ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240504

PromptCD: Coupled and Decoupled Prompt Learning for Vision-Language Models

Junjie Wu[a], **Mingjie Sun**[a,b], **Chen Gong**[a,b], **Nan Yu**[a] **and Guohong Fu**[a,b,*]

[a]School of Computer Science and Technology, Soochow University, Suzhou, China
[b]Institute of Artificial Intelligence, Soochow University
{20224027010, nyu}@stu.suda.edu.cn, {mjsun, gongchen18, ghfu}@suda.edu.cn

Abstract. Large-scale pre-trained vision-language models (VLMs), like CLIP, have presented striking generalizability for adapting to image classification in a few shot setting. Most existing methods explore a set of learnable tokens, such as prompt learning, on data-efficient utilization for task adaptation. However, they focus on either the coupled-modality property by prompt projection or decoupled-modality characteristic by prompt consistency, which ignores effective interaction between prompts. To model the deep yet sufficient cross-modal interaction and enhance the generalization between both seen and unseen tasks, in this paper, we propose a novel coupled and decoupled prompt learning framework, dubbed PromptCD, for vision-language models. Specifically, we introduce a bi-directional coupled-modality mechanism to intensify the interaction between both vision and language branches. Additionally, we propose mixture consistency to further improve the generalization and discrimination of the models on unseen tasks. The integration of such a mechanism and consistency facilitates the proposed framework adaptation for various downstream tasks. We conduct extensive experiments on 11 image classification datasets under a range of evaluation protocols, including base-to-novel and domain generalization, and cross-dataset recognition. Experimental results demonstrate that our proposed PromptCD overall outperforms state-of-the-art methods.

1 Introduction

With remarkable advances in the community of artificial intelligence (AI), large-scale pre-trained vision-language foundation models, such as ALIGN [15] and CLIP [34], have owned powerful generalization to capture open-vocabulary visual concepts for domain-specific image classification. A pioneering work by CLIP leverages a contrastive loss [37] to train a two-tower architecture consisting of image and text encoders, which aligns features of the paired image-text in a common latent space. Unlike conventional fine-tuning methods, a more efficient alternative, like prompt learning, usually introduces a few learnable tokens to optimize the CLIP model with most of pre-trained parameters frozen. Such paradigm better adapts the pre-trained models to various downstream tasks.

In the literature, according to the type of modality signals perceived by prompts, mainstream CLIP-based prompt learning approaches can be roughly categorized into the following three types: vision branch, language branch, and both. The first two approaches [11, 16, 54, 55] target to learn uni-modal prompts, which neglects the distributions of new classes embedding. The latter one approaches [18, 19, 23, 27, 36, 40, 44, 47] are to simultaneously treat a continuous set of learnable prompts in both modality branches. Typically, due to the inherent heterogeneity between modalities, dual-branch prompt learning approaches stimulate the models to learn discriminative representations for vision-language tasks.

