting textual representations from the text input and image encoder drawing
 151 visual representations from the image input. Textual and visual representations are aligned by con-
 152 ducting contrastive learning based on the language-image data pairs. For the sake of simplicity,
 153 we denote the text encoder as f and image encoder as g in CLIP. With a handcrafted prompt (e.g.,
 154 a photo of a [CLASS]) input into the frozen text encoder, the pre-trained CLIP can be deployed
 155 to downstream image classification tasks. Specifically, input images are fed to the image encoder,
 156 while the text prompt is input into the text encoder. Suppose “[CLASS]” has K categories in current
 157 downstream task, the pre-trained CLIP can make a probability prediction for input image x by

$$p(k | x) = \frac{\exp(\text{sim}(z_t^k, g(x))/\tau)}{\sum_{j=1}^K \exp(\text{sim}(z_t^j, g(x))/\tau)} \quad (1)$$

161 where z_t^j , $j \in 1, 2, \dots, K$ denotes text feature generated for class j by the text encoder f . $\text{sim}(a, b)$
 160 denotes the cosine similarity between two vector a and b while τ is the temperature parameter.

In order to improve the performance of pre-trained CLIP in downstream tasks, CoOp (Zhou et al., 2022b) introduces learnable text prompt to amend the mapping from text labels to textual representations. Suppose the learnable context is denoted as $Q = [q_1, q_2, \dots, q_N]$, the complete text input can be written as $Q_C = [q_1, q_2, \dots, q_N, \text{CLASS}]$. When the text input $Q_C = [Q, \text{CLASS}]$ is fed to the frozen text encoder, the corresponding textual feature vector for class k can be written by $z_t^k = f([Q, k])$. For each instance (x_i, y_i) in the tuning dataset $D_S := \{(x_i, y_i)\}_{i=1}^m$, the model can provide a prediction by $p(y_i | x_i) = \frac{\exp(\text{sim}(f([Q, y_i]), g(x_i))/\tau)}{\sum_{j=1}^K \exp(\text{sim}(f([Q, j]), g(x_i))/\tau)}$. The learnable text prompt is optimized by solving the following objective:

$$\min_Q \mathcal{L}_{CE-\text{logit}} := - \sum_{(x_i, y_i) \in D_S} y_i \log p(y_i | x_i). \quad (2)$$

Since only text prompt is learnable while both text and image encoder are frozen during the tuning stage, prompt tuning is a parameter-efficient tuning scheme and has gained great success.

3.2 PROBABILITY OF NECESSITY AND SUFFICIENCY (PNS)

Probability of Necessity and Sufficiency (PNS) describe the probability with which a variable is the necessary and sufficient cause of another variable. The formal definition of PNS is given as follows.

Definition 3.1 (Probability of Necessity and Sufficiency (Pearl, 2009)). *Let the specific implementations of causal variable Φ as ϕ and $\bar{\phi}$, where $\phi \neq \bar{\phi}$. The probability with which variable Φ is the necessary and sufficient cause of variable Y on test data distribution P_T is given by:*

$$\begin{aligned} PNS(Y, \Phi) := & \underbrace{P_T(Y_{do(\Phi=\phi)} = y | \Phi = \bar{\phi}, Y \neq y)}_{\text{sufficiency}} P_T(\Phi = \bar{\phi}, Y \neq y) \\ & + \underbrace{P_T(Y_{do(\Phi=\bar{\phi})} \neq y | \Phi = \phi, Y = y)}_{\text{necessity}} P_T(\Phi = \phi, Y = y), \end{aligned} \quad (3)$$

where $do(\Phi = \phi)$ (do-operator) means the manipulable variable Φ is forced to be a fixed value ϕ .

Since the probability of necessity and sufficiency is defined based on counterfactual distributions, it is usually intractable to estimate the PNS of two variables. However, with two assumptions (Exogeneity and Monotonicity) proposed and utilized in (Pearl, 2009; Yang et al., 2023b), we can obtain a useful lemma as follows. **Considering the limited length of main text, we put more detailed explanations about Exogeneity and Monotonicity assumption in Appendix C.**

Lemma 3.2 (Pearl (2009); Yang et al. (2023b)). *If variable Φ is exogenous relative to variable Y , and Y is monotonic relative to Φ , we can get*

$$PNS(Y, \Phi) = \underbrace{P_T(Y = y | \Phi = \phi)}_{\text{sufficiency}} - \underbrace{P_T(Y = y | \Phi = \bar{\phi})}_{\text{necessity}}. \quad (4)$$

3.3 PNS RISK MODELING

According to definition 3.1, PNS risk is based on the measure of ϕ and $\bar{\phi}$. As $\bar{\phi}$ represents the intervention value, it is not necessary for it to be a sample from the same distribution as the causal variable Φ . Thus, we need an auxiliary variable $\bar{\Phi} \in \mathcal{Z}$ (within the same space as variable Φ). The intervention value $\bar{\phi}$ is sampled from the distribution $P_T(\bar{\Phi} | X = x)$. To calculate the probability of necessity and sufficiency between the representations and the target in neural networks, we need to construct three networks parameterized by θ and ξ to estimate the distributions $P_T(\Phi | X = x) =$ and $P_T(\bar{\Phi} | X = x)$ by $P_T^\theta(\Phi | X = x) =$ and $P_T^\xi(\bar{\Phi} | X = x)$, respectively. Additionally, we need to build a linear classifier ω to parameterize the mapping from causal representations to target. That is, the target can be obtained by $y = \text{sign}(\omega^\top \phi)$ (Yang et al., 2023b).

Let $\mathcal{I}(A)$ be an indicator function, where $\mathcal{I}(A) = 1$ if A is true; otherwise, $\mathcal{I}(A) = 0$. PNS risk based on Definition 3.1 and Lemma 3.2 can be calculated by

$$\mathcal{R}_S(\omega, \theta, \xi) := \mathbb{E}_{(x, y) \sim D_S} [\mathbb{E}_{\phi \sim P_S(\Phi | X=x)} \mathcal{I}[\text{sign}(\omega^\top \phi) \neq y] + \mathbb{E}_{\bar{\phi} \sim P_S(\bar{\Phi} | X=x)} \mathcal{I}[\text{sign}(\omega^\top \bar{\phi}) = y]] \quad (5)$$

For practical modeling convenience, a recent study (Yang et al., 2023b) proposed an effective approximation scheme for PNS risk by deriving an upper bound of Equation 5.

216 **Proposition 3.3** (Proposition 3.1 in (Yang et al., 2023b)). *Given a source domain \mathcal{S} , we define the*
 217 *sufficient and necessary risks as:*

$$218 \quad SF_{\mathcal{S}}(\omega, \theta) := \mathbb{E}_{(x,y) \sim D_{\mathcal{S}}} \mathbb{E}_{\phi \sim P_{\mathcal{S}}^{\theta}(\Phi|X=x)} \mathcal{I}[\text{sign}(\omega^T \phi) \neq y],$$

$$220 \quad NC_{\mathcal{S}}(\omega, \xi) := \mathbb{E}_{(x,y) \sim D_{\mathcal{S}}} \mathbb{E}_{\bar{\phi} \sim P_{\mathcal{S}}^{\xi}(\bar{\Phi}|X=x)} \mathcal{I}[\text{sign}(\omega^T \bar{\phi}) = y],$$

221 *and let the Monotonicity measurement be defined as*

$$222 \quad M_{\mathcal{S}}^{\omega}(\theta, \xi) := \mathbb{E}_{(x,y) \sim D_{\mathcal{S}}} \mathbb{E}_{\phi \sim P_{\mathcal{S}}^{\theta}(\Phi|X=x)} \mathbb{E}_{\bar{\phi} \sim P_{\mathcal{S}}^{\xi}(\bar{\Phi}|X=x)} \mathcal{I}[\text{sign}(\omega^T \phi) = \text{sign}(\omega^T \bar{\phi})],$$

223 *then we have*

$$225 \quad \mathcal{R}_{\mathcal{S}}(\omega, \theta, \xi) = M_{\mathcal{S}}^{\omega}(\theta, \xi) + 2SF_{\mathcal{S}}(\omega, \theta)NC_{\mathcal{S}}(\omega, \xi) \leq M_{\mathcal{S}}^{\omega}(\theta, \xi) + 2SF_{\mathcal{S}}(\omega, \theta). \quad (6)$$

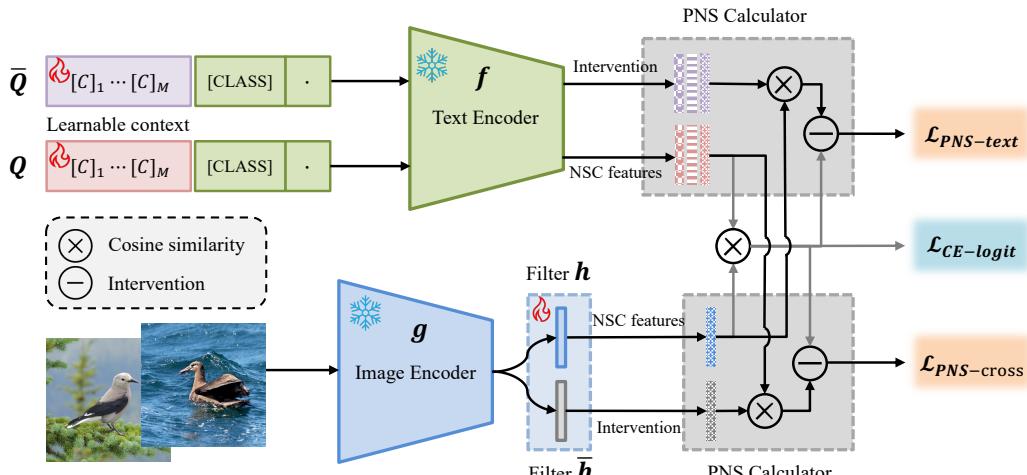
226 Based on the upper bound derived in Proposition 3.3, CaSN (Yang et al., 2023b) maximizes the PNS
 227 between variable Φ and variable Y by solving the following optimization problem:

$$228 \quad \min_{\omega, \theta} \max_{\xi} \mathcal{L}_{PNS}(\omega, \theta, \xi) := M_{\mathcal{S}}^{\omega}(\theta, \xi) + SF_{\mathcal{S}}(\omega, \theta) + \lambda \mathcal{R}_{KL}, \quad \text{subject to } \|\phi - \bar{\phi}\| \geq \delta, \quad (7)$$

230 where $\mathcal{R}_{KL} := \mathbb{E}_{D_{\mathcal{S}}} KL(P_{\mathcal{S}}^{\theta}(\Phi | X=x) \| \pi_{\Phi}) + \mathbb{E}_{D_{\mathcal{S}}} KL(P_{\mathcal{S}}^{\xi}(\bar{\Phi} | X=x) \| \pi_{\bar{\Phi}})$. $KL(\cdot, \cdot)$ denotes
 231 the KL-divergence between two probability distributions. $\pi_{\Phi} := P_{\mathcal{S}}(\Phi)$ and $\pi_{\bar{\Phi}} := P_{\mathcal{S}}(\bar{\Phi})$ describe
 232 the prior distributions of Φ and $\bar{\Phi}$, respectively.

234 4 METHODOLOGY

236 In this section, we first discuss the detailed design of the proposed framework LogicAI-PT in Section
 237 4.1 and then provide theoretical analysis on generalization error bound to demonstrate the effectiveness
 238 of the proposed method from the theoretical perspective in chapter 4.2.



