CoDA: Co-Adaptive Dual-Path Alignment for Vision-Language Models

Yi Zhang[1], **Rui Zhu**[2*], **Channi Li**[1], **Xiaoxu Li**[1*], **Zhanyu Ma**[3,4] and **Jing-Hao Xue**[5]
[1]Lanzhou University of Technology
[2]City St George's, University of London
[3]Beijing University of Posts and Telecommunications
[4]Beijing Key Laboratory of Multimodal Data Intelligent Perception and Governance
[5]University College London
yizhang.cv.ac@gmail.com, rui.zhu@city.ac.uk, ChanniLi0404@outlook.com, lixiaoxu@lut.edu.cn, mazhanyu@bupt.edu.cn, jinghao.xue@ucl.ac.uk

Abstract

Parameter-efficient adaptation methods, such as adapters and prompt learning, have become popular for transferring pre-trained vision-language models (VLMs) to downstream tasks. However, under limited supervision during adaptation, overly aggressive cross-modal alignment can distort the intrinsic structure of modality-specific representations, leading to degraded generalization. In this paper, we propose Co-adaptive Dual-path Alignment (CoDA), a new adaptation framework that explicitly disentangles and coordinates cross-modal semantic alignment and intra-modal structural consistency. In CoDA, we also propose a parent class labeling strategy that injects hierarchical semantic priors into textual prompts, further stabilizing alignment under limited supervision. Extensive experiments on eleven benchmark datasets show that CoDA outperforms state-of-the-art parameter-efficient methods, particularly under few-shot learning and distribution-shift scenarios. Our code is available at https://github.com/yizhang-ac/CoDA.

1 Introduction

Vision-language models (VLMs), modeling visual and textual information jointly, have enabled unified vision-language understanding. By contrastively training on large-scale image-text pairs, VLMs like CLIP [Radford *et al.*, 2021] learn a shared embedding space, where semantically related image-text pairs are pulled closer while unrelated ones are pushed apart. Trained on over 400 million image-text pairs, CLIP shows remarkable zero-shot generalization across diverse downstream tasks.

Despite their strong generalization ability, the massive amount of parameters of VLMs makes full fine-tuning computationally expensive and prone to overfitting, especially in low-data regimes such as few-shot learning. To address this issue, adaptation strategies that freeze the backbone and introduce a small set of task-specific parameters, including adapter-based methods [Gao *et al.*, 2024; Yu *et al.*, 2023; Li *et al.*, 2023] and prompt-learning methods [Zhou *et al.*, 2022a; Khattak *et al.*, 2023a; Ouali *et al.*, 2023], have become the dominant choice under limited supervision.

During adaptation, the importance of explicitly modeling cross-modal alignment has been stressed by recent studies. For shared representation of modalities, representative works like MMA [Yang *et al.*, 2024] introduce a multi-modal adapter that shares weights between the image and text branches. However, under limited supervision during adaptation, overly aggressive cross-modal alignment can distort the intrinsic structure of intra-modal representations, leading to degraded generalization.

We argue that robust VLM adaptation requires *coordinately* addressing two coupled yet distinct objectives: (i) *cross-modal alignment*, which governs how semantic information should be adequately exchanged between modalities, and (ii) *intra-modal consistency*, which governs how the intra-modal structure learned from large-scale pre-training should be preserved. Neglecting either objective may result in over-coupled representations or structurally degenerate features, particularly in few-shot and distribution-shift settings.

To this end, we propose *Co-adaptive Dual-path Alignment (CoDA)*, a new adaptation framework that explicitly disentangles and coordinates cross-modal semantic alignment and intra-modal structural consistency. Rather than collapsing the two objectives into a single one during adaptation, CoDA simultaneously promotes cross-modal semantic interaction and calibrates intra-modal feature structure.

Specifically, we first introduce a channel-wise module, *Decoupled Squeeze-and-Excitation (dcSE)*, that reduces modality discrepancy prior to our proposed dual-path alignment. dcSE aggregates token features into global channel statistics and projects them to a shared semantic space, enabling cross-modal interaction at the level of semantic responses rather than explicit token or instance alignment. This design avoids fragile one-to-one token-level alignments under ambiguity and domain shift.

Then building on dcSE, we propose a dual-path module, *Distribution-Structure Co-adaptation (DiSCo)*, to refine intermediate representations. Concretely, a β-path computes a distribution-aware cross-modal bias (an additive semantic offset) via entropy-regularized optimal transport, enabling soft cross-modal semantic emphasis without enforcing hard