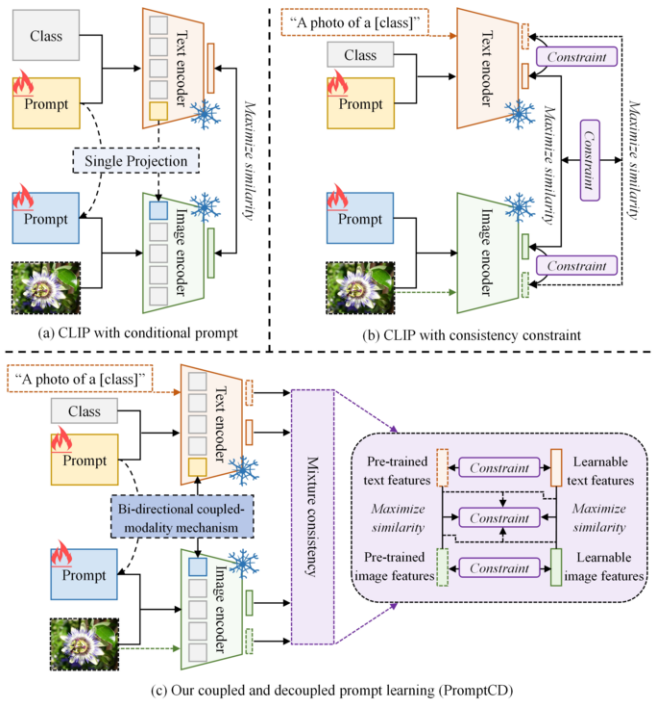

Figure 1. Architecture comparison between PromptCD and existing CLIP-based prompt learning approaches (likewise MapLe[18] and PromptSRC [19]). (a) CLIP with conditional prompt paradigm performs the interaction between branches via uni-direction prompt projection. (b) CLIP with consistency constraint eliminates the discrepancy between seen and unseen tasks by prompt regularization. (c) PromptCD integrates the bi-directional interaction of learnable prompts and mixture consistency for both coupled- and decoupled-modality characteristics.

However, there are two critical challenges in dual-branch prompt learning studies based on the CLIP model. Most of existing CLIP-based dual-branch prompt learning approaches overfocus on either the coupled-modality property or decoupled-modality characteristic of CLIP. Typically, the former [18, 28] refers to pushing one modality to align with another through a coupling function, which aims

* Corresponding author.

to eliminate the discrepancy between modalities. However, it utilizes prompts to describe task-specific objectives [55], which prevents the models from learning task-agnostic knowledge. Moreover, such methods solely consider uni-direction prompt projection, and easily induce the issue of one-way path learning. This ignores the interaction of learnable prompts that include global semantic information. A comparison of the conditional prompt paradigm for CLIP and our proposed PromptCD framework is presented in Figure 1.

Secondly, previous CLIP-based prompt learning works [19, 49, 52, 54, 56] propose to employ various anti-overfitting strategies for exploring decoupled-modality characteristic of the CLIP model, which captures discriminative representations for both seen and unseen tasks. Among them, prompt regularization has emerged as an efficient way to reach this anticipation. Nevertheless, these works fail to establish the deep interaction between modality branches. It is not easy to sufficiently perform alignment between modality features, especially in scenarios with a small number of training samples. An architectural comparison of the consistency constraint for CLIP and the proposed framework is illustrated in Figure 1.

To mitigate the aforementioned challenges, based on the CLIP model, we propose a novel Coupled and Decoupled Prompt learning framework for image classification, dubbed PromptCD, which aims to model vision-language interaction and further boost the generalization on both seen and unseen tasks. More specifically, for the first challenge, we enforce a bi-directional coupled-modality mechanism on both vision and language branches to elegantly align embeddings, as shown in Figure 1. Unlike prompt tuning by language-to-vision projection, where one modality is dominant, our method focuses on informative knowledge between modalities. Such mechanism effectively enables the correlation between learnable prompts, and integrates them during the process of cross-modal interaction. Furthermore, we enforce mixture consistency between both learnable and pre-trained features on the perspectives of modality and task, which effectively overcomes the overfitting problem when the models combat a new class. We enforce consistency between both learnable and pre-trained encoders on both vision and language branches. On the other hand, we impose consistency between learnable and pre-trained output logits on the combination of task-exclusive and task-agnostic knowledge. By integrating these consistency constraints, as well as a bi-directional coupled-modality mechanism, we can effectively enhance the generalization and discrimination of the models for both seen and unseen tasks.

In summary, the main contributions of this work are three-fold as follows:[1]

- We propose a novel coupled and decoupled prompt learning framework (PromptCD) for pre-trained vision-language foundation models, which effectively improves the performance of image classification on both seen and unseen classes.
- We integrate the bi-directional coupled-modality mechanism and mixture consistency into the CLIP-based prompt learning framework. Concretely, the bi-directional coupled-modality mechanism is to intensify the interaction between both vision and language branches for modality alignment. The mixture consistency proposes to optimize the compromise between dataset distributions.
- Quantitative and qualitative experiments are conducted on 11 image classification datasets for a series of evaluation protocols, including base-to-novel and domain generalization, and cross-dataset recognition, which demonstrates the superiority of our proposed framework over several state-of-the-art methods.