257 **Figure 1: Overall framework of LogicAI-PT.** “NSC” represents “necessary and sufficient cause”.
 258 Two filters behind the image encoder are implemented using two linear layer, respectively. The
 259 NSC features in textual and visual modalities are given by $f([Q, \text{CLASS}])$ and $h(g(X))$, respec-
 260 tively. The interventions in textual and visual modalities are given by $f([\bar{Q}, \text{CLASS}])$ and $h(g(X))$,
 261 respectively. Only $f([Q, \text{CLASS}])$ and $h(g(X))$ are utilized for predicting at inference phase.

263 4.1 OVERVIEW OF LOGICAL-PT

265 In order to achieve effective cross-modal mitigation of spurious correlations for prompt-tuning in
 266 vision-language models, we design a practical framework which can be divided into two compo-
 267 nents: 1) $Y \Leftrightarrow \Phi_t$ eliminates the spurious correlations and enhances logic alignment in textual
 268 modality; 2) $\Phi_t \Leftrightarrow \Phi_v$ integrates mitigation of spurious correlations in visual modality and cross-
 269 modal alignment of representations organically when $Y \Leftrightarrow \Phi_t$ excludes spurious features in Φ_t .
 The overall framework of the proposed method LogicAI-PT is displayed in Figure 1.

Cross-modal logic alignment. As shown in objective (7), constructing the parameterized mapping ω , θ and ξ is necessary for calculating the PNS risk. When we aim at achieving cross-modal logic alignment, we need to maximize the probability of necessity and sufficiency between visual representations and textual representations. In the design framework, two filters h and \bar{h} serve as the parameterized mapping θ and ξ , respectively. Moreover, $f([Q, \text{CLASS}])$ can work as the classifier ω . Therefore, the PNS risk corresponding to cross-modal logic alignment is given by

$$\mathcal{L}_{PNS-cross} = \mathcal{L}_{PNS}(f([Q, \text{CLASS}]), h, \bar{h}). \quad (8)$$

Textual logic alignment. When we calculate textual PNS risk to achieve textual logic alignment, $f([Q, \text{CLASS}])$ and $f([\bar{Q}, \text{CLASS}])$ serve as the parameterized mapping θ and ξ , respectively. To construct the classifier ω for textual representations, we draw the prototype of each class from the visual representation space $h(g(X))$. These prototypes can serve as a classifier for the textual representations by calculating cosine similarity-based logit. In this way, the PNS risk corresponding to textual logic alignment is given by

$$\mathcal{L}_{PNS-text} = \mathcal{L}_{PNS}(h(g(X)), f([Q, \text{CLASS}]), f([\bar{Q}, \text{CLASS}])). \quad (9)$$

Overall objective. As shown in Figure 1, the cross-modal cross-entropy loss $\mathcal{L}_{CE-logit}$ is computed utilizing the cosine similarity between textual representations $f([Q, \text{CLASS}])$ and visual representations $h(g(X))$. Therefore, the overall train objective can be written as:

$$\min_{Q, h} \max_{\bar{Q}, \bar{h}} \mathcal{L}_{CE-logit} + \alpha \mathcal{L}_{PNS}(f([Q, \text{CLASS}]), h, \bar{h}) + \beta \mathcal{L}_{PNS}(h(g(X)), f([Q, \text{CLASS}]), f([\bar{Q}, \text{CLASS}])). \quad (10)$$

During the inference phase, the probability prediction for an input image is calculated by using the cosine similarity between textual and visual "NSC" features, i.e., $f([Q, \text{CLASS}])$ and $h(g(X))$.

4.2 THEORETICAL ANALYSIS

Along the information flow from visual representations Φ_v to text label Y in a vision-language model, we can evaluate the effectiveness of the visual feature extractor Φ_v in predicting the target Y using the mutual information $I(Y; \Phi_v(X))$. In practice, we can acquire the empirical estimation of $I(Y; \Phi_v(X))$ on the source dataset D_S , represented as $\hat{I}_S(Y; \Phi_v(X))$. When the learning model is ready for deployment, we prioritize the performance of Φ_v on some unknown target data distribution, denoted by $I_T(Y; \Phi_v(X))$. Since $I_T(Y; \Phi_v(X))$ is inaccessible, bounding the generalization error $I_T(Y; \Phi_v(X)) - \hat{I}_S(Y; \Phi_v(X))$ is critical for analysing the generalization performance of the proposed method in learning theory.

Before starting to the theoretical analysis on generalization error bound, we first introduce a useful assumption for the following theoretical analysis.

Assumption 4.1. In the textual modal, the textual representations Φ_t are fully informative for determining the target Y . That is, we have $Y \perp\!\!\!\perp \Phi_v | \Phi_t$.

Theorem 4.2. Suppose the source and target data distributions are denoted by $\mathbb{P}_S(X, Y)$ and $\mathbb{P}_T(X, Y)$, respectively, and the size of the source dataset D is m . Then, there exists a finite constant C such that the following inequality holds with a probability at least $1 - \delta$:

$$\begin{aligned} |I_T(Y; \Phi_v(X)) - \hat{I}_S(Y; \Phi_v(X))| &\leq \underbrace{\frac{\sqrt{C \log(|\mathcal{Y}|/\delta)} \left(|\mathcal{X}| \log(m) + |\mathcal{Y}| \log(|\mathcal{Z}|) \right) + \frac{2}{e} |\mathcal{X}|}{\sqrt{m}}}_{\text{Empirical error term}} \\ &+ \underbrace{\mathcal{J}(Y|\Phi_t) + \sqrt{C|\mathcal{Y}|\mathcal{J}(Y|\Phi_t)}}_{\text{Textual error term}} + \underbrace{\mathcal{J}(\Phi_t|\Phi_v) + \sqrt{C|\mathcal{Y}|\mathcal{J}(\Phi_t|\Phi_v)}}_{\text{Alignment error term}}, \end{aligned}$$

where $m \geq \frac{C}{4} \log(|\mathcal{Y}|/\delta) |\mathcal{X}| e^2$. The term 'Textual error term' is caused by distribution shift in textual modality while 'Alignment error term' stems from the misalignment between textual and visual modalities. $\mathcal{J}(Y|\Phi_t)$ denotes the Jeffrey's divergence defined by

$$\mathcal{J}(Y|\Phi_t) \triangleq \mathcal{KL}(\mathbb{P}_T(Y | \Phi_t) \| \mathbb{P}_S(Y | \Phi_t)) + \mathcal{KL}(\mathbb{P}_S(Y | \Phi_t) \| \mathbb{P}_T(Y | \Phi_t))$$

where $\mathcal{KL}(\cdot \| \cdot)$ denotes the Kullback–Leibler divergence between two probability distributions. Similarly, the term $\mathcal{J}(\Phi_t|\Phi_v)$ is given by

$$\mathcal{J}(\Phi_t|\Phi_v) \triangleq \mathcal{KL}(\mathbb{P}_T(\Phi_t | \Phi_v(X)) \| \mathbb{P}_S(\Phi_t | \Phi_v(X))) + \mathcal{KL}(\mathbb{P}_S(\Phi_t | \Phi_v(X)) \| \mathbb{P}_T(\Phi_t | \Phi_v(X))).$$

324 **Remark 4.3.** The first term ‘Empirical error term’ stems from limited number of data samples
 325 and will approach 0 as the size of source dataset grows towards infinity. As regard to the second
 326 term ‘Textual error term’ caused by spurious correlations in textual modality, it can be unbounded
 327 and equals to 0 if and only if $\mathbb{P}_{\mathcal{T}}(Y|\Phi_t) = \mathbb{P}_{\mathcal{S}}(Y|\Phi_t)$. When the textual representations encode
 328 spurious correlations, the second term is always strictly larger than 0. As comparison, the third
 329 term ‘Alignment error term’ is caused by the misalignment between textual and visual represen-
 330 tations. Similarly, the ‘Alignment error term’ is always non-negative and equals 0 if and only if
 331 $\mathbb{P}_{\mathcal{T}}(\Phi_t|\Phi_v) = \mathbb{P}_{\mathcal{S}}(\Phi_t|\Phi_v)$. According to the results in Theorem 4.3 in (Yang et al., 2023b), we know
 332 that optimizing the PNS risk in equation 8 can guarantee $Y \perp\!\!\!\perp Q | \Phi_t$ and optimizing the PNS risk
 333 in equation 9 can enable $\Phi_t \perp\!\!\!\perp X | \Phi_v$. Therefore, the proposed method can render both ‘Textual
 334 error term’ and ‘Alignment error term’ approach 0. In other words, our method can guarantee a
 335 tighter generalization error bound compared with the state-of-the-art prompt-tuning schemes for
 336 vision-language models. Detailed proof of Theorem 4.2 is provided in Appendix B.
 337

338 5 EXPERIMENTS

339 5.1 EXPERIMENTAL SETUP

341 **Datasets** To evaluate the performance of the proposed LogicAI-PT, we conduct experiments on
 342 four commonly used datasets: Waterbird (Sagawa et al., 2019), CelebA (Liu et al., 2015), ImageNet-
 343 1K (Russakovsky et al., 2015), and PACS Li et al. (2017). Detailed setup is explained as follows.
 344

345 Waterbirds is a commonly used benchmark dataset for studying spurious correlations. The task is to
 346 classify whether an image shows a landbird or a waterbird. The background (land and water) serve as
 347 a spurious attribute for classification of bird images. Images in Waterbird dataset can be divided into
 348 four groups: landbirds on land background (G1), landbirds on water background (G2), waterbirds
 349 on land background (G3) and waterbirds on water background (G4). The number of pictures within
 350 these four groups account for 73.0%, 3.8%, 1.2%, and 22.0% of the data, respectively. Group G3
 351 is the minority group. In the training set, landbirds appeared more often on land backgrounds,
 352 while waterbirds appeared more often on water backgrounds, so models fine-tuned on this dataset
 353 tended to rely on backgrounds rather than birds to make prediction. However, in the testing set, both
 354 landbirds and waterbirds have the same probability of appearing on a land background as on a water
 355 background, which leads to a degradation of the model’s performance.

356 Similar to Waterbirds, CelebA is a hair color prediction dataset, which also has 4 groups: non-blond
 357 females (G1), non-blond males (G2), blond females (G3) and blond males (G4) with proportions
 358 3.9%, 73.9%, 21.1%, and 1.1% of the data, respectively. Group G4 is the minority group.
 359

360 In ImageNet-1K, there are features spuriously correlated with some categories (Singla et al., 2021).
 361 For example, for Baby pacifier class, the spurious attribute is baby face. Samples without babies in
 362 the image are susceptible to being classified as water bottles rather than baby pacifier. CLIP using
 363 ResNet-50 has a 98.2% classification accuracy for samples with babies in the image, but only 36.1%
 364 for samples without babies. We use the water bottle class and the baby pacifier class in ImageNet-
 365 1K as the training set, which has three groups: water bottles (G1), baby pacifier without baby (G2),
 366 baby pacifier with baby (G3), accounting for 73.9%, 5.2%, and 20.9% of the data, respectively; the
 367 group G2 is the minority group. Note that since the validation set for ImageNet contains only 50
 368 images per class, we transferred a portion of the data from the original training set to the test set.
 369

370 PACS is a larger real-world dataset commonly used for evaluating out-of-distribution (OOD) gener-
 371 alization. It consists of 7 classes distributed across 4 domains. We adopt the “leave-one-domain-out”
 372 strategy to evaluate OOD generalization performance. For example, when evaluating performance
 373 on ‘Art Painting’ domain, the remaining three domains are used as train domains.
 374

375 **Baseline Methods.** We compare the performance of our LogicAI-PT with the state-of-the-art com-
 376 petitors, including the zero-shot CLIP (Radford et al., 2021); CoOp (Zhou et al., 2022b), a widely
 377 adopted prompt tuning method, which only minimize the contrastive loss $\mathcal{L}_{CE-logit}$; Empirical
 378 Risk Minimization (ERM), the standard technique for minimizing classification loss which also
 379 only minimize the cross-entropy loss; and CoOPood (Zhang et al., 2024b) which aligns the textual
 380 representations with the decoupled invariant representations. It is noted that, different from CoOp,
 381 under our model framework, the ERM method will use the causal projection layer (i.e., h in Fig-
 382

ure 1). Besides, we also introduce two state-of-the-art prompt tuning methods as competitors: 1) PromptSRC (Khattak et al., 2023b) which designs a self-regulating framework for prompt learning and DePT (Zhang et al., 2024a) which decouples the base-specific knowledge from feature channels into an isolated feature space during prompt tuning of VLMs.