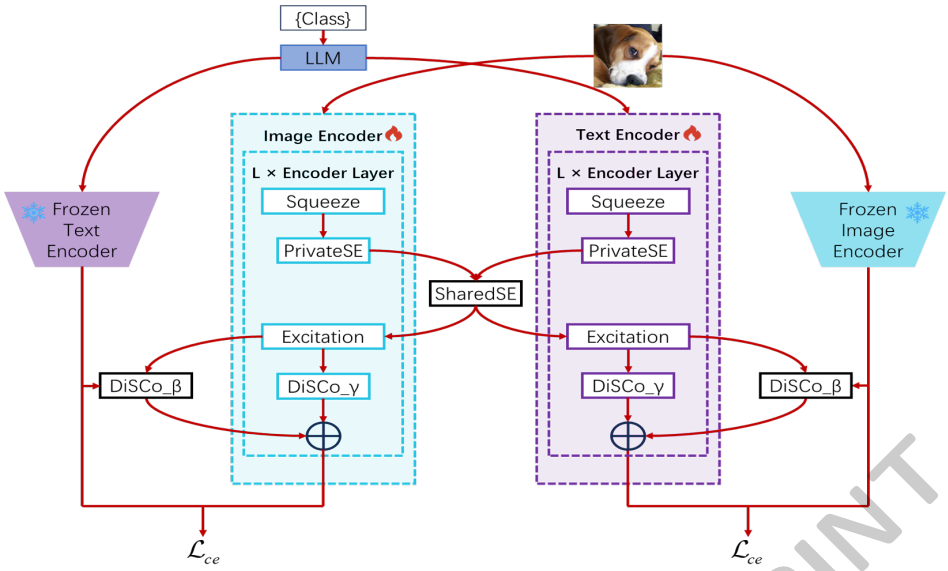

Figure 1: An overview of CoDA. Given an image-text pair, we insert dcSE and DiSCo into the frozen CLIP encoders to coordinate cross-modal semantic alignment and intra-modal structural consistency during adaptation. dcSE performs channel-wise squeeze-excitation in a shared semantic subspace to reduce modality discrepancy via distributed channel modulation. DiSCo then refines intermediate features with two complementary signals: the β-path computes an OT-based soft semantic emphasis over tokens from the other modality and injects it as an additive bias, while the γ-path extracts intra-modal principal components and applies channel-wise multiplicative calibration to preserve intra-modal structure. PCG augments text prompts with hierarchical priors for stable alignment under limited supervision.

token alignment. In parallel, a γ-path summarizes dominant intra-modal structure using principal component analysis and converts it into channel-wise calibration signals, preserving intra-modal structure during adaptation.

Finally, inspired by hierarchical human perception, we introduce a *Parent Class Generation (PCG)* strategy that augments labels with LLM-generated parent concepts. This hierarchical prompt design injects coarse-to-fine semantic priors into textual representations and improves alignment stability under limited supervision.

Our contributions are summarized as follows:

- We propose CoDA, a new vision-language adaptation framework achieving both cross-modal semantic alignment and intra-modal structural consistency.
- We introduce dcSE and DiSCo, two complementary modules that operate at channel and feature levels, respectively, enabling controlled semantic exchange and structure-aware representation refinement.
- We further propose a PCG mechanism that injects hierarchical semantic priors via LLMs, improving robustness under few-shot and distribution-shift scenarios.

2 Related Work

2.1 Vision-Language Models

Large-scale VLMs pre-trained on web-scale image-text pairs have become a cornerstone for multi-modal understanding. A representative example is CLIP [Radford *et al.*, 2021], which learns a shared image-text embedding space via contrastive learning, exhibiting strong zero-shot generalization across diverse downstream tasks. Building upon CLIP, subsequent models such as ALIGN [Jia *et al.*, 2021], FILIP [Yao *et al.*, 2021], and SigLIP [Zhai *et al.*, 2023] further improve vision-language representations by scaling training data and model capacity, refining contrastive objectives, or enhancing token-level interactions. Despite these advances, fine-tuning large VLMs to downstream tasks remains challenging due to their scale and sensitivity to limited supervision, motivating parameter-efficient adaptation strategies.

2.2 Parameter-efficient Adaptation of VLMs

Parameter-efficient adaptation techniques originally developed in NLP have been extended to VLMs [Zhou *et al.*, 2022a; Khattak *et al.*, 2023a; Gao *et al.*, 2024]. Prompt-learning methods adapt VLMs by optimizing learnable context tokens in the text encoder. CoOp [Zhou *et al.*, 2022b] replaces handcrafted prompts with continuous learnable vectors, while CoCoOp [Zhou *et al.*, 2022a] further conditions prompts on image features to improve generalization. Subsequent works explore multi-modal prompts [Khattak *et al.*, 2023a] and regularization strategies to mitigate overfitting under limited supervision. Adapter-based methods instead introduce lightweight trainable modules into frozen backbones. CLIP-Adapter [Gao *et al.*, 2024] and Tip-Adapter [Zhang *et al.*, 2021] focus on refining visual representations, whereas more recent approaches inject adapters into both visual and textual branches to enable multi-modal adaptation [Jiang *et al.*, 2022; Yang *et al.*, 2024]. However, most of these methods focus merely on final embedding alignment, without explicitly regulating cross-modal distribution discrepancies or intra-modal representation structure.

2.3 Modality Alignment for Adaptation of VLMs

Recent studies have highlighted the importance of explicitly modeling modality alignment during VLM adaptation. Representative methods such as MMA [Yang *et al.*, 2024] enforce mutual alignment between visual and textual representations by encouraging consistency across modalities during adaptation. Multi-modal prompt learning approaches [Khattak *et al.*, 2023a] also improve alignment by jointly optimizing prompts in both visual and textual branches. Other works implicitly promote alignment through instance-level conditioning or shared embedding supervision. Despite their effectiveness, existing alignment-based methods predominantly assume that the internal structure of each modality remains stable during adaptation. Such assumptions may be fragile under few-shot or distribution-shifted settings, where aggressive cross-modal coupling can distort intra-modal representations. In contrast, our method disentangles cross-modal alignment from intra-modal structural calibration and coordinates them, enabling controlled semantic interaction while preserving the intrinsic intra-modal structure during adaptation.

3 Methodology

As illustrated in Figure 1, we propose CoDA, a new unified vision-language adaptation framework that explicitly coordinates both cross-modal semantic alignment and intra-modal structural consistency. Coordinating these two objectives is critical for robust adaptation under limited supervision.

3.1 Decoupled Squeeze-and-Excitation (dcSE)

In multi-modal adaptation, naïve feature fusion or tightly coupled attention often leads to over-coupling between modalities, where modality-specific representations are suppressed. While conventional Squeeze-and-Excitation (SE) layers [Hu *et al.*, 2018] effectively recalibrate channel responses within a single modality, they neither support semantic exchange across modalities nor preserve intra-modal structure during adaptation.