[1] https://github.com/xiaofen623/PromptCD

2 Related Work

In this section, we review pre-trained vision-language foundation models, parameter efficient fine-tuning, and prompt learning in vision-language models, respectively.

Vision-Language Models. Recent years have witnessed the unprecedented growth of VLMs in computer vision (CV) and natural language processing (NLP). VLMs bridge the gap between vision and language from image-text pairs, which shows the impressive generalization on downstream tasks [6, 39, 43, 48]. Typically, existing VLMs involve two perspectives: model architecture and pre-training objective. The former commonly includes fusion-encoder [24, 51] and dual-encoder [26, 40]. Among them, dual-encoder architecture, such as CLIP [34], employs uni-modal encoders to capture their respective semantic features for modality alignment. Such streamlined architecture can mitigate the computation burden.

Moreover, pre-training objectives gradually converge to two types, i.e., generative and discriminative modelling. More specifically, generative modelling includes masked/prefix language prediction [29] and masked region/image prediction [21]. Discriminative modelling contains image-text matching [50] and contrastive learning [1]. Among them, contrastive learning can better align different modality representations in a common vector space. Thus, we focus on CLIP that performs alignment through contrastive learning.

Parameter Efficient Fine-Tuning. Parameter efficient fine-tuning (PEFT) has emerged as a low-cost and effective technique that adapts VLMs to domain-specific downstream tasks [27, 36]. Such technique optimizes a small portion of model parameters, rather than re-training VLMs from scratch. Thereby, it can maximize the compromise between computation burden and information capacity, and achieve significant generalization for various downstream tasks. According to tuning types, existing PEFT approaches can be roughly divided into three types, i.e., Low-Rank Adaption (LoRA) [14], adapters [10] and prompt learning [25, 53]. The first two kinds [17] devise additional modules with randomly initialized parameters, which may perturb the knowledge from the raw models, resulting in sub-optimal learning. The third kind [40] designs a few learnable prompt tokens to adapt model's features towards downstream tasks. Thus, based on its simple design and scaling capacity, we focus on prompt learning in the classic vision-language model, such as CLIP.

Prompt Learning in VLMs. Inspired by the successful applications of prompt learning in the realm of multimodal learning, a series of relevant works [49, 56] explore the potential capacity of prompt learning in VLMs, e.g., CLIP [34], for downstream tasks. For example, CoOp [55] proposes to use learnable prompt vectors for the text branch. VPT [16] proposes to introduce prompt tokens for the image branch of the CLIP model. However, they face the problem of the performance degradation on unseen tasks. To this end, CoCoOp [54] designs a neural network to implement instance-conditional prompts adaptation on CLIP. Unlike the above mentioned approaches that learn prompts via a separate side, MaPLe [18] proposes to align vision-language features by simultaneously optimizing learnable prompts for each branch of CLIP. Beyond that, PromptSRC [19] proposes to employ self-regulate prompts to absorb task-specific and task-agnostic knowledge for CLIP adaptation.

Although these approaches achieve nontrivial improvements to the task of image classification, they neglect the deep and sufficient interaction between vision and language branches. Thus, in this work, we propose a novel coupled and decoupled prompt learning method to stimulate the CLIP model to sufficiently improve generalization and discriminative capabilities between both seen and unseen tasks.