5.2 OVERALL PERFORMANCE

Table 1: Overall performance comparison among LogicAI-PT and the state-of-the-art competitors.

Datasets	ResNet-50								ViT-B/32							
	Waterbird		CelebA		ImageNet		PACS		Waterbird		CelebA		ImageNet		PACS	
Test Acc (%)	Worst	Avg														
CLIP	43.6	70.7	67.8	84.1	36.6	68.2	80.2	91.5	41.4	65.3	69.7	85.2	51.4	75.8	81.7	93.8
CoOp	49.3	79.1	28.9	80.6	77.3	87.7	81.3	92.4	43.5	77.4	26.2	77.0	87.1	92.8	82.4	94.5
ERM	54.7	84.1	26.7	78.2	80.5	88.5	80.0	92.6	49.6	78.3	25.9	76.8	86.7	93.3	82.9	94.1
CoOPood	60.3	86.3	31.6	78.6	85.8	92.9	81.5	92.8	52.5	79.2	27.1	76.5	89.9	94.6	82.7	94.4
PromptSRC	57.2	85.5	68.2	85.3	81.6	89.4	81.7	93.6	50.8	79.5	69.3	85.9	87.8	94.1	83.4	94.8
DePT+PromptSRC	57.9	86.0	68.3	85.7	82.0	90.1	81.6	93.9	51.7	80.0	70.2	86.3	87.4	94.3	83.5	95.1
LogicAI-PT	67.5	<u>86.2</u>	69.9	87.3	90.2	95.1	82.4	<u>93.7</u>	61.2	80.3	73.1	86.9	91.8	95.4	84.3	95.2

To assess OOD generalization performance, we evaluate the test accuracy of the obtained models across a range of diverse test data distributions (4 test domains in Waterbird, CelebA, 3 test domains in ImageNet-1K dataset, and 4 test distributions in PACS). Among them, the worst-case (Worst) accuracy and average (Avg) accuracy are summarized in Table 1. Since the test data distribution is unknown in practical scenarios, both the worst-case and average accuracy are significant for reflecting the OOD generalization performance of a model. As shown in Table 1, our method LogicAI-PT outperforms the competitors on both worst-case and average test accuracy in four commonly used datasets. In particular, LogicAI-PT achieves around 7% / 9%, 2% / 3%, 4% / 2% and 1% / 1% higher worst-case accuracy than the second best algorithm on Waterbird, CelebA, ImageNet-1K and PACS when ResNet-50 / ViT-B/32 is used as backbone model, respectively.

5.3 VISUALIZATION

For the purpose of verifying that the tuned models developed by our method LogicAI-PT exploit the necessary and sufficient features rather than spurious features, we sample some data instances to generate visual explanations for the selected model using Grad-CAM (Selvaraju et al., 2017). The commonly used Grad-CAM can produce a localization map which highlights the important regions in the input image that a deep learning model depends on for predicting the label. As shown in Figure 2, the pivotal features employed by various prompt tuning methods and zero-shot CLIP for predicting WaterBird (Figure 2(a)) and BabyPacifier (Figure 2(b)) are highlighted in red.

The visualization results reveal that the proposed LogicAI-PT demonstrates three notable advantages over existing prompt-tuning methods: **1) LogicAI-PT can effectively eliminate the non-causal spurious features** that are associated with the label (i.e., ‘background’ in WaterBird dataset and ‘baby’ in ImageNet-1K dataset). **2) LogicAI-PT can mitigate the ‘sufficient but not necessary’ features** that demonstrate inconsistent presence across different data instances. For example, the shape of feet is a ‘sufficient but not necessary’ feature for classifying the picture of a bird as ‘waterbird’ or ‘landbird’ because its feet can retract or remain hidden when the bird is lying down or in flight. **3) As shown in Figure 2(a), LogicAI-PT can mitigate the ‘necessary but not sufficient’ features** which can impact the classification performance when the distribution of these ‘necessary but not sufficient’ features varies. For example, the wings of birds are ‘necessary but not sufficient’ features for distinguishing ‘waterbird’ from ‘landbird’. From the visualization results in Figure 2(a), we can find that LogicAI-PT avoids utilizing the wings to categorize the pictures of birds.

In summary, visualization results demonstrate the proposed LogicAI-PT can effectively exploit the ‘sufficient and necessary’ features and mitigate the unstable features, including non-causal spurious features, ‘sufficient but not necessary’ features and ‘necessary but not sufficient’ features. This explains why LogicAI-PT achieves superior out-of-distribution generalization performance, delivering more consistent results across diverse data distributions compared to its competitors.

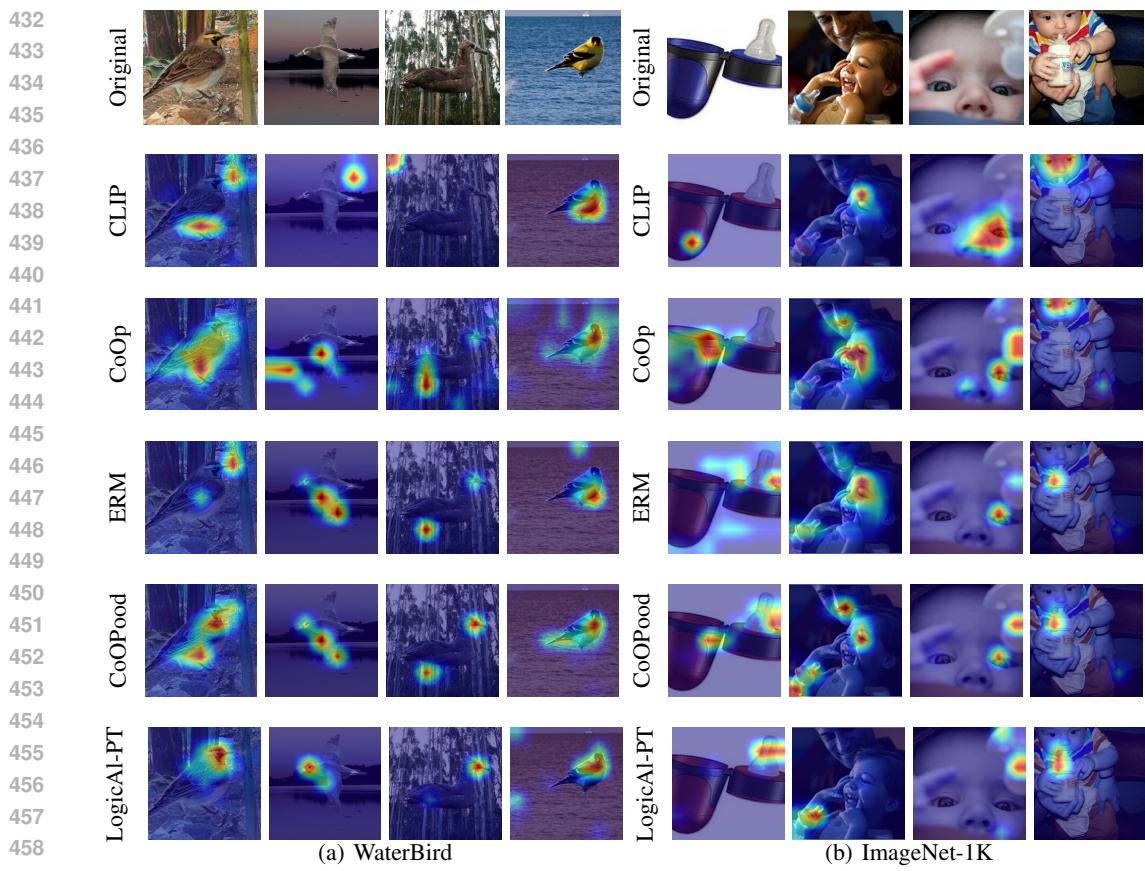


Figure 2: Visualization results of various prompt tuning approaches and zero-shot CLIP when predicting in WaterBird and ImageNet-1K datasets are generated by using Grad-CAM.

5.4 ABLATION STUDY

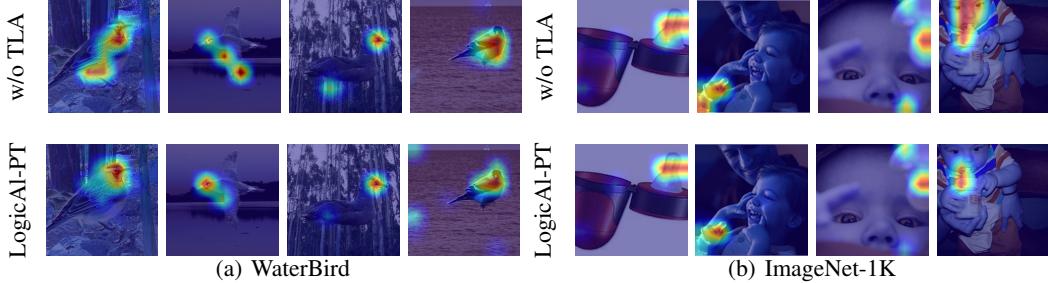
Table 2: The effect of the two separate regularization terms in the overall objective.

Datasets	Waterbird		CelebA		ImageNet-1K		PACS	
	Worst	Avg	Worst	Avg	Worst	Avg	Worst	Avg
Test Acc (%)	51.56	78.24	30.17	79.32	78.59	87.23	80.66	92.18
LogicAI-PT ($\alpha = 0$)	65.45	85.72	67.51	85.74	88.64	94.31	80.75	93.27
LogicAI-PT	67.52	86.23	69.85	87.31	90.24	95.12	82.41	93.65

Effect of Logic Alignments As discussed in Section 4.1, there are two significant regularization terms corresponding to the cross-modal logic alignment and textual logic alignment in the proposed optimization objective 10. We evaluate the isolated effects of them by independently setting $\alpha = 0$ and $\beta = 0$ in the objective 10, respectively. As displayed in Table 2, the results indicate that the cross-modal logic alignment is more important for cross-modal mitigation of spurious correlations than textual logic alignment. However, combining textual logic alignment with cross-modal logic alignment can further improve the out-of-distribution generalization performance. In this case, a natural question arises: ‘Is textual alignment necessary, and what role does it serve during prompt tuning?’ We assess the necessity of textual logic alignment in the following paragraph.