To address these limitations, we propose *Decoupled Cross-modal Squeeze-and-Excitation (dcSE)*, a channel-wise module that enables controlled semantic exchange between visual and textual representations while avoiding aggressive cross-modal alignment. As illustrated in Fig. 1, dcSE consists of two stages. First, a decoupled squeeze that summarizes intra-modal channel statistics and projects them into a shared feature space, making their representations more comparable. Then, a decoupled excitation that transforms this shared representation back into modality-specific channel responses.

Let $X^{(l)} \in \mathbb{R}^{B\times T\times C}$ denote the intermediate features from the l-th Transformer block, where B is the batch size, T is the number of tokens, and C is the channel dimension. dcSE first aggregates token-level features into a global channel descriptor by average pooling over the token dimension:

$$d^{(l)} = \frac{1}{T}\sum_{i=1}^{T} X^{(l)}_{:,i,:} \in \mathbb{R}^{B\times C}. \tag{1}$$

The descriptor is then projected into a shared semantic subspace of dimension d_{common}:

$$c^{(l)} = d^{(l)} W_s^\top \in \mathbb{R}^{B\times d_{\text{common}}}, \quad W_s \in \mathbb{R}^{d_{\text{common}}\times C}. \tag{2}$$

Within the shared semantic subspace, dcSE applies two consecutive SE-style transformations to balance modality-specific adaptation and cross-modal semantic consistency. Both *PrivateSE* and *SharedSE* adopt the standard two-layer bottleneck structure of SE blocks. Specifically, $W_{p1}, W_{s1} \in \mathbb{R}^{d_r\times d_{\text{common}}}$ and $W_{p2}, W_{s2} \in \mathbb{R}^{d_{\text{common}}\times d_r}$ are learnable projection matrices, where d_r denotes the reduced channel dimension. PrivateSE uses modality-specific parameters, whereas SharedSE shares parameters across modalities to enforce semantic consistency.

The *PrivateSE* operation uses modality-specific parameters and is defined as

$$p^{(l)} = \text{PrivateSE}(c^{(l)}) = \phi\Big(c^{(l)} W_{p1}^\top\Big) W_{p2}^\top, \tag{3}$$

where $\phi(\cdot)$ denotes the ReLU activation and $p^{(l)} \in \mathbb{R}^{B\times d_{\text{common}}}$.

The subsequent *SharedSE* operation shares parameters across modalities to enforce semantic consistency:

$$q^{(l)} = \text{SharedSE}(p^{(l)}) = \sigma\Big(\phi\Big(p^{(l)} W_{s1}^\top\Big) W_{s2}^\top\Big), \tag{4}$$

where $\sigma(\cdot)$ denotes the sigmoid function and $q^{(l)} \in \mathbb{R}^{B\times d_{\text{common}}}$.

Finally, the shared semantic signal is mapped back to the original channel space through a modality-specific excitation:

$$w^{(l)} = \sigma\Big(q^{(l)} W_e^\top\Big) \in \mathbb{R}^{B\times C}, \quad W_e \in \mathbb{R}^{C\times d_{\text{common}}}, \tag{5}$$

which is broadcast to all tokens and applied channel-wise:

$$\widetilde{X}^{(l)}_{b,i,:} = X^{(l)}_{b,i,:} \odot w^{(l)}_{b,:}, \quad \forall b \in [1,B],\ \forall i \in [1,T]. \tag{6}$$

By decoupling semantic aggregation and channel excitation, dcSE enables soft cross-modal semantic interaction at the level of global channel statistics, rather than enforcing explicit token- or instance-level alignment, while preserving intra-modal feature modulation.

3.2 Distribution-Structure Co-adaptation (DiSCo)

To further refine intermediate representations after dcSE, we introduce *Distribution-Structure Co-adaptation (DiSCo)*, a dual-path module that operates directly in the feature space. Rather than explicitly fusing visual and textual features, DiSCo produces complementary features from two parallel paths: (i) a β-path that captures cross-modal semantic bias through distribution-aware aggregation; and (ii) a γ-path that performs structure-aware channel calibration to preserve intra-modal structure. These features are subsequently injected back into the original representations.

β-path. For each sample, we extract a global semantic descriptor from the current modality using the `[CLS]` token, denoted as $h^{(b)}_{\text{self}} \in \mathbb{R}^C$, $b \in \{1,\dots,B\}$. From the other modality, we collect the token-level features as a set of semantic bases $H^{(b)}_{\text{other}} = \{h^{(b,j)}_{\text{other}}\}_{j=1}^{T}$, where T denotes the number of tokens and j indexes tokens in the other modality. These bases are taken from the intermediate token features of the other modality at the same block after dcSE.

Although dcSE brings visual and textual features into a shared semantic space, correspondence between individual tokens across modalities remains inherently ambiguous. Different modalities may emphasize different attributes or components of the same concept, and spurious correlations under domain shift further undermine the reliability of one-to-one token matching. As a result, enforcing explicit or hard alignment risks amplifying early mismatches and destabilizing adaptation. Moreover, since $h_{\text{self}}^{(b)}$ summarizes global semantics, collapsing it onto a single token basis is unnecessarily restrictive; a distributed association better reflects that multiple tokens may jointly represent the same concept.

The β-path models cross-modal interaction at the level of *semantic emphasis*. Specifically, we treat the global descriptor $h_{\text{self}}^{(b)}$ as distributing its semantic contribution over the token bases $\{h_{\text{other}}^{(b,j)}\}$ from the other modality, yielding a soft and distributed form of alignment.