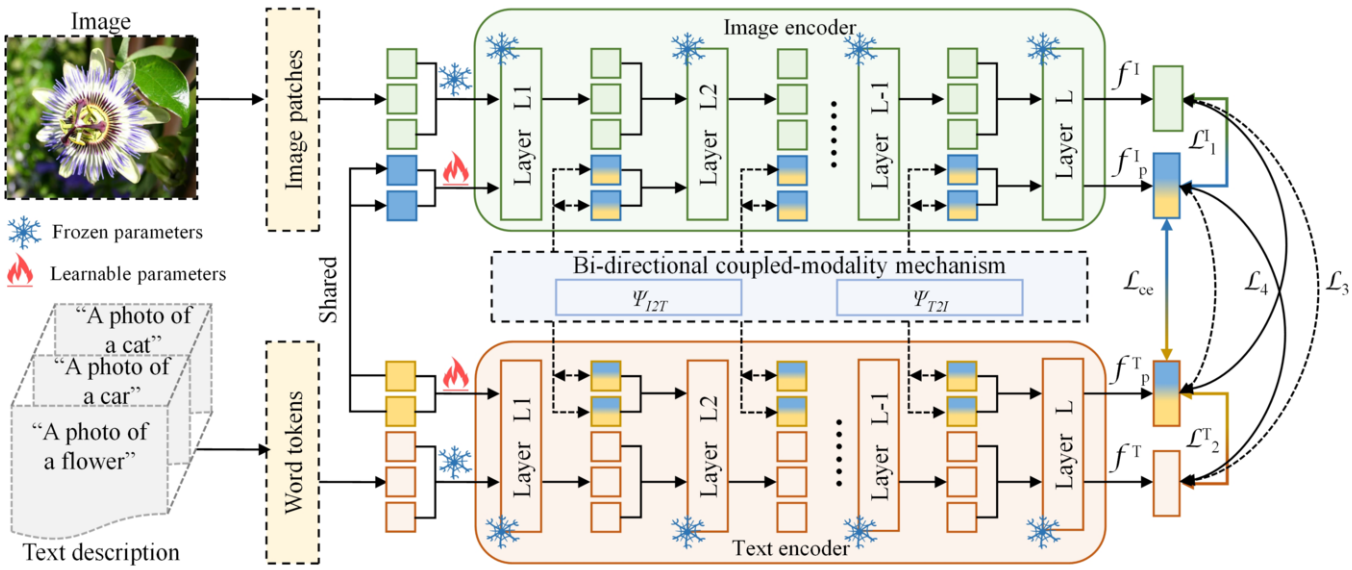

Figure 2. The framework of coupled and decoupled prompt learning (PromptCD) for image classification. For the given image-text pair, PromptCD utilizes image and text encoders to capture multiple embeddings, including learnable text and image embeddings, pre-trained text and image embeddings. In the stage of training, learnable embeddings are integrated by a bi-directional coupled-modality mechanism. Pre-trained embeddings are fed into their respective transformer layers. Then, different modality branches achieve multiple features, including learnable text and image features, pre-trained text and image features. PromptCD pulls learnable features and pre-trained features close by using L1 loss as a constraint. PromptCD employs the Kullback-Leibler divergence loss as a constraint for output logits. In the stage of inference, PromptCD just calculates an output logit between prompt text and image features.

3 Methodology

In this section, we present an overview of the proposed PromptCD framework, as illustrated in Figure 2. First, we give the preliminaries over the prevalent vision-language foundation model, such as CLIP, followed by the details of our proposed PromptCD. Notably, PromptCD can be extended to mainstream pre-trained vision-language foundation models.