Necessity of Textual Logic Alignment Before studying the role textual logic alignment serves through the lens of visualization, we start from a qualitative analysis. Since the textual representations (corresponding to variable Φ_t) are the class-wise mapping from the text labels, the sufficiency of variable Y for variable Φ_t (i.e., $Y \Rightarrow \Phi_t$) is naturally guaranteed while the reverse $Y \Leftarrow \Phi_t$ is not

486 ensured. In other words, textual representations (Φ_t) must be necessary causes for variable Y , but
 487 they don't have to be sufficient causes for variable Y . Therefore, textual logic alignment is proposed
 488 to enhance the sufficiency of text representations (Φ_t) for label Y . Accordingly, when cross-modal
 489 logic alignment (i.e., $\Phi_t \Leftrightarrow \Phi_v$) is achieved, combining textual logic alignment can mitigate the
 490 visual features that are not sufficient for variable Y .



501 **Figure 3:** Visualization results for assessing the necessity of textual logic alignment.

502 To investigate the actual role that textual logic alignment serves, we visualize the features which is
 503 utilized by the model tuned without textual logic alignment (w/o TLA), i.e., $\beta = 0$. In particular,
 504 when we set $\beta = 0$, α is tuned to its optimal value, i.e., the cross-modal logic alignment ($\Phi_t \Leftrightarrow \Phi_v$)
 505 is enhanced. The visualization results are displayed in Figure 3. Comparing the results, we can
 506 find that adding textual logic alignment can mitigate the visual features which are not sufficient for
 507 predicting Y . For example, adopting textual logic alignment mitigates the ‘background’ feature
 508 (on 3rd picture in Figure 3(a)) and ‘wing’ feature (on 2nd picture in Figure 3(a)) which are not
 509 sufficient features for making classification in WaterBird dataset, and mitigate the ‘bottle’ feature
 510 (on 2nd picture in Figure 3(b)) and ‘baby face’ feature (on 4th picture in Figure 3(b)) that are not
 511 sufficient features for predicting ‘babypacifier’ in ImageNet dataset. Therefore, we can conclude
 512 that the visualization results support the qualitative analysis.

513 **Table 3: Performance of LogicAI-PT with different values of α and β on ImageNet-1K.**

α	0.0	1.0	10.0	20.0	30.0	50.0
worst-case (%)	78.6	80.9	86.1	90.2	88.7	79.5
average (%)	87.2	89.4	93.5	95.1	94.0	87.9
β	0.0	0.10	1.00	10.0	20.0	30.0
worst-case (%)	88.6	89.7	90.2	89.3	87.2	85.5
average (%)	94.3	94.8	95.1	94.5	93.2	91.9

523 **Sensitivity of Hyper-parameters** We evaluate the effects of two significant hyper-parameters in
 524 the proposed objective (i.e., α and β) on model performance here. Since the results on other datasets
 525 present the similar tendency as on ImageNet, we herein focus on ImageNet. When evaluating the
 526 effect of α , we fix $\beta = 1.0$. When evaluating the effect of α , we fix $\alpha = 20.0$. The results are
 527 shown in Table 3. We can find the performance of LogicAI-PT is more sensitive to the selection
 528 of α than the selection of β . To effectively mitigate spurious correlations in VLMs, careful tuning
 529 of α is essential. Regarding β , a small value is safer in practice, as a large β may compromise the
 530 discriminative capability of the extracted features.

531 6 CONCLUSION

532 This paper investigates the cross-modal mitigation of spurious correlations in prompt tuning of
 533 vision-language models. We exploit causally motivated *logic alignment* (i.e., alignment with
 534 necessity and sufficiency) to integrate mitigation of spurious correlations and cross-modal alignment
 535 of representations organically. Theoretical analysis is provided to prove that our method can yield a
 536 tighter generalization error bound than existing approaches. Experimental results across diverse
 537 datasets demonstrate the superiority of the proposed framework, termed LogicAI-PT, in out-of-
 538 distribution generalization performance, compared with the state-of-the-art competitors.

540 REFERENCES
541

- 542 Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio,
543 Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-
544 distribution generalization. *Advances in Neural Information Processing Systems*, 34:3438–3450,
545 2021.
- 546 Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization.
547 *arXiv preprint arXiv:1907.02893*, 2019.
- 548 Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. PLOT:
549 Prompt learning with optimal transport for vision-language models. In *The Eleventh International
550 Conference on Learning Representations*, 2023.
- 551 Yimeng Chen, Ruixin Xiong, Zhi-Ming Ma, and Yanyan Lan. When does group invariant learning
552 survive spurious correlations? *Advances in Neural Information Processing Systems*, 35:7038–
553 7051, 2022a.
- 554 Yining Chen, Elan Rosenfeld, Mark Sellke, Tengyu Ma, and Andrej Risteski. Iterative feature
555 matching: Toward provable domain generalization with logarithmic environments. *Advances in
556 Neural Information Processing Systems*, 35:1725–1736, 2022b.
- 557 Yongqiang Chen, Yonggang Zhang, Yatao Bian, Han Yang, MA Kaili, Binghui Xie, Tongliang Liu,
558 Bo Han, and James Cheng. Learning causally invariant representations for out-of-distribution
559 generalization on graphs. *Advances in Neural Information Processing Systems*, 35:22131–22148,
560 2022c.
- 561 Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant
562 learning. In *International Conference on Machine Learning*, pp. 2189–2200. PMLR, 2021.
- 563 Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale.
564 *arXiv preprint arXiv:2010.11929*, 2020.
- 565 Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel,
566 Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature
567 Machine Intelligence*, 2(11):665–673, 2020.
- 568 Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recog-
569 nition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp.
570 770–778, 2016.
- 571 Bo-Wei Huang, Keng-Te Liao, Chang-Sheng Kao, and Shou-De Lin. Environment diversification
572 with multi-head neural network for invariant learning. *Advances in Neural Information Processing
573 Systems*, 35:915–927, 2022.
- 574 Dongsung Huh and Avinash Baidya. The missing invariance principle found – the reciprocal twin
575 of invariant risk minimization. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and
576 A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 23023–23035.
577 Curran Associates, Inc., 2022.
- 578 Yibo Jiang and Victor Veitch. Invariant and transportable representations for anti-causal domain
579 shifts. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances
580 in Neural Information Processing Systems*, volume 35, pp. 20782–20794. Curran Associates, Inc.,
581 2022.
- 582 Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shah-
583 baz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference
584 on Computer Vision and Pattern Recognition*, pp. 19113–19122, 2023a.
- 585 Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan
586 Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without
587 forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp.
588 15190–15200, 2023b.

- 594 Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain
 595 generalization. In *Proceedings of the IEEE international conference on computer vision*, pp.
 596 5542–5550, 2017.
- 597
- 598 Haoyang Li, Ziwei Zhang, Xin Wang, and Wenwu Zhu. Learning invariant graph representations for
 599 out-of-distribution generalization. In *Advances in Neural Information Processing Systems*, 2022.
- 600 Jinshuo Liu, Haopeng Li, Sarah Monazam Erfani, Lei Feng, James Bailey, and Feng Liu. Visual-text
 601 cross alignment: Refining the similarity score in vision-language models. In *Forty-first Interna-*

602 *tional Conference on Machine Learning*, 2024.

603

604 Yong Lin, Shengyu Zhu, Lu Tan, and Peng Cui. Zin: When and how to learn invariance without
 605 environment partition? *Advances in Neural Information Processing Systems*, 35:24529–24542,
 606 2022.

607

608 Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, and Zheyuan Shen. Heterogeneous risk minimization. In
 609 *International Conference on Machine Learning*, pp. 6804–6814. PMLR, 2021a.

610

611 Jiashuo Liu, Zheyuan Hu, Peng Cui, Bo Li, and Zheyuan Shen. Integrated latent heterogeneity and
 612 invariance learning in kernel space. *Advances in Neural Information Processing Systems*, 34:
 21720–21731, 2021b.

613

614 Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild.
 615 In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

616

617 Chaochao Lu, Yuhuai Wu, José Miguel Hernández-Lobato, and Bernhard Schölkopf. Invariant
 618 causal representation learning for out-of-distribution generalization. In *International Conference*

619 *on Learning Representations*, 2022.

620 Judea Pearl. *Causality*. Cambridge university press, 2009.

621

622 Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant pre-
 623 dictions: identification and confidence intervals. *Journal of the Royal Statistical Society. Series B*

(Statistical Methodology), pp. 947–1012, 2016.

624

625 Maxime Peyrard, Sarvejeet Ghotra, Martin Josifoski, Vidhan Agarwal, Barun Patra, Dean Carignan,
 626 Emre Kiciman, Saurabh Tiwary, and Robert West. Invariant language modeling. In *Proceedings*

627 *of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5728–5743,
 Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

628

629 GuanWen Qiu, Da Kuang, and Surbhi Goel. Complexity matters: Feature learning in the presence
 630 of spurious correlations. In *Forty-first International Conference on Machine Learning*, 2024.

631

632 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal,
 633 Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual
 634 models from natural language supervision. In *International conference on machine learning*, pp.
 8748–8763. PMLR, 2021.

635

636 Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng
 637 Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual
 638 recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

639

640 Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust
 641 neural networks for group shifts: On the importance of regularization for worst-case generaliza-

642 tion. *arXiv preprint arXiv:1911.08731*, 2019.

643

644 Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh,
 645 and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based local-
 646 ization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626,
 2017.

647

648 Ohad Shamir, Sivan Sabato, and Naftali Tishby. Learning and generalization with the information
 bottleneck. *Theoretical Computer Science*, 411(29-30):2696–2711, 2010.

- 648 Sheng Shen, Shijia Yang, Tianjun Zhang, Bohan Zhai, Joseph E Gonzalez, Kurt Keutzer, and Trevor
 649 Darrell. Multitask vision-language prompt tuning. In *Proceedings of the IEEE/CVF Winter Con-*
 650 *ference on Applications of Computer Vision*, pp. 5656–5667, 2024.
- 651 Sahil Singla, Besmira Nushi, Shital Shah, Ece Kamar, and Eric Horvitz. Understanding failures
 652 of deep networks via robust feature extraction. In *Proceedings of the IEEE/CVF Conference on*
 653 *Computer Vision and Pattern Recognition*, pp. 12853–12862, 2021.
- 654 Xueyang Tang, Song Guo, Jingcai Guo, Jie Zhang, and Yue Yu. Causally motivated personalized
 655 federated invariant learning with shortcut-averse information-theoretic regularization. In *Forty-*
 656 *first International Conference on Machine Learning*, 2024.
- 657 Jin Tian and Judea Pearl. Probabilities of causation: Bounds and identification. *Annals of Mathe-*
 658 *matics and Artificial Intelligence*, 28(1):287–313, 2000.
- 659 Yixin Wang and Michael I Jordan. Desiderata for representation learning: A causal perspective.
 660 *arXiv preprint arXiv:2109.03795*, 2021.
- 661 Mengyue Yang, Xinyu Cai, Furui Liu, Weinan Zhang, and Jun Wang. Specify robust causal rep-
 662 resentation from mixed observations. In *Proceedings of the 29th ACM SIGKDD Conference on*
 663 *Knowledge Discovery and Data Mining*, pp. 2978–2987, 2023a.
- 664 Mengyue Yang, Zhen Fang, Yonggang Zhang, Yali Du, Furui Liu, Jean-Francois Ton, Jianhong
 665 Wang, and Jun Wang. Invariant learning via probability of sufficient and necessary causes. *Ad-*
 666 *vances in Neural Information Processing Systems*, 36:79832–79857, 2023b.
- 667 Yu Yang, Besmira Nushi, Hamid Palangi, and Baharan Mirzasoleiman. Mitigating spurious cor-
 668 relations in multi-modal models during fine-tuning. In *International Conference on Machine*
 669 *Learning*, pp. 39365–39379. PMLR, 2023c.
- 670 Wenqian Ye, Guangtao Zheng, Xu Cao, Yunsheng Ma, Xia Hu, and Aidong Zhang. Spurious corre-
 671 lations in machine learning: A survey. *arXiv preprint arXiv:2402.12715*, 2024.
- 672 Ji Zhang, Shihan Wu, Lianli Gao, Heng Tao Shen, and Jingkuan Song. Dept: Decoupled prompt tun-
 673 ing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*,
 674 pp. 12924–12933, 2024a.
- 675 Jie Zhang, Xiaosong Ma, Song Guo, Peng Li, Wenchao Xu, Xueyang Tang, and Zicong Hong.
 676 Amend to alignment: Decoupled prompt tuning for mitigating spurious correlation in vision-
 677 language models. In *Forty-first International Conference on Machine Learning*, 2024b.
- 678 Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks:
 679 A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024c.
- 680 Zixiang Zhao, Lilun Deng, Haowen Bai, Yukun Cui, Zhipeng Zhang, Yulun Zhang, Haotong Qin,
 681 Dongdong Chen, Jiangshe Zhang, PENG WANG, and Luc Van Gool. Image fusion via vision-
 682 language model. In *Forty-first International Conference on Machine Learning*, 2024.
- 683 Fan Zhou, Yuzhou Mao, Liu Yu, Yi Yang, and Ting Zhong. Causal-debias: Unifying debiasing in
 684 pretrained language models and fine-tuning via causal invariant learning. In *Proceedings of the*
 685 *61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,
 686 pp. 4227–4241, 2023.
- 687 Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for
 688 vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and*
 689 *pattern recognition*, pp. 16816–16825, 2022a.
- 690 Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-
 691 language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022b.
- 692 Luwei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified
 693 vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI conference*
 694 *on artificial intelligence*, volume 34, pp. 13041–13049, 2020.