To this end, we first measure semantic discrepancy using cosine distance:

$$M_{bj} = 1 - \cos\big(h_{\text{self}}^{(b)},\, h_{\text{other}}^{(b,j)}\big), \quad b \in [1, B],\ j \in [1, T], \quad (7)$$

where smaller values indicate higher semantic affinity. We then formulate semantic emphasis alignment as a mass redistribution problem, in which a unit amount of semantic contribution from $h_{\text{self}}^{(b)}$ is softly allocated over the token bases according to M_{bj}.

This formulation naturally leads to entropy-regularized optimal transport (OT) [Cuturi, 2013], which yields a smooth and normalized *transport plan* under a mass-conservation constraint, preventing degenerate concentration on a single token and providing an interpretable distribution of semantic emphasis. For each sample b, we solve

$$\Gamma^{*(b)} = \underset{\Gamma^{(b)} \in \mathbb{R}_+^{1\times T}}{\arg\min} \left\langle \Gamma^{(b)},\, M^{(b)} \right\rangle - \varepsilon H(\Gamma^{(b)}) \ \text{ s.t. } \Gamma^{(b)}\mathbf{1} = 1. \quad (8)$$

where $M^{(b)} \in \mathbb{R}^{1\times T}$ is the b-th row of M, $\mathbf{1} \in \mathbb{R}^T$ is an all-ones vector, and $H(\Gamma^{(b)}) = -\sum_{j=1}^{T} \Gamma_j^{(b)} \log\big(\Gamma_j^{(b)}\big)$ is the entropy regularizer with $\varepsilon > 0$ controlling the smoothness.

The resulting transport weights $\Gamma^{*(b)}$ encode how strongly each token basis contributes to cross-modal adaptation. We aggregate them into a distribution-aware semantic bias:

$$\beta^{(b)} = \sum_{j=1}^{T} \Gamma_j^{*(b)}\, h_{\text{other}}^{(b,j)} \in \mathbb{R}^C, \qquad \beta = \{\beta^{(b)}\}_{b=1}^{B} \in \mathbb{R}^{B\times C}. \quad (9)$$

Injected as an additive bias, β encourages the two modalities to emphasize consistent semantic content without enforcing hard alignment.

γ-path. While the β-path encourages cross-modal semantic alignment, adaptation may still distort the intrinsic intra-modal structure learned from pre-training, especially under domain shift. The γ-path therefore derives a structure-aware calibration signal from token statistics to stabilize channel responses within each modality.

Let $X \in \mathbb{R}^{B\times T\times C}$ denote the intermediate token features of the current modality. For each sample b, we view $X^{(b)} = X_{b,:,:} \in \mathbb{R}^{T\times C}$ as T observations in a C-dimensional channel space and compute a covariance statistic $\Sigma^{(b)} = \frac{1}{T}\big(\hat{X}^{(b)}\big)^\top \hat{X}^{(b)} \in \mathbb{R}^{C\times C}$, where $\hat{X}^{(b)} = X^{(b)} - \mathbf{1}(\bar{x}^{(b)})^\top$ and $\bar{x}^{(b)} = \frac{1}{T}\sum_{i=1}^{T} X_{i,:}^{(b)}$. We obtain the dominant structural direction $u^{(b)} \in \mathbb{R}^C$ as the leading eigenvector of $\Sigma^{(b)}$, approximated by a few steps of power iteration [Golub and Van Loan, 2013]. Its strength is measured by the Rayleigh quotient $\lambda^{(b)} = (u^{(b)})^\top \Sigma^{(b)} u^{(b)}$, and the channel-wise calibration vector is defined as

$$\gamma^{(b)} = \lambda^{(b)} u^{(b)} \in \mathbb{R}^C, \qquad \gamma = \{\gamma^{(b)}\}_{b=1}^{B} \in \mathbb{R}^{B\times C}. \quad (10)$$

The resulting γ is broadcast over tokens and applied as a multiplicative modulation signal, highlighting structurally salient channels and reducing intra-modal structural drift during adaptation.

Feature Integration. Finally, the outputs of the two paths are injected into the original token-level features X:

$$\widetilde{X}_{b,i,:} = X_{b,i,:} \odot \big(1 + \gamma^{(b)}\big) + \beta^{(b)}, \quad \forall b \in [1, B],\ \forall i \in [1, T]. \quad (11)$$

Through the complementary effects of β and γ, DiSCo refines representations by coordinately leveraging cross-modal semantic distributions and intra-modal structural cues, without enforcing hard feature alignment.

3.3 Parent Class Generation

To further enhance textual semantics with hierarchical knowledge, we propose a parent class generation strategy based on LLMs. This design is motivated by human hierarchical cognition: when recognizing unfamiliar objects, humans often reason from coarse categories to fine-grained concepts.

Given a label set $\mathcal{C} = \{c_1, \ldots, c_N\}$, we employ DeepSeek [Liu *et al.*, 2024] to infer a semantically meaningful parent concept for each class using carefully designed prompts, e.g., "Based on function or form, assign {class_name} to a broad object category." For each class c_n, the LLM generates a concise parent label p_n, such as

$$\text{samoyed} \to \text{Canidae}, \quad \text{ant} \to \text{insect}.$$

Each class is paired with its parent concept to construct a hierarchical textual prompt:

$$\begin{aligned}\text{Prompt}(c_n) = \big[&\texttt{SOS}\ \textit{a photo of}\ c_n,\ \textit{a type of}\\ &v_1 \ldots v_K\ p_n\ \texttt{EOS}\big],\end{aligned} \quad (12)$$

where $\{v_k\}$ are learnable context tokens.