3.1 Preliminaries

Since the widely-used pre-trained vision-language foundation model such as CLIP [34] has already shown significant results in various downstream tasks [30, 43], thus we employ such foundation model as the backbone in this work. More specifically, CLIP proposes to align the matched image and text within a common vector space, which mainly consists of an image encoder θ_I and a text encoder θ_T. These modality encoders are based on the transformer network [42]. To better understand, suppose that we have an image classification dataset $D = \{x_i^I, y_i\}_{i=1}^M$, where x_i^I and M refer to an image and the number of samples, respectively. Additionally, each image is assigned with a ground truth y_i from a set of classes $c = \{c_j\}_{j=1}^C$, where C is the number of classes. CLIP generates a range of text descriptions x_j^T by leveraging the hand-crafted template, i.e., "a photo of a $\{c_j\}$", for each class. Thereby, for each input text description x_j^T, the text encoder θ_T encodes them as $z_j^T = \theta_T(x_j^T)$. These text features are combined as the complete text embedding $z^T = [z_1^T, z_2^T, ..., z_C^T]$. For an input image x_i^I, the image encoder θ_I encodes it as the patch embedding z^I. Finally, CLIP calculates the highest similarity between all the text embeddings and image embedding, which can be formulated as follows:

$$P(y_i|x_i^I) = \frac{\exp(sim(z^I, z_{y_i}^T)/\tau)}{\sum_{j=1}^C \exp(sim(z^I, z_j^T)/\tau)}, \quad (1)$$

where, sim is the cosine similarity, and $sim(z^I, z_j^T) = z^I(z_j^T)^{\mathbf{T}}$ denotes the output logit between text and image embeddings. τ refers to a temperature hype-parameter, which is utilized to control the softness of distribution.

Considering the limitation of the hand-crafted template across domains, the recent CLIP-based prompt learning work such as CoOP [55] employs a continuous set of learnable vectors to extract task-specific text embeddings. Typically, CoOp utilizes m learnable prompt vectors $\{v_1, v_2, ..., v_m\}$ to replace the raw sentence "a photo of a $\{c_j\}$", thereby yielding $w_j = \{v_1, v_2, ..., v_m, c_j\}$. On the vision branch, an input image is mapped as d image patches $\{q_1, q_2, ..., q_d\}$. These patches are concatenated with learnable prompt vectors, generating $u = \{v_1, v_2, ..., v_m, q_1, q_2, ..., q_d\}$. Therefore, the optimized objective of CLIP is defined as follows:

$$P(y_i|x_i^I) = \frac{\exp(sim(\theta_I(u), \theta_T(w_{y_i}))/\tau)}{\sum_{j=1}^C \exp(sim(\theta_I(u), \theta_T(w_j))/\tau)}, \quad (2)$$

These aforementioned methods leverage a set of learnable prompt vectors to learn informative knowledge for the task of image classification. However, they have a major limitation that lacks the sufficient interaction between modality prompts. This prevents the pre-trained vision-language foundation models from performing the synergy between image and text features. In the following subsections, we detail the proposed PromptCD, a novel framework that overcomes this limitation and effectively improves the performance of the models on both seen and unseen tasks.

3.2 Bi-directional Coupled-modality Mechanism

Bi-directional coupled-modality mechanism aims to establish the deep and sufficient interaction between both vision and language branches, which intensifies the coupled-modality property of CLIP

to align the pairs of image and text for task-specific datasets distribution. Based on the streamlined two-tower architecture, i.e., independent modality branches of CLIP, we iteratively perform learnable prompt projection for task-specific perception without introducing excessive model parameters. More specifically, on the first layer of the vision branch, the prompt vectors are initialized from text prompts by a linear layer. Text prompts are randomly initialized, which are concatenated into the fixed template. For subsequent transformer layers, we explicitly condition prompts on the interaction between both vision and language branches. That is, we leverage a language-to-vision coupling function to transfer knowledge between branches, followed by adding a vision-to-language coupling function to perform the same target. Formally, the whole process can be expressed as follows:

$$\hat{f}_l^T = f_l^T + \psi_{I2T}(f_l^I),\ \hat{f}_l^I = f_l^I + \psi_{T2I}(f_l^T), \tag{3}$$

where, l denotes the l-th transformer layer of modality encoders. ψ denotes the coupling operation, i.e., linear layers, maintaining the same dimension of features as another branch. $\hat{f}$ are the intergrated embeddings with m learnable prompt vectors.