A MOTIVATION FOR UTILIZING PNS

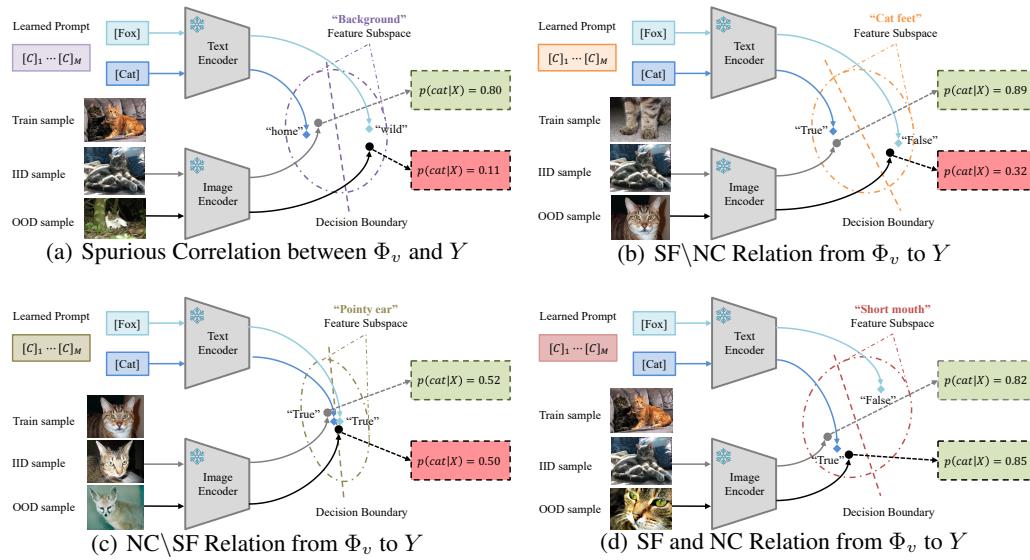


Figure 4: Illustration for three possible relations that are unstable across diverse data distributions (non-causal spurious correlation, SF\NC relation, and NC\SF relation) in vision-language models, where ‘SF\NC’ denotes ‘sufficient but not necessary’ and ‘NC\SF’ indicates ‘necessary but not sufficient’. Besides, ‘SF and NC’ means ‘sufficient and necessary’ in Figure 4(d). ‘IID’ indicates ‘in-distribution’ while ‘OOD’ means ‘out-of-distribution’. Φ_v represents the visual representation while Y indicates text label.

In these examples, the task is a binary classification problem aimed at distinguishing ‘cat’ class from ‘fox’ class. The learned prompt together with the frozen text encoder works as a projector which projects the two text labels onto a specific feature subspace. When the text labels are projected into the ‘Background’ feature subspace (as shown in Figure 4(a)), the ‘background’ feature component in visual representation space determines the prediction result because prediction is made using cosine similarity between visual features and text features. In this way, a spurious correlation between visual representation and text label is built by this learned prompt. Similarly, the learned prompt in Figure 4(b) builds a SF\NC relation from Φ_v to Y , since ‘cat feet’ is a sufficient but not necessary feature for predicting ‘cat’; the learned prompt in Figure 4(c) builds a NC\SF relation from Φ_v to Y , since ‘pointy ear’ is a necessary but not sufficient feature for predicting ‘cat’; the learned prompt in Figure 4(d) builds a SF and NC relation (i.e., logic alignment) from Φ_v to Y , since ‘short mouth’ is a sufficient and necessary feature for predicting ‘cat’.

As illustrated in Figure 4(a), 4(b) and 4(c), all these three relations (non-causal spurious correlation, SF\NC causal relation, and NC\SF causal relation) are unstable when data distribution varies. Therefore, apart from mitigation of cross-modal spurious correlations, cross-modal logic alignment (i.e., sufficiency and necessary) is also essential for enhancing the out-of-distribution generalization performance in vision-language models. This is why we utilize PNS risk in the prompt tuning of VLMs to achieve better out-of-distribution generalization performance.

B THEORETICAL PROOF: GENERALIZATION ERROR BOUND

In this paper, we denote the true data distribution of source and target datasets as p_S and p_T , respectively. In practical scenarios, the number of available data instances in a specific dataset is limited. We describe the empirical data distributions estimated from the source dataset and target dataset by \hat{p}_S and \hat{p}_T , respectively. Without loss of generality, we use notations with subscripts S and T to represent metrics on the source and target data, respectively, while notations with the overscript $\hat{\cdot}$ denote empirical estimates (e.g., the empirical distribution \hat{p} and the true distribution p).

Proposition B.1 (Lemma 11 Shamir et al. (2010)). *Let p be a distribution vector of arbitrary (possibly countably infinite) cardinality, and \hat{p} be an empirical estimation of p based on a dataset of size m . Then with a probability of at least $1 - \delta$ over the samples, the following inequality holds:*

$$\|p - \hat{p}\| \leq \frac{2 + \sqrt{2 \log(1/\delta)}}{\sqrt{m}} \quad (11)$$

Theorem 4.7. Suppose the source and target data distributions are denoted by $\mathbb{P}_S(X, Y)$ and $\mathbb{P}_T(X, Y)$, respectively, and the size of the source dataset D is m . Then, there exists a finite constant C such that the following inequality holds with a probability at least $1 - \delta$:

$$\begin{aligned} |I_T(Y; \Phi_v(X)) - \hat{I}_S(Y; \Phi_v(X))| &\leq \underbrace{\frac{\sqrt{C \log(|\mathcal{Y}|/\delta)}(|\mathcal{X}| \log(m) + |\mathcal{Y}| \log(|\mathcal{Z}|)) + \frac{2}{e}|\mathcal{X}|}{\sqrt{m}}}_{\text{Empirical error term}} \\ &+ \underbrace{\mathcal{J}(Y|\Phi_t) + \sqrt{C|\mathcal{Y}|\mathcal{J}(Y|\Phi_t)}}_{\text{Textual error term}} + \underbrace{\mathcal{J}(\Phi_t|\Phi_v) + \sqrt{C|\mathcal{Y}|\mathcal{J}(\Phi_t|\Phi_v)}}_{\text{Alignment error term}}, \end{aligned}$$

where $m \geq \frac{C}{4} \log(|\mathcal{Y}|/\delta)|\mathcal{X}|e^2$. The term ‘Textual error term’ is caused by distribution shift in textual modality while ‘Alignment error term’ stems from the misalignment between textual and visual modalities. $\mathcal{J}(Y|\Phi_t)$ denotes the Jeffrey’s divergence defined by

$$\mathcal{J}(Y|\Phi_t) \triangleq \mathcal{KL}(\mathbb{P}_T(Y|\Phi_t)\|\mathbb{P}_S(Y|\Phi_t)) + \mathcal{KL}(\mathbb{P}_S(Y|\Phi_t)\|\mathbb{P}_T(Y|\Phi_t))$$

where $\mathcal{KL}(\cdot\|\cdot)$ denotes the Kullback–Leibler divergence between two probability distributions. Similarly, $\mathcal{J}(\Phi_t|\Phi_v)$ is given be

$$\mathcal{J}(\Phi_t|\Phi_v) \triangleq \mathcal{KL}(\mathbb{P}_T(\Phi_t|\Phi_v(X))\|\mathbb{P}_S(\Phi_t|\Phi_v(X))) + \mathcal{KL}(\mathbb{P}_S(\Phi_t|\Phi_v(X))\|\mathbb{P}_T(\Phi_t|\Phi_v(X))).$$

Proof. At the beginning of the proof, we denote the mutual information between X and Y which is computed on data distribution $\hat{p}_S, \hat{p}_T, p_S$ and p_T by $\hat{I}_S(Y; X), \hat{I}_T(Y; X), I_S(Y; X)$ and $I_T(Y; X)$, respectively. We will derive the generalization error bound using the similar schemes as in (Shamir et al., 2010; Yang et al., 2023a; Tang et al., 2024).

Before starting the process of proof, we define a useful real-valued function ξ as follows:

$$\xi(x) = \begin{cases} 0, & x = 0 \\ x \log(\frac{1}{x}), & 0 < x \leq \frac{1}{e} \\ \frac{1}{e}, & x > \frac{1}{e} \end{cases} \quad (12)$$

It is noted that $\xi(x)$ is a continuous, monotonically increasing and concave real-valued function.