4 Experiments

To evaluate our method, we follow the experiment setup established in previous works such as MaPLe [Khattak *et al.*, 2023a] and HPT [Wang *et al.*, 2024]. We compare the performance of CoDA with the baseline zero-shot CLIP [Radford *et al.*, 2021] and state-of-the-art (SOTA) fine-tuning methods, including Co-CoOp [Zhou *et al.*, 2022a], MaPLe [Khattak *et al.*, 2023a], promptSRC [Khattak *et al.*, 2023b], MetaPrompt [Zhao *et al.*, 2024], HPT [Wang *et al.*, 2024], 2SFS [Farina *et al.*, 2025] and FCPrompt-G [Liu *et al.*, 2025]. Comprehensive ablation studies are also conducted to demonstrate the effectiveness of each component in CoDA.

Method		ImageNet	Caltech	Flowers	Pets	Cars	Food101	Aircraft	SUN397	DTD	EuroSAT	UCF101	Avg.
CLIP	B	72.43	96.84	72.08	91.17	63.37	90.10	27.19	69.36	53.24	56.48	70.53	69.34
	N	68.14	94.00	77.80	97.26	74.89	91.22	36.29	75.35	59.90	64.05	73.85	74.22
	H	70.22	95.40	74.83	94.12	68.65	90.66	31.09	72.23	56.37	60.03	73.85	71.70
CoCoOp (CVPR 2022)	B	75.98	97.96	94.87	95.20	70.49	90.70	33.41	79.74	77.01	87.49	82.33	80.47
	N	70.43	93.81	71.75	97.69	73.59	91.29	23.71	76.86	56.00	60.04	73.45	71.69
	H	73.10	95.84	81.71	96.43	72.01	90.99	27.74	78.27	64.85	71.21	77.64	75.83
MaPLe (CVPR 2023)	B	76.66	97.74	95.92	95.43	72.94	90.71	37.44	80.82	80.36	94.07	83.00	82.28
	N	70.54	94.36	72.46	97.76	74.00	92.05	35.61	78.70	59.18	73.23	78.66	75.14
	H	73.47	96.02	82.56	96.58	73.47	91.38	36.50	79.75	68.16	82.35	80.77	78.55
PromptSRC (ICCV 2023)	B	77.60	98.10	98.07	95.33	78.27	90.67	42.73	82.67	83.37	92.90	87.10	84.26
	N	70.73	94.03	76.50	97.30	**74.97**	91.53	37.87	78.47	62.97	73.90	78.80	76.10
	H	74.01	96.02	85.95	96.30	76.58	91.10	40.15	80.52	71.75	82.32	82.74	79.97
MetaPrompt (TIP 2024)	B	77.39	98.28	97.53	95.71	75.43	**90.76**	39.38	82.10	82.52	93.37	84.70	83.38
	N	71.06	94.58	74.54	96.98	74.43	91.77	37.59	79.01	60.10	78.34	78.56	76.09
	H	74.09	96.39	84.50	96.34	74.93	91.26	38.46	80.53	72.79	85.20	81.51	79.57
HPT (AAAI 2024)	B	77.95	98.37	98.17	95.78	76.95	90.46	42.68	82.57	83.84	94.24	86.52	84.32
	N	70.74	**94.98**	**78.37**	97.65	74.23	91.57	38.13	78.20	63.33	77.12	80.06	76.86
	H	74.17	96.42	**87.16**	96.71	75.57	91.01	40.28	80.35	68.25	84.82	83.16	80.42
2SFS (CVPR 2025)	B	77.71	**98.71**	98.29	95.32	82.50	89.11	**47.48**	82.59	84.60	**96.91**	87.85	**85.55**
	N	70.99	94.43	76.17	97.82	74.80	91.34	35.51	78.91	65.01	67.09	78.19	75.48
	H	**74.20**	**96.52**	85.83	96.55	78.46	90.21	40.63	80.70	73.52	79.29	82.74	80.20
FCPrompt-G (TPAMI 2025)	B	77.33	98.41	98.32	**96.40**	80.87	**90.76**	43.68	82.93	**84.61**	96.80	**87.99**	85.28
	N	**71.11**	94.07	74.82	**97.89**	73.03	**92.19**	36.97	**79.26**	60.31	75.21	78.28	75.72
	H	74.09	96.19	84.98	**97.14**	76.75	**91.47**	40.05	**80.88**	70.42	84.65	82.85	80.22
CoDA	B	**78.87**	98.47	**98.43**	95.73	**82.80**	90.47	44.70	**82.97**	83.27	94.27	87.17	85.20
	N	70.30	93.60	77.80	97.13	74.89	91.23	**39.53**	77.97	**66.47**	**82.27**	**80.63**	**77.44**
	H	74.11	95.97	86.91	96.42	**78.64**	90.85	**41.96**	80.39	**73.93**	**87.86**	**83.77**	**81.13**

Table 1: Comparison with SOTA methods on 11 datasets under base-to-new generalization setting. B denotes the results on the base set, N denotes those on the novel set, and H denotes their harmonic means. The average performances are shown in the last column. Best results are labeled by bold fonts.

4.1 Datasets

We report results on 11 datasets used in CoOp [Zhou *et al.*, 2022b]: ImageNet [Deng *et al.*, 2009], Caltech [Fei-Fei *et al.*, 2004], OxfordPets [Parkhi *et al.*, 2012], StanfordCars [Krause *et al.*, 2013], Flowers [Nilsback and Zisserman, 2008], Food101 [Bossard *et al.*, 2014], FGVCAircraft [Maji *et al.*, 2013], EuroSAT [Helber *et al.*, 2019], UCF101 [Soomro *et al.*, 2012], DTD [Cimpoi *et al.*, 2014] and SUN397 [Xiao *et al.*, 2010].