3.3 Mixture Consistency

Although prompt coupling enables the model to maximize information communication between modalities, it may easily fit task-specific datasets. To enhance representation learning of the model regarding unseen tasks, we impose mixture consistency to improve the generalization of the model for new classes. To distinguish the discrepancy between modality features, we employ the L1 loss as a consistency constraint between learnable and pre-trained features. Notably, other variants, likewise MSE and cosine similarity, can also be utilized to replace L1 loss as the consistency constraint. This constraint is implemented on both vision and language branches, which can be calculated as follows:

$$\mathcal{L}_1^I = |f_p^I - f^I|,\ \mathcal{L}_2^T = |f_p^T - f^T|, \tag{4}$$

where, f_p^I and f_p^T are the learnable features with their respective prompts. f^I and f^T are the pre-trained features from the frozen encoders of the CLIP model.

On the other hand, the discrepancy between output logits can be formulated as follows:

$$\mathcal{L}_3 = KL(P(f_p^I, f_p^T), P(f^I, f^T)), \tag{5}$$

$$\mathcal{L}_4 = KL(P(f_p^I, f^T), P(f^I, f_p^T)), \tag{6}$$

where, the symbol P is declared in preliminaries. KL denotes the Kullback-Leibler (KL) divergence loss for calculating the probability distribution between predictions. Among these output logits, $P(f_p^I, f_p^T)$ is the prompt output logit, and $P(f^I, f^T)$ is the general output logit. $P(f_p^I, f^T)$ denotes the output logit between prompt image features and general text features. $P(f^I, f_p^T)$ denotes the output logit between general image features and prompt text features.

Overall, the proposed mixture consistency is imposed by combining all the constraints. Therefore, the final objective is expressed by:

$$\mathcal{L} = \mathcal{L}_{ce} + \mathcal{L}_1^I + \mathcal{L}_2^T + \alpha\mathcal{L}_3 + (1-\alpha)\mathcal{L}_4, \tag{7}$$

where, $\mathcal{L}_{ce}$ denotes a supervised loss that represents the task of image classification. The sign α denotes a balancing factor.

4 Experiments

4.1 Experimental Settings

Datasets. We evaluate PromptCD on 11 benchmark datasets, i.e., ImageNet [5], Caltech101 [8], OxfordPets [33], StanfordCars [22], Flowers102 [32], Food101 [2], FGVCAircraft [31], SUN397 [46], DTD [4], EuroSAT [12], and UCF101 [38]. For evaluating domain generalization, we adopt four variants of ImageNet, i.e., ImgNet-V2 [35], ImgNet-Sketch [45], ImgNet-A [9], and ImgNet-R [13].

Baselines. We compare PromptCD with a broad spectrum of baselines, including CLIP [34], CoOP [55], CoCoOp [54], ProGrad [56], KgCoOP[49], AAPL[20], MaPLe [18], and PromptSRC [19].

Implementation Details. We implement all the experiments based on the CLIP model, which regards ViT-B/16 [7] as the backbone. Following previous works [18, 55], we use a few-shot training strategy in all experiments at 16 samples per class for all seen classes. The length m and depth l of prompts are set as 4 and 9, respectively. The value of α is set as 0.3. In the training stage, an SGD optimizer is employed for updating learnable parameters with a learning rate of 0.0025, as well as a batch size of 4. The maximal training epoch is set to 25. The reported accuracy and harmonic mean (HM) are an average over three random runs. All experiments are conducted by using PyTorch on a single NVIDIA GeForce RTX 3090 GPU.