In general, we consider a deterministic Visual feature extractor denoted by Φ_v . To enhance conciseness in written expression, we will use Φ_v to represent $\Phi_v(X)$ in this proof without further elaboration. Thus, we can write that

$$\begin{aligned} |\hat{I}_S(Y; \Phi_v(X)) - I_T(Y; \Phi_v(X))| &\triangleq |\hat{I}_S(Y; \Phi_v) - I_T(Y; \Phi_v)| \\ &= |\hat{I}_S(Y; \Phi_v) - I_S(Y; \Phi_v) + I_S(Y; \Phi_v) - I_T(Y; \Phi_v)| \\ &\leq \underbrace{|\hat{I}_S(Y; \Phi_v) - I_S(Y; \Phi_v)|}_{\mathcal{A}_1} + \underbrace{|I_S(Y; \Phi_v) - I_T(Y; \Phi_v)|}_{\mathcal{A}_2} \end{aligned} \quad (13)$$

We know that the mutual information $I(Y; \Phi)$ is defined by:

$$I(Y; \Phi) \triangleq H(\Phi) - H(\Phi | Y) \quad (14)$$

where $H(\cdot)$ represents the Shannon information entropy. We firstly deal with the first term in the above inequality:

$$\begin{aligned} \mathcal{A}_1 &= |\hat{H}_S(\Phi_v) - H_S(\Phi_v) + H_S(\Phi_v | Y) - \hat{H}_S(\Phi_v | Y)| \\ &\leq |H_S(\Phi_v | Y) - \hat{H}_S(\Phi_v | Y)| + |\hat{H}_S(\Phi_v) - H_S(\Phi_v)| \end{aligned} \quad (15)$$

810 For the first term on the right side of Eq. 15, we can write that
 811

$$\begin{aligned}
 & |H_S(\Phi_v | Y) - \hat{H}_S(\Phi_v | Y)| \\
 &= \left| \sum_y (p_S(y)H_S(\Phi_v | y) - \hat{p}_S(y)\hat{H}_S(\Phi_v | y)) \right| \\
 &= \left| \sum_y (p_S(y)H_S(\Phi_v | y) - p_S(y)\hat{H}_S(\Phi_v | y) + p_S(y)\hat{H}_S(\Phi_v | y) - \hat{p}_S(y)\hat{H}_S(\Phi_v | y)) \right| \\
 &\leq \left| \sum_y p_S(y)(H_S(\Phi_v | y) - \hat{H}_S(\Phi_v | y)) \right| + \left| \sum_y (p_S(y) - \hat{p}_S(y))\hat{H}_S(\Phi_v | y) \right|
 \end{aligned}$$

822
 823 The first term on the right side of the above inequality can be bounded by
 824

$$\begin{aligned}
 & \left| \sum_y p_S(y)(H_S(\Phi_v | y) - \hat{H}_S(\Phi_v | y)) \right| \\
 &\leq \left| \sum_y p_S(y) \sum_{\phi_v} (p_S(\phi_v | y) \log(p_S(\phi_v | y)) - \hat{p}_S(\phi_v | y) \log(\hat{p}_S(\phi_v | y))) \right| \\
 &\leq \sum_y p_S(y) \sum_{\phi_v} \xi(|p_S(\phi_v | y) - \hat{p}_S(\phi_v | y)|) \\
 &= \sum_y p_S(y) \sum_{\phi_v} \xi\left(\left| \sum_x p_S(\phi_v | x) (p_S(x | y) - \hat{p}_S(x | y)) \right| \right) \\
 &= \sum_y p_S(y) \sum_{\phi_v} \xi\left(\left| \sum_x (p_S(\phi_v | x) - A) (p_S(x | y) - \hat{p}_S(x | y)) \right| \right) \\
 &\leq \sum_y p_S(y) \sum_{\phi_v} \xi\left(\|p_S(X | y) - \hat{p}_S(X | y)\| \|p_S(\phi_v | X) - A\| \right)
 \end{aligned}$$

841 where A can be any constant. When we set $A \triangleq \frac{1}{|X|} \sum_x p_S(\phi_v | x)$, we can get
 842

$$\left| \sum_y p_S(y)(H_S(\Phi_v | y) - \hat{H}_S(\Phi_v | y)) \right| \leq \sum_y p_S(y) \sum_{\phi_v} \xi\left(\|p_S(X | y) - \hat{p}_S(X | y)\| \cdot \sqrt{V(p_S(\phi_v | X))} \right) \quad (16)$$

843 where $\frac{1}{|X|} V(p_S(\phi_v | X))$ describes the variance of the vector $p_S(\phi_v | X)$. It is known that $\hat{H}_S(\Phi_v) \geq$
 844 $\hat{H}_S(\Phi_v | y)$ for any y , since conditioning cannot increase entropy Shamir et al. (2010). Therefore,
 845

$$\begin{aligned}
 & \left| \sum_y (p_S(y) - \hat{p}_S(y))\hat{H}_S(\Phi_v | y) \right| \leq \|p_S(Y) - \hat{p}_S(Y)\| \left| \sum_y \hat{H}_S(\Phi_v) \right| \\
 &= \|p_S(Y) - \hat{p}_S(Y)\| (|Y| \hat{H}_S(\Phi_v))
 \end{aligned} \quad (17)$$

856 Because $\Phi_v(X) \in \mathcal{Z}$, we can get that $\hat{H}_S(\Phi_v) \leq \log(|\mathcal{Z}|)$ according to the definition of Shannon
 857 Information Entropy. Combining Eq. (16) and Eq. (17), we can get
 858

$$\begin{aligned}
 & |H_S(\Phi_v | Y) - \hat{H}_S(\Phi_v | Y)| \leq \sum_y p_S(y) \sum_{\phi_v} \xi\left(\|p_S(X | y) - \hat{p}_S(X | y)\| \cdot \sqrt{V(p_S(\phi_v | X))} \right) \\
 &+ (|Y| \cdot \log(|\mathcal{Z}|)) \cdot \|p_S(Y) - \hat{p}_S(Y)\|
 \end{aligned} \quad (18)$$

864 On the other hand, we have

$$\begin{aligned}
 866 \quad |H_S(\Phi_v) - \hat{H}_S(\Phi_v)| &= \left| \sum_{\phi_v} (p_S(\phi_v) \log(p_S(\phi_v)) - \hat{p}_S(\phi_v) \log(\hat{p}_S(\phi_v))) \right| \\
 867 \\
 868 \quad &\leq \sum_{\phi_v} \xi(|p_S(\phi_v) - \hat{p}_S(\phi_v)|) \\
 869 \\
 870 \quad &= \sum_{\phi_v} \xi \left(\left| \sum_x p_S(\phi_v|x)(p_S(x) - \hat{p}_S(x)) \right| \right) \\
 871 \\
 872 \quad &= \sum_{\phi_v} \xi \left(\left| \sum_x (p_S(\phi_v|x) - A)(p_S(x) - \hat{p}_S(x)) \right| \right) \\
 873 \\
 874 \quad &\leq \sum_{\phi_v} \xi \left(\|p_S(X) - \hat{p}_S(X)\| \cdot \sqrt{V(p_S(\phi_v|X))} \right)
 \end{aligned} \tag{19}$$

875 where the constant A is chosen as $A \triangleq \frac{1}{|\mathcal{X}|} \sum_x p_S(\phi_v|x)$. Plugging Eq. (18) and Eq. (19) into
 876 Eq. (15), we can get

$$\begin{aligned}
 877 \quad \mathcal{A}_1 &\leq \sum_y p_S(y) \sum_{\phi_v} \xi \left(\|p_S(X|y) - \hat{p}_S(X|y)\| \cdot \sqrt{V(p_S(\phi_v|X))} \right) \\
 878 \\
 879 \quad &\quad + (|Y| \log(|\mathcal{Z}|)) \cdot \|p_S(Y) - \hat{p}_S(Y)\| + \sum_{\phi_v} \xi \left(\|p_S(X) - \hat{p}_S(X)\| \cdot \sqrt{V(p_S(\phi_v|X))} \right)
 \end{aligned} \tag{20}$$

880 Subsequently, we can apply the concentration bound given in Proposition B.1 to $\|p_S(X|y) - \hat{y}_S(X|y)\|$, $\|p_S(X) - \hat{p}_S(X)\|$ and $\|p_S(Y) - \hat{p}_S(Y)\|$ for any y in Eq. (20). To make sure the
 881 bounds hold simultaneously over these $|Y| + 2$ quantities, we replace δ in Eq. (11) by $\delta/(|Y| + 2)$
 882 as in the proof of Theorem 3 in Shamir et al. (2010). Hence, with a probability at least $1 - \delta$ we have
 883

$$\begin{aligned}
 884 \quad \mathcal{A}_1 &\leq 2 \sum_{\phi_v} \xi \left(\left(2 + \sqrt{2 \log((|Y| + 2)/\delta)} \right) \sqrt{\frac{V(p_S(\phi_v|X))}{m}} \right) \\
 885 \\
 886 \quad &\quad + \frac{2 + \sqrt{2 \log((|Y| + 2)/\delta)}}{\sqrt{m}} \cdot (|Y| \log(|\mathcal{Z}|))
 \end{aligned} \tag{21}$$

887 There exists a small constant C that makes the following inequality hold:

$$2 + \sqrt{2 \log((|Y| + 2)/\delta)} \leq \sqrt{C \log(|Y|/\delta)}$$

888 In addition, we know that the variance of any random variable that takes value in the range $[0, 1]$
 889 is at most $\frac{1}{4}$. Since $\frac{1}{|\mathcal{X}|} \sum_x V(p_S(\phi_v|X))$ is the variance of the distribution vector $p_S(\phi_v|X)$, we
 890 have that $V(p_S(\phi_v|X)) \leq |\mathcal{X}|/4, \forall \phi_v$.

891 Suppose that the size of training dataset (i.e., $m = |\mathcal{D}_u|$) satisfying that

$$m \geq \frac{C}{4} \log(|Y|/\delta) |\mathcal{X}| e^2 \tag{22}$$

892 Then, we can get

$$\sqrt{\frac{C \log(|Y|/\delta) V(p_S(\phi_v|X))}{m}} \leq \sqrt{\frac{C \log(|Y|/\delta) |\mathcal{X}|}{4m}} \leq \frac{1}{e}, \forall \phi_v.$$

893 We define that $\mathcal{V}(\phi_v) \triangleq C \log(|Y|/\delta) V(p_S(\phi_v|X))$, then we have that

$$\begin{aligned}
 894 \quad \sum_{\phi_v} \xi \left(\sqrt{\frac{\mathcal{V}(\phi_v)}{m}} \right) &= \sum_{\phi_v} \sqrt{\frac{\mathcal{V}(\phi_v)}{m}} \log \left(\sqrt{\frac{\mathcal{V}(\phi_v)}{m}} \right) \\
 895 \\
 896 \quad &= \sum_{\phi_v} \sqrt{\frac{\mathcal{V}(\phi_v)}{m}} \log(\sqrt{m}) + \sqrt{\frac{1}{m}} \sqrt{\mathcal{V}(\phi_v)} \log \left(\frac{1}{\sqrt{\mathcal{V}(\phi_v)}} \right) \\
 897 \\
 898 \quad &\leq \sum_{\phi_v} \left(\sqrt{\frac{\mathcal{V}(\phi_v)}{m}} \log(\sqrt{m}) + \frac{1}{\sqrt{m}e} \right)
 \end{aligned}$$

Using the results proved in the proof of Theorem 3 in Shamir et al. (2010), we can have that $\sum_{\phi_v} \sqrt{\mathcal{V}(\phi_v)} \leq \sqrt{|\mathcal{X}| |\Phi_v|}$. Therefore, we can write that

$$\sum_{\phi_v} \xi \left(\sqrt{\frac{C \log(|Y|/\delta) V(p_S(\phi_v|X))}{m}} \right) \leq \frac{\sqrt{C \log(|Y|/\delta) |X| |\Phi_v|} \log(m) + \frac{2}{e} |\Phi_v|}{2\sqrt{m}} \quad (23)$$

where $|\Phi_v|$ denote the size of the feature space from which Φ_v takes value. Recalling that Φ_v is used to represent $\Phi_v(X)$ where Φ_v itself is a deterministic feature extractor, we can conclude that $|\Phi_v| \leq |X|$. Thus, we can get

$$\begin{aligned} \mathcal{A}_1 &\leq \frac{\sqrt{C \log(|Y|/\delta) |X| \log(m) + \frac{2}{e} |X|}}{\sqrt{m}} + \frac{\sqrt{C \log(|Y|/\delta) |Y| \log(|Z|)}}{\sqrt{m}} \\ &= \frac{\sqrt{C \log(|Y|/\delta)} (|X| \log(m) + |Y| \log(|Z|)) + \frac{2}{e} |X|}{\sqrt{m}} \end{aligned} \quad (24)$$