4.2 Implementation Details

We use ViT-B/16 as the vision backbone and perform prompt tuning on the pre-trained CLIP model. For base-to-new generalization, we optimize the model using SGD with an initial learning rate of 0.017 for 10 epochs and a batch size of 32. Following prior work, we compute image-text similarity scores using the CLIP logit scale and apply a softmax to obtain class probabilities, and train with cross-entropy loss. In accordance with CoOp, we use 16 shots for training and evaluate on the full test set.

4.3 Quantitative Performance on Various Tasks

Base-to-novel generalization. Similar to existing CoOp-based methods, an important evaluation of fine-tuning involves a base-to-new generalization setting composed of two disjoint subsets: base classes and novel classes. The trainable parameters are obtained by using labeled base classes and evaluated on both base classes and unseen novel classes. Table 1 reports the base-to-new generalization results across 11 datasets. CoDA consistently improves performance on novel classes while maintaining competitive accuracy on base classes, leading to a superior harmonic mean compared with existing methods.

Few-shot learning. Table 2 presents the few-shot learning results under 1, 4 and 16-shot settings. It shows that as the number of shots increases, our method achieves a higher average accuracy across 11 datasets. While our method does not always achieve the highest accuracy on every dataset, it generally ranks among the top performers, demonstrating strong overall performance. These results show that CoDA performs robustly in low-shot settings. By leveraging distribution-aware semantic cues and structure-aware calibration, the proposed method mitigates overfitting under limited supervision, and exhibits steady performance gains as more training samples become available.

Domain generalization. In Table 3, we compare the performances on the domain generalization task following prior benchmarks [Roy and Etemad, 2023; Park *et al.*, 2024]. CoDA achieves the best average accuracy of 60.97%, demonstrating its generalization capability across different domain shifts.

4.4 Ablation Study

Effects of DiSCo, dcSE, and PCG. We conduct ablation studies to examine the individual and combined effects of the proposed components including DiSCo, dcSE, and PCG.

Shots	Method	ImageNet	Caltech	Pets	Cars	Flowers	Food101	Aircraft	SUN397	DTD	EuroSAT	UCF101	Avg.
1-shot	CoOp	65.7	92.5	90.2	67.5	78.3	84.3	20.8	67.0	50.1	56.4	71.2	67.6
	MaPLe	69.7	93.6	91.4	67.6	74.9	85.4	28.1	69.3	50.0	29.1	71.1	66.4
	PLOT	66.5	**94.3**	91.9	68.0	80.5	**86.2**	28.6	66.8	**54.6**	65.4	74.3	70.7
	CLIP-LoRA	**70.4**	93.7	**92.3**	70.1	83.2	84.3	30.2	**70.4**	54.3	72.3	**76.3**	**72.5**
	Ours	69.1	93.7	90.8	**71.9**	**84.6**	85.4	**31.3**	69.7	53.2	**72.4**	74.8	**72.5**
4-shot	CoOp	68.8	94.5	92.5	74.4	92.2	84.5	30.9	69.7	59.5	69.7	77.6	74.0
	MaPLe	70.6	95.0	93.3	70.1	84.9	**86.7**	30.1	71.4	59.0	69.9	77.1	73.5
	PLOT	70.4	95.1	92.6	76.3	92.9	86.5	35.3	71.7	62.4	83.2	79.8	76.9
	CLIP-LoRA	**71.4**	**95.2**	91.0	77.4	93.7	82.7	37.9	72.8	63.8	**84.9**	81.1	77.4
	Ours	71.2	**95.2**	**93.3**	**78.8**	**94.0**	86.3	**38.9**	**73.9**	**64.7**	78.8	**82.1**	**77.9**
16-shot	CoOp	71.9	95.8	92.0	82.9	96.8	84.2	43.2	74.9	69.7	85.0	83.1	80.0
	MaPLe	71.9	95.4	93.2	74.3	94.2	87.4	36.8	74.5	68.4	87.5	81.4	78.6
	PLOT	72.6	96.0	93.6	84.6	97.6	87.1	46.7	76.0	71.4	92.0	85.3	82.1
	CLIP-LoRA	73.6	**96.4**	92.4	86.3	98.0	84.2	**54.7**	76.1	72.0	**92.1**	86.7	83.0
	Ours	**74.4**	**96.4**	**94.0**	**86.4**	**98.4**	**87.6**	51.4	**77.6**	**74.2**	90.2	**87.0**	**83.4**

Table 2: Performance comparison on few-shot setting on the ViT-B/16 backbone.

	Source	Target				
Method	ImageNet	-V2	-Sketch	-A	-R	Avg.
CoOp	71.51	64.20	47.99	49.71	75.21	59.28
MaPLe	70.72	64.07	49.15	50.90	76.98	60.26
PromptSRC	71.27	64.35	49.55	50.90	77.80	60.65
CoPrompt	70.80	64.25	49.43	50.50	77.51	60.42
ProMetaR	71.29	64.39	49.55	**51.25**	**77.89**	60.77
CoDA	**73.43**	**66.17**	**50.40**	49.83	77.47	**60.97**

Table 3: Performance comparison on domain generalization.

DiSCo	dcSE	PCG	Base	New	HM
			83.35	67.14	74.37
✓			85.07	70.93	77.36
	✓		**85.25**	71.35	77.68
		✓	84.99	68.56	75.90
✓	✓		85.14	73.59	78.94
✓		✓	82.96	75.11	78.84
	✓	✓	83.33	75.80	79.39
✓	✓	✓	85.20	**77.44**	**81.13**

Table 4: The effects of DiSCo, dcSE, and PCG.