4.2 Base-to-novel Generalization

To verify the effectiveness of the proposed PromptCD in the base-to-novel generalization task, where the model is trained on the seen classes in a few-shot setting and evaluated on both seen (base) and unseen (novel) classes in a zero-shot setting, we divide each dataset into base and novel classes following [18, 19]. Table 1 summaries a comparison of PromptCD with baseline methods in the base-to-novel generalization task. From the average results over all classification datasets, we obverse that PromptCD achieves the best performance than other methods. It demonstrates that PropmtCD has strong generalization between seen and unseen tasks. Typically, in the accuracy of base classes, PropmtCD achieves an improvement of 2.16% over MaPLe, and 0.18% over PromptSRC. In the accuracy of novel classes, PropmtCD gains more than 1.27% and 0.31% over MaPLe and PromptSRC, respectively. Meanwhile, we notice that PromptCD obtains a nontrivial improvement regarding the satellite-image dataset EuroSAT. We found the fact that EuroSAT contains only 10 classes, and consists of 13,500 samples, resulting in overfitting on seen classes and performance degradation on unseen classes. PromptCD sufficiently integrates the characteristics of coupled- and decoupled-modality to improve the generalization and discrimination, especially in datasets with fewer classes. Notably, CLIP outperforms most existing methods on unseen tasks which is a challenge for them. Prompt learning makes these models fit seen classes, but leads to an obvious performance degeneration on learning task-agnostic knowledge in the setting of a few-shot samples. However, the utilization of prompt projection or regularization can improve the generalization on both seen and unseen tasks. In the harmonic mean, PromptCD achieves more than 8.53%, 1.68%, and 0.26% improvements over CLIP, MaPLe and PromptSRC, which demonstrates that our proposed method possesses strong generalization and discrimination between seen and unseen tasks. Moreover, from the harmonic mean of each dataset, we can obverse that PromptCD is superior to existing baseline approaches on 6 out of 11 classification datasets. The significant performance illustrates that PromptCD has robust generalization ability between both seen and unseen tasks.

Table 1. Comparison of PromptCD with existing methods in the base-to-new generalization task on 11 classification datasets. The best results are **bolded**.