As regard to the second term in Eq. (13), we can write that

$$\begin{aligned} \mathcal{A}_2 &= |I_T(Y; \Phi_v) - I_S(Y; \Phi_v)| \\ &= \left| \sum_y \sum_{\phi_v} p_T(y, \phi_v) \log \left(\frac{p_T(y, \phi_v)}{p_T(y)p_T(\phi_v)} \right) - p_S(y, \phi_v) \log \left(\frac{p_S(y, \phi_v)}{p_S(y)p_S(\phi_v)} \right) \right| \\ &= \left| \sum_y \sum_{\phi_v} \left(p_T(y, \phi_v) \log(p_T(y|\phi_v)) - p_S(y, \phi_v) \log(p_S(y|\phi_v)) \right) + H_T(Y) - H_S(Y) \right| \end{aligned} \quad (25)$$

As is commonly stated in the machine learning literature, the target variable Y is an exogenous variable, which indicates that $p_S(Y) = p_T(Y)$. Therefore, we have that $|H_S(Y) - H_T(Y)| = 0$. In this way, we can write that

$$\begin{aligned} \mathcal{A}_2 &\leq \left| \sum_y \sum_{\phi_v} \left(p_T(y, \phi_v) \log(p_T(y|\phi_v)) - p_S(y, \phi_v) \log(p_S(y|\phi_v)) \right) \right| \\ &= \left| \sum_y \sum_{\phi_v} \left(p_T(y, \phi_v) \log(p_T(y|\phi_v)) - p_T(y, \phi_v) \log(p_S(y|\phi_v)) + p_T(y, \phi_v) \log(p_S(y|\phi_v)) - p_S(y, \phi_v) \log(p_S(y|\phi_v)) \right) \right| \\ &\leq \left| \sum_y \sum_{\phi_v} p_T(y, \phi_v) \log \left(\frac{p_T(y|\phi_v)}{p_S(y|\phi_v)} \right) \right| + \left| \sum_y \sum_{\phi_v} (p_T(y, \phi_v) - p_S(y, \phi_v)) \log(p_S(y|\phi_v)) \right| \\ &= \mathcal{KL}(p_T(Y|\Phi_v) \| p_S(Y|\Phi_v)) + \underbrace{\left| \sum_y \sum_{\phi_v} (p_T(y, \phi_v) - p_S(y, \phi_v)) \log(p_S(y|\phi_v)) \right|}_B \end{aligned}$$

According to the above equation, we have that

$$\mathcal{B}^2 = \left\| \sum_y \sum_{\phi_v} (p_T(y, \phi_v) - p_S(y, \phi_v)) \log(p_S(y|\phi_v)) \right\|^2$$

Using the Jensen's inequality, we can get

$$\begin{aligned} \mathcal{B}^2 &\leq |Y| \sum_y \left\| \sum_{\phi_v} (p_T(y, \phi_v) - p_S(y, \phi_v)) \log(p_S(y|\phi_v)) \right\|^2 \\ &\leq |Y| \sum_y \sum_{\phi_v} p(\phi_v) \left\| (p_T(y|\phi_v) - p_S(y|\phi_v)) \log(p_S(y|\phi_v)) \right\|^2, \\ &\leq |Y| C_S^2 \sum_y \sum_{\phi_v} p(\phi_v) \|p_T(y|\phi_v) - p_S(y|\phi_v)\|^2 \end{aligned}$$

972 where C_S denotes a constant satisfying that $C_S = \max_{(\phi_v, y) \in (\Phi_v, Y)} |\log(p_S(y|\phi_v))|$. We know
 973 that $\log(\cdot)$ is a concave function, therefore we can get
 974

$$\begin{aligned} 975 \quad & \mathcal{B}^2 \leq |Y|C_S^2 \sum_y \sum_{\phi_v} p(\phi_v) \|p_T(y|\phi_v) - p_S(y|\phi_v)\| \|\log(p_T(y|\phi_v)) - \log(p_S(y|\phi_v))\| \\ 976 \quad & = |Y|C_S^2 \sum_y \sum_{\phi_v} p(\phi_v) (p_T(y|\phi_v) - p_S(y|\phi_v)) \left(\log(p_T(y|\phi_v)) - \log(p_S(y|\phi_v)) \right) \\ 977 \quad & = |Y|C_S^2 \sum_y \sum_{\phi_v} p(\phi_v) \left(p_T(y|\phi_v) \log\left(\frac{p_T(y|\phi_v)}{p_S(y|\phi_v)}\right) - p_S(y|\phi_v) \log\left(\frac{p_T(y|\phi_v)}{p_S(y|\phi_v)}\right) \right) \\ 978 \quad & = |Y|C_S^2 \left(\mathcal{KL}(p_T(Y|\Phi_v)\|p_S(Y|\Phi_v)) + \mathcal{KL}(p_S(Y|\Phi_v)\|p_T(Y|\Phi_v)) \right). \\ 979 \end{aligned}$$

980 Consequently, we can get that
 981

$$\begin{aligned} 982 \quad & \mathcal{A}_2 \leq \mathcal{KL}(p_T(Y|\Phi_v)\|p_S(Y|\Phi_v)) \\ 983 \quad & + \sqrt{|Y|C_S^2 \left(\mathcal{KL}(p_T(Y|\Phi_v)\|p_S(Y|\Phi_v)) + \mathcal{KL}(p_S(Y|\Phi_v)\|p_T(Y|\Phi_v)) \right)} \\ 984 \quad & \leq \mathcal{J}(p_T(Y|\Phi_v), p_S(Y|\Phi_v)) + \sqrt{|Y|C_S^2 \mathcal{J}(p_T(Y|\Phi_v), p_S(Y|\Phi_v))} \\ 985 \end{aligned} \quad (26)$$

986 where $\mathcal{J}(p, q)$ denotes the Jeffrey's divergence between probability p and q which is defined by
 987

$$\mathcal{J}(p_T(Y|\Phi_v), p_S(Y|\Phi_v)) \triangleq \mathcal{KL}(p_T(Y|\Phi_v)\|p_S(Y|\Phi_v)) + \mathcal{KL}(p_S(Y|\Phi_v)\|p_T(Y|\Phi_v))$$

988 With Equation (24) and Equation (26), we can conclude that
 989

$$\begin{aligned} 990 \quad & |\hat{I}_S(Y; \Phi_v(X)) - I_T(Y; \Phi_v(X))| \leq \frac{\sqrt{C \log(|Y|/\delta)} (|X| \log(m) + |Y| \log(|Z|)) + \frac{2}{e}|X|}{\sqrt{m}} \\ 991 \quad & + \mathcal{J}(p_T(Y|\Phi_v), p_S(Y|\Phi_v)) + \sqrt{|Y|C_S^2 \mathcal{J}(p_T(Y|\Phi_v), p_S(Y|\Phi_v))} \\ 992 \end{aligned} \quad (27)$$

993 When Assumption 4.1 is satisfied, we have that $Y \perp\!\!\!\perp \Phi_v \mid \Phi_t$. Thus, we can get that $p_S(Y|\Phi_v, \Phi_t) = p_S(Y|\Phi_t)$, $\forall \Phi_t, \Phi_v$ and $p_T(Y|\Phi_v, \Phi_t) = p_T(Y|\Phi_t)$, $\forall \Phi_t, \Phi_v$. In other words, we
 994 can derive that
 995

$$996 \quad p_S(Y, \Phi_v) = \sum_{\phi_t} p_S(Y, \phi_t, \Phi_v) = \sum_{\phi_t} p_S(Y|\phi_t, \Phi_v) p_S(\phi_t, \Phi_v) = \sum_{\phi_t} p_S(Y|\phi_t) p_S(\phi_t, \Phi_v).$$

997 That is, $p_S(Y|\Phi_v) = \sum_{\phi_t} p_S(Y|\phi_t) p_S(\phi_t|\Phi_v)$. Similarly, the probability distribution $p_T(Y|\Phi_v)$ can be rewrite as $p_T(Y|\Phi_v) = \sum_{\phi_t} p_T(Y|\phi_t) p_T(\phi_t|\Phi_v)$. Plugging these two equations
 998 into $\mathcal{KL}(p_S(Y|\Phi_v)\|p_T(Y|\Phi_v))$, we can obtain that
 999

$$\begin{aligned} 1000 \quad & \mathcal{KL}(p_S(Y|\Phi_v)\|p_T(Y|\Phi_v)) = \sum_y \sum_{\phi_v} p_S(y, \phi_v) \log\left(\frac{p_S(y|\phi_v)}{p_T(y|\phi_v)}\right) \\ 1001 \quad & = \sum_y \sum_{\phi_v} p(\phi_v) \sum_{\phi_t} p_S(y|\phi_t) p_S(\phi_t|\phi_v) \log\left(\frac{\sum_{\phi_t} p_S(y|\phi_t) p_S(\phi_t|\phi_v)}{\sum_{\phi_t} p_T(y|\phi_t) p_T(\phi_t|\phi_v)}\right). \\ 1002 \end{aligned}$$

1003 Here we consider a real-valued function $\zeta(x) = x \log(x)$ which is a convex function. Then, we can
 1004 write that
 1005

$$\begin{aligned} 1006 \quad & \mathcal{KL}(p_S(Y|\Phi_v)\|p_T(Y|\Phi_v)) = \sum_y \sum_{\phi_v} p(\phi_v) \sum_{\phi_t} p_T(y|\phi_t) p_T(\phi_t|\phi_v) \zeta\left(\frac{\sum_{\phi_t} p_S(y|\phi_t) p_S(\phi_t|\phi_v)}{\sum_{\phi_t} p_T(y|\phi_t) p_T(\phi_t|\phi_v)}\right) \\ 1007 \quad & = \sum_y \sum_{\phi_v} p(\phi_v) \sum_{\phi_t} p_T(y|\phi_t) p_T(\phi_t|\phi_v) \zeta\left(\frac{p_T(y|\phi_t) p_T(\phi_t|\phi_v)}{\sum_{\phi_t} p_T(y|\phi_t) p_T(\phi_t|\phi_v)} \cdot \frac{p_S(y|\phi_t) p_S(\phi_t|\phi_v)}{p_T(y|\phi_t) p_T(\phi_t|\phi_v)}\right) \\ 1008 \quad & \leq \sum_y \sum_{\phi_v} p(\phi_v) \sum_{\phi_t} p_T(y|\phi_t) p_T(\phi_t|\phi_v) \sum_{\phi_t} \frac{p_T(y|\phi_t) p_T(\phi_t|\phi_v)}{\sum_{\phi_t} p_T(y|\phi_t) p_T(\phi_t|\phi_v)} \zeta\left(\frac{p_S(y|\phi_t) p_S(\phi_t|\phi_v)}{p_T(y|\phi_t) p_T(\phi_t|\phi_v)}\right) \\ 1009 \quad & = \sum_y \sum_{\phi_v} p(\phi_v) \sum_{\phi_t} p_T(y|\phi_t) p_T(\phi_t|\phi_v) \zeta\left(\frac{p_S(y|\phi_t) p_S(\phi_t|\phi_v)}{p_T(y|\phi_t) p_T(\phi_t|\phi_v)}\right) \\ 1010 \end{aligned}$$