As shown in Table 4, each component contributes positively to overall performance, demonstrating their effectiveness. When all components are jointly enabled, CoDA achieves the best harmonic mean. In particular, DiSCo and dcSE yield substantial improvements on novel classes, indicating enhanced robustness under distribution shift.

Notably, PCG provides moderate gains when used alone, but becomes significantly more effective when combined with DiSCo or dcSE, suggesting that hierarchical semantics are best leveraged under explicit cross-modal interaction.

Effect of cross-modal interaction. To further examine the role of cross-modal interaction, we evaluate two additional variants that restrict adaptation to a single modality: *text-only*, where only the textual branch is adapted, and *vision-only*, where adaptation is applied exclusively to the visual branch. Both variants remove explicit cross-modal interaction while preserving comparable parameter budgets.

As reported in Table 5, single-modality adaptation consistently underperforms the full CoDA model. In particular, vision-only adaptation suffers from noticeable degradation on novel classes, highlighting the difficulty of generalization without textual guidance. Text-only performs better than vision-only on novel classes, suggesting textual adaptation is particularly important for generalizing to unseen categories. In contrast, jointly adapting both modalities leads to the best overall performance, confirming that CoDA's gains stem from coordinated vision-language co-adaptation rather than stronger uni-modal tuning.

	Base	New	HM
Text-only	80.34	74.32	77.21
Vision-only	83.24	71.71	77.05
CoDA	**85.20**	**77.44**	**81.13**

Table 5: The effect of cross-modal interaction.

Effect of the depth of tuning layers. We further study how the depth of applying CoDA affects base-to-new generalization. Table 6 reports results from the baseline model (without CoDA) to progressively deeper adaptation ranges, where "12" denotes adapting only the last Transformer block and "10→12" (resp. "7→12", "2→12") denotes adapting the last three (resp. six, eleven) blocks in both branches.

Without CoDA, the model achieves strong base accuracy but generalizes poorly to novel classes. Adapting only the last block substantially improves novel accuracy, but causes a notable drop on base classes, indicating that shallow adaptation can overly perturb the decision boundary learned for base classes. As the adaptation depth increases (10→12, 7→12, 2→12), base accuracy steadily recovers while novel accuracy continues to improve, yielding consistent gains in harmonic mean. Applying CoDA to all layers achieves the best overall results, suggesting that CoDA benefits from stable and cumulative refinement across depth rather than relying on a single-layer correction.

More effect of PCG. We further analyze PCG to verify that its gains come from meaningful semantic guidance rather than a superficial prompt change. Table 7 compares three variants: (i) removing PCG (w/o PCG), (ii) using "xxx" as

Layer	Baseline	12	10→12	7→12	2→12	CoDA
Base	83.35	79.31	82.36	84.28	85.09	**85.20**
Novel	67.14	73.80	75.87	76.96	77.40	**77.44**
HM	74.37	76.46	78.98	80.45	81.06	**81.13**

Table 6: Comparison of CoDA with the baseline by fine-tuning the last few layers on 11 datasets under the base-to-novel generalization setting.

Setting	Base	New	HM
w/o PCG	85.14	73.59	78.94
Incorrect PCG	85.18	75.15	79.85
CoDA	**85.20**	**77.44**	**81.13**

Table 7: The effect of PCG with correct and incorrect hierarchical semantics. To create the incorrect setting, we set all parent classes to "xxx".

the incorrect parent token for all classes (Incorrect PCG), and (iii) the full CoDA with semantically correct PCG.

Introducing an incorrect PCG already improves over w/o PCG, suggesting that a hierarchical prompt scaffold can act as a useful inductive bias. Replacing incorrect PCGs with semantically meaningful ones consistently yields further gains on novel classes and the harmonic mean, indicating that semantic correctness is crucial for robust generalization. We further visualize class-center embeddings with t-SNE in Figure 2, in which correct PCG produces more compact clusters, supporting the quantitative improvements observed in Table 7.

4.5 Visualization of Saliency Consistency

To further analyze how CoDA adapts vision-language representations at the perceptual level, we visualize saliency heatmaps produced by the visual encoder under CoDA. Figure 3 presents representative examples across four datasets.

Despite notable variations in pose, background and appearance, CoDA exhibits highly consistent saliency patterns within each category. For instance, in *OxfordPets-Beagle*, the attention is consistently concentrated on the head and facial regions, while in *Caltech101-Face* and *ImageNet-Tench*, the model focuses on the most discriminative semantic regions across different instances. Similarly, for *OxfordFlowers-Wallflower*, saliency maps highlight characteristic petal and floral structures rather than background textures.

These observations suggest that CoDA does not rely on spurious visual cues, but instead learns category-level semantic features that are stable across samples. Such intra-class saliency consistency complements the quantitative gains reported earlier, indicating that CoDA improves generalization by encouraging structured and semantically grounded feature adaptation, rather than instance-specific memorization.

5 Conclusion

We proposed CoDA, a new dual-path adaptation framework that effectively coordinates cross-modal semantic alignment and intra-modal structural consistency for adapting VLMs.

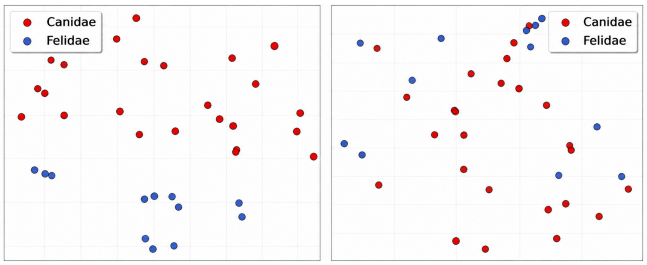

Figure 2: Class-center t-SNE visualization with PCG (left) and without PCG (right). Each point represents a class prototype obtained by averaging image features of the corresponding class. Incorporating PCG leads to more compact and semantically structured class-level clustering.