1026 According to the definition of $\zeta(x) = x \log(x)$, we can get that
 1027

$$\begin{aligned}
 & \mathcal{KL}(p_{\mathcal{S}}(Y | \Phi_v) \| p_{\mathcal{T}}(Y | \Phi_v)) \\
 & \leq \sum_y \sum_{\phi_v} p(\phi_v) \sum_{\phi_t} p_{\mathcal{S}}(y | \phi_t) p_{\mathcal{S}}(\phi_t | \phi_v) \log \left(\frac{p_{\mathcal{S}}(y | \phi_t) p_{\mathcal{S}}(\phi_t | \phi_v)}{p_{\mathcal{T}}(y | \phi_t) p_{\mathcal{T}}(\phi_t | \phi_v)} \right) \\
 & = \sum_y \sum_{\phi_v} p(\phi_v) \sum_{\phi_t} p_{\mathcal{S}}(y | \phi_t) p_{\mathcal{S}}(\phi_t | \phi_v) \log \left(\frac{p_{\mathcal{S}}(y | \phi_t)}{p_{\mathcal{T}}(y | \phi_t)} \right) \\
 & \quad + \sum_y \sum_{\phi_v} p(\phi_v) \sum_{\phi_t} p_{\mathcal{S}}(y | \phi_t) p_{\mathcal{S}}(\phi_t | \phi_v) \log \left(\frac{p_{\mathcal{S}}(\phi_t | \phi_v)}{p_{\mathcal{T}}(\phi_t | \phi_v)} \right) \\
 & = \sum_y \sum_{\phi_v} \sum_{\phi_t} p_{\mathcal{S}}(y, \phi_t, \phi_v) \log \left(\frac{p_{\mathcal{S}}(y | \phi_t)}{p_{\mathcal{T}}(y | \phi_t)} \right) \\
 & \quad + \sum_y \sum_{\phi_v} \sum_{\phi_t} p_{\mathcal{S}}(y, \phi_t, \phi_v) \log \left(\frac{p_{\mathcal{S}}(\phi_t | \phi_v)}{p_{\mathcal{T}}(\phi_t | \phi_v)} \right) \\
 & = \sum_y \sum_{\phi_t} p_{\mathcal{S}}(y, \phi_t) \log \left(\frac{p_{\mathcal{S}}(y | \phi_t)}{p_{\mathcal{T}}(y | \phi_t)} \right) + \sum_{\phi_t} \sum_{\phi_v} p_{\mathcal{S}}(\phi_t, \phi_v) \log \left(\frac{p_{\mathcal{S}}(\phi_t | \phi_v)}{p_{\mathcal{T}}(\phi_t | \phi_v)} \right) \\
 & = \mathcal{KL}(p_{\mathcal{S}}(Y | \Phi_t) \| p_{\mathcal{T}}(Y | \Phi_t)) + \mathcal{KL}(p_{\mathcal{S}}(\Phi_t | \Phi_v) \| p_{\mathcal{T}}(\Phi_t | \Phi_v))
 \end{aligned}$$

1048 Similarly, we can also derive that
 1049

$$\mathcal{KL}(p_{\mathcal{T}}(Y | \Phi_v) \| p_{\mathcal{S}}(Y | \Phi_v)) \leq \mathcal{KL}(p_{\mathcal{T}}(Y | \Phi_t) \| p_{\mathcal{S}}(Y | \Phi_t)) + \mathcal{KL}(p_{\mathcal{T}}(\Phi_t | \Phi_v) \| p_{\mathcal{S}}(\Phi_t | \Phi_v))$$

1050 Therefore, we conclude that
 1051

$$\mathcal{J}(p_{\mathcal{T}}(Y | \Phi_v) \| p_{\mathcal{S}}(Y | \Phi_v)) \leq \mathcal{J}(p_{\mathcal{T}}(Y | \Phi_t) \| p_{\mathcal{S}}(Y | \Phi_t)) + \mathcal{J}(p_{\mathcal{T}}(\Phi_t | \Phi_v) \| p_{\mathcal{S}}(\Phi_t | \Phi_v)).$$

1052 Plugging this inequality into inequality 27, we can finally get
 1053

$$\begin{aligned}
 |I_{\mathcal{T}}(Y; \Phi_v(X)) - \hat{I}_{\mathcal{S}}(Y; \Phi_v(X))| & \leq \frac{\sqrt{C \log(|\mathcal{Y}|/\delta)} \left(|\mathcal{X}| \log(m) + |\mathcal{Y}| \log(|\mathcal{Z}|) \right) + \frac{2}{e} |\mathcal{X}|}{\sqrt{m}} \\
 & \quad + \mathcal{J}(Y | \Phi_t) + \sqrt{C |\mathcal{Y}| \mathcal{J}(Y | \Phi_t)} + \mathcal{J}(\Phi_t | \Phi_v) + \sqrt{C |\mathcal{Y}| \mathcal{J}(\Phi_t | \Phi_v)},
 \end{aligned}$$

1054 Thus, we complete the proof of Theorem 4.2. \square
 1055

C MORE DETAILS ABOUT PNS AND PNS MODELING

1056 Probability of Necessity and Sufficiency (PNS) describe the probability with which a variable is the
 1057 necessary and sufficient cause of another variable. The formal definition of PNS is given as follows.
 1058

1059 **Definition C.1** (Probability of Necessity and Sufficiency (Pearl, 2009)). *Let the specific implementations of causal variable Φ as ϕ and $\bar{\phi}$, where $\phi \neq \bar{\phi}$. The probability with which variable Φ is the necessary and sufficient cause of variable Y on test data distribution $P_{\mathcal{T}}$ is given by:*

$$\begin{aligned}
 PNS(Y, \Phi) := & \underbrace{P_{\mathcal{T}}(Y_{do(\Phi=\phi)} = y | \Phi = \bar{\phi}, Y \neq y) P_{\mathcal{T}}(\Phi = \bar{\phi}, Y \neq y)}_{\text{sufficiency}} \\
 & + \underbrace{P_{\mathcal{T}}(Y_{do(\Phi=\bar{\phi})} \neq y | \Phi = \phi, Y = y) P_{\mathcal{T}}(\Phi = \phi, Y = y)}, \\
 & \quad \text{necessity}
 \end{aligned} \tag{28}$$

1060 where $do(\Phi = \phi)$ (*do-operator*) indicates that the manipulable variable Φ is forced to be a fixed
 1061 value $\Phi = \phi$.
 1062

1063 Since the probability of necessity and sufficiency is defined based on counterfactual distributions, it
 1064 is usually intractable to estimate the PNS of two variables. Therefore, we need some assumptions to
 1065 facilitate the practical calculation of PNS.
 1066

Assumption C.2 (Exogeneity (Pearl, 2009; Yang et al., 2023b)). *Variable Φ is exogenous relative to variable Y with respect to the source domain S and target domain T , if the intervention probability is identified by conditional probability, i.e., $P_S(Y_{do(\Phi=\phi)} = y) = P_S(Y = y \mid \Phi = \phi)$ and $P_T(Y_{do(\Phi=\phi)} = y) = P_T(Y = y \mid \Phi = \phi)$.*

Assumption C.3 (Monotonicity (Pearl, 2009; Yang et al., 2023b)). *Variable Y is monotonic relative to variable Φ if and only if either $P(Y_{do(\Phi=\phi)} = y, Y_{do(\Phi=\bar{\phi})} \neq y) = 0$ or $P(Y_{do(\Phi=\phi)} \neq y, Y_{do(\Phi=\bar{\phi})} = y) = 0$ holds.*

Exogeneity defined in Assumption C.2 bridges the gap between the intractable intervention probability and the computable conditional probability, while monotonicity defined in Assumption C.3 guarantees that the causal variable Φ has monotonic effect on variable Y . With these two assumptions, we can obtain a useful lemma as follows.

Lemma C.4 (Pearl (2009); Yang et al. (2023b)). *If variable Φ is exogenous relative to variable Y , and Y is monotonic relative to Φ , we can get*

$$PNS(Y, \Phi) = \underbrace{P_T(Y = y \mid \Phi = \phi)}_{\text{sufficiency}} - \underbrace{P_T(Y = y \mid \Phi = \bar{\phi})}_{\text{necessity}}. \quad (29)$$

D MORE EXPERIMENTAL RESULTS

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, 2020) as the backbone models. The prompt used in all methods has 8 learnable tokens and initialized as the default one “a photo of”. 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 LogicAI-PT, unless otherwise specified, the value of hyper-parameters α and β are 10.0 and 1.0 for CelebA; 20.0, 1.0 for ImageNet-1K; 3.0 and 2.0 for WaterBird.

Computational Efficiency We analyze and compare the computational overhead of our method with several existing methods to verify the computational efficiency of the proposed LogicAI-PT. The results are presented in the following table.

Table 4: Evaluation results on computational overhead of our method LogicAI-PT and the state-of-the-art competitors. ‘Params’ denotes the number of learnable parameters while ‘FLOPS’ represents ‘Floating Point Operations’.

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%
LogicAI-PT	1.028M	2.10%	354.57G	0.02%

We can see the the overall parameters and Floating Point Operations (FLOPS) of our LogicAI-PT are only 2.1% and 0.02% higher than those of CLIP and CoOp, respectively. Compared with the improvement in out-of-distribution generalization performance, our method LogicAI-PT impressive computational efficiency in terms of the number of parameters and FLOPS.

Adaptation from Causal Representation Learning To enhance the motivation for utilizing PNS modeling to improve out-of-distribution generalization during prompt tuning of VLMs, we adapt two representative causal representation learning methods from invariant learning: IRM (Arjovsky et al., 2019) and IB-IRM (Ahuja et al., 2021). They mitigate spurious correlations by ensuring the invariance of the conditional probability of the label Y given the causal representation across varied training environments. The evaluation is conducted using ResNet-50 as the backbone model. The experimental results on four datasets are list as follows:

We can find that LogicAI-PT outperforms the typical causally invariant representation learning methods. The underlying reason stems from the advantage of ‘sufficient and neccesary’ causal representa-

1134

1135 Table 5: Performance comparison among our method LogicAI-PT and the prevalent schemes
1136 adapted from two single-modal causal presentation learning methods.

Datasets	Waterbird		CelebA		ImageNet-1K		PACS		
	Test Acc (%)	Worst	Avg	Worst	Avg	Worst	Avg	Worst	Avg
ERM		54.7	84.1	26.7	78.2	80.5	88.5	80.0	92.6
IRM		64.7	83.9	67.1	86.2	87.9	93.6	80.7	93.8
IB-IRM		65.3	84.3	67.9	85.8	88.3	93.9	81.2	93.4
LogicAI-PT		67.5	86.2	69.9	87.3	90.2	95.1	82.4	93.7

1143

1144

1145

tion over traditional causal representation, which also forms the motivation for proposing LogicAI-PT for prompt tuning of VLMs. Prevalent causal representation learning methods primarily aim to mitigate non-causal spurious correlations. In contrast, the concept of ‘sufficiency and necessity’ goes further by excluding not only non-causal spurious correlations but also causal relationships that are ‘sufficient but not necessary’ or ‘necessary but not sufficient’. We provide specific examples to clarify these types of relationships and explain why only ‘sufficient and necessary’ relations remain stable across diverse data distributions in Figure 4.

1152

1153

1154

1155

1156

1157

1158

1159

1160

1161

1162

1163

1164

1165

1166

1167

1168

1169

1170

1171

1172

1173

1174

1175

1176

1177

1178

1179

1180

1181

1182

1183

1184

1185

1186

1187