Figure 3: Visual comparison of saliency maps. Columns correspond to dataset-category pairs (Caltech101, OxfordFlowers, ImageNet, and OxfordPets), and rows show the original image and saliency maps produced by HPT, CoPrompt and our CoDA, respectively. CoDA yields more consistent and category-relevant activation patterns across samples under appearance and background variations.

CoDA consists of three components: dcSE performs decoupled squeeze-excitation to enable controlled semantic exchange via global channel modulation; DiSCo refines intermediate representations with distribution-aware cross-modal semantic bias and structure-aware intra-modal calibration; and PCG injects hierarchical semantic priors into textual prompts.

Extensive experiments show that CoDA achieves consistent improvements over SOTA methods on base-to-new generalization and few-shot learning across multiple datasets. Ablation studies validate the contribution of each component and their complementary effects. Overall, our results suggest that coordinately modeling cross-modal alignment and intra-modal structure is key to robust and transferable adaptation of VLMs.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant 62225601, U23B2052), the Beijing Natural Science Foundation Project No.L242025, the Gansu Province Top Leading Talents Foundation, the Scientific Research Innovation Capability Support Project for Young Faculty under Grant SRICSPYF-BS2025009, the Gansu Provincial Key Talent Project under Grant 2025RCXM002, the Fundamental Research Funds for the Beijing University of Posts and Telecommunications under Grant 2025AI4S15, the Guizhou Province Science and Tech-

nology Plan Project (No.QianKeHeZhongDa [2025]031), the Beijing Key Laboratory of Multimodal Data Intelligent Perception and Governance, and the Royal Society under International Exchanges Award IEC\NSFC\201071.

References

[Bossard *et al.*, 2014] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 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*, pages 446–461. Springer, 2014.

[Cimpoi *et al.*, 2014] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014.

[Cuturi, 2013] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[Farina *et al.*, 2025] Matteo Farina, Massimiliano Mancini, Giovanni Iacca, and Elisa Ricci. Rethinking few-shot adaptation of vision-language models in two stages. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29989–29998, 2025.

[Fei-Fei *et al.*, 2004] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004.

[Gao *et al.*, 2024] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024.

[Golub and Van Loan, 2013] Gene H Golub and Charles F Van Loan. *Matrix computations*. JHU press, 2013.

[Helber *et al.*, 2019] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

[Hu *et al.*, 2018] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[Jia *et al.*, 2021] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[Jiang *et al.*, 2022] Haojun Jiang, Jianke Zhang, Rui Huang, Chunjiang Ge, Zanlin Ni, Jiwen Lu, Jie Zhou, Shiji Song, and Gao Huang. Cross-modal adapter for text-video retrieval. *arXiv preprint arXiv:2211.09623*, 2022.

[Khattak *et al.*, 2023a] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19113–19122, 2023.

[Khattak *et al.*, 2023b] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15190–15200, 2023.

[Krause *et al.*, 2013] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013.

[Li *et al.*, 2023] Xin Li, Dongze Lian, Zhihe Lu, Jiawang Bai, Zhibo Chen, and Xinchao Wang. Graphadapter: Tuning vision-language models with dual knowledge graph. *Advances in Neural Information Processing Systems*, 36:13448–13466, 2023.

[Liu *et al.*, 2024] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[Liu *et al.*, 2025] Liangchen Liu, Nannan Wang, Chen Chen, Decheng Liu, Xi Yang, Xinbo Gao, and Tongliang Liu. Frequency-based comprehensive prompt learning for vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[Maji *et al.*, 2013] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[Nilsback and Zisserman, 2008] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008.

[Ouali *et al.*, 2023] Yassine Ouali, Adrian Bulat, Brais Matinez, and Georgios Tzimiropoulos. Black box few-shot adaptation for vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15534–15546, October 2023.

[Park *et al.*, 2024] Jinyoung Park, Juyeon Ko, and Hyunwoo J Kim. Prompt learning via meta-regularization. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 26940–26950, 2024.

[Parkhi *et al.*, 2012] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012.

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

[Roy and Etemad, 2023] Shuvendu Roy and Ali Etemad. Consistency-guided prompt learning for vision-language models. *arXiv preprint arXiv:2306.01195*, 2023.

[Soomro *et al.*, 2012] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[Wang *et al.*, 2024] Yubin Wang, Xinyang Jiang, De Cheng, Dongsheng Li, and Cairong Zhao. Learning hierarchical prompt with structured linguistic knowledge for vision-language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 5749–5757, 2024.

[Xiao *et al.*, 2010] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010.

[Yang *et al.*, 2024] Lingxiao Yang, Ru-Yuan Zhang, Yanchen Wang, and Xiaohua Xie. Mma: Multi-modal adapter for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23826–23837, June 2024.

[Yao *et al.*, 2021] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021.

[Yu *et al.*, 2023] Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang. Task residual for tuning vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10899–10909, 2023.

[Zhai *et al.*, 2023] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023.

[Zhang *et al.*, 2021] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021.

[Zhao *et al.*, 2024] Cairong Zhao, Yubin Wang, Xinyang Jiang, Yifei Shen, Kaitao Song, Dongsheng Li, and Duoqian Miao. Learning domain invariant prompt for vision-language models. *IEEE Transactions on Image Processing*, 33:1348–1360, 2024.

[Zhou *et al.*, 2022a] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022.

[Zhou *et al.*, 2022b] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022.