Dataset	Sets	CLIP[34]	CoOP[55]	CoCoOp[54]	ProGrad[56]	KgCoOP[49]	AAPL[20]	MaPLe[18]	PromptSRC[19]	PromptCD
Average	Base	69.34	82.69	80.47	82.48	80.73	80.27	82.28	84.26	**84.44**
	Novel	74.22	63.22	71.69	70.75	73.06	72.17	75.14	76.10	**76.41**
	HM	71.70	71.66	75.83	76.16	77.00	76.01	78.55	79.97	**80.23**
ImageNet	Base	72.43	76.47	75.98	77.02	75.83	76.53	76.66	77.60	**77.80**
	Novel	68.14	67.88	70.43	66.66	69.96	70.57	70.54	70.73	**70.75**
	HM	70.22	71.92	73.10	71.46	72.78	73.43	73.47	74.01	**74.11**
Caltech101	Base	96.84	98.00	97.96	98.02	97.72	97.87	97.74	**98.10**	97.97
	Novel	94.00	89.81	93.81	93.89	94.39	**95.10**	94.36	94.03	94.63
	HM	95.40	93.73	95.84	95.91	96.03	**96.46**	96.02	96.02	96.27
OxfordPets	Base	91.17	93.67	95.20	95.07	94.65	**95.63**	95.43	95.33	95.43
	Novel	97.26	95.29	97.69	97.63	97.76	97.40	97.76	97.30	**97.80**
	HM	94.12	94.47	96.43	96.33	96.18	96.51	96.58	96.30	**96.60**
StanfordCars	Base	63.37	78.12	70.49	77.68	71.76	70.33	72.94	**78.27**	77.73
	Novel	74.89	60.40	73.59	68.63	75.04	73.50	74.00	74.97	**75.43**
	HM	68.65	68.13	72.01	72.88	73.36	71.88	73.47	**76.58**	76.56
Flowers102	Base	72.08	97.60	94.87	95.54	95.00	95.10	95.92	**98.07**	97.77
	Novel	77.80	59.67	71.75	71.87	74.73	70.63	72.46	**76.50**	75.67
	HM	74.83	74.06	81.71	82.03	83.65	81.06	82.56	**85.95**	85.31
Food101	Base	90.10	88.33	90.70	90.37	90.50	90.70	90.71	90.67	**90.73**
	Novel	91.22	82.26	91.29	89.59	91.70	91.60	**92.05**	91.53	91.87
	HM	90.66	85.19	90.99	89.98	91.09	91.15	**91.38**	91.10	91.30
FGVCAircraft	Base	27.19	40.44	33.41	40.54	36.21	34.07	37.44	**42.73**	42.53
	Novel	36.29	22.30	23.71	27.57	33.55	24.17	35.61	**37.87**	36.33
	HM	31.09	28.75	27.74	32.82	34.83	28.28	36.50	**40.15**	39.19
SUN397	Base	69.36	80.60	79.74	81.26	80.29	79.65	80.82	82.67	**82.80**
	Novel	75.35	65.89	76.86	74.17	76.53	76.90	**78.70**	78.47	78.37
	HM	72.23	72.51	78.27	77.55	78.36	78.25	79.75	80.52	**80.52**
DTD	Base	53.24	79.44	77.01	77.35	77.55	73.90	80.36	83.37	**83.60**
	Novel	59.90	41.18	56.00	52.35	54.99	53.43	59.18	62.97	**63.00**
	HM	56.37	54.24	64.85	62.45	64.35	62.02	68.16	71.75	**71.85**
EuroSAT	Base	56.48	92.19	87.49	90.11	85.64	87.00	94.07	92.90	**96.10**
	Novel	64.05	54.74	60.04	60.89	64.34	66.30	73.23	73.90	**76.27**
	HM	60.03	68.69	71.21	72.67	73.48	75.25	82.35	82.32	**85.04**
UCF101	Base	70.53	84.69	82.33	84.33	82.89	82.20	83.00	**87.10**	86.40
	Novel	77.50	56.05	73.45	74.94	76.67	74.27	78.66	78.80	**80.40**
	HM	73.85	67.46	77.64	79.35	79.65	78.03	80.77	82.74	**83.29**

Table 2. Comparison with existing methods in the domain generalization task on four variants of ImageNet (ImNet). The advanced results are **bolded**.

Method	Source	Target				
	ImNet	Avg.	-V2	-SK.	-A	-R
CLIP	66.73	57.18	60.83	46.15	47.77	73.96
CoCoOP	71.02	59.91	64.07	48.75	50.63	76.18
ProGrad	**72.24**	59.07	64.73	47.61	49.39	74.58
KgCoOp	71.20	60.11	64.10	48.97	50.69	76.70
MaPLe	70.72	60.27	64.07	49.15	**50.90**	76.98
PromptSRC	71.27	60.65	64.35	49.55	**50.90**	77.80
PromptCD	72.10	**60.75**	**65.25**	**49.60**	50.35	**77.80**

4.3 Domain Generalization

To illustrate the performance of PromptCD in the domain generalization task which evaluates performance on target domains while training on the source domain, we treat ImageNet as the source domain, and regard other variants as target domains. Notably, source and target domains contain the same classes, but there are different distributions between them. Table 2 presents a comparison between PromptCD and existing methods in the domain generalization task. From this table, PromptCD achieves comparable performance with ProGras on the source domain, and performs 1.38% and 0.83% higher than MaPLe and PromptSRC. For target domains, PromptCD gets overall favourable performance, with an improvement of 1.68% over ProGra. Additionally, PromptCD outperforms all baseline methods on 3 out of 4 target domains. The significant results show powerful generalization of PromptCD for datasets with domain shifts.

4.4 Cross-dataset Recognition

In the above evaluation suites, the base-to-novel generalization task has similar dataset distributions between seen and unseen classes, and the domain generalization task has the same classes between the source domain and target domains. In order to further demonstrate the generalization and discrimination capabilities of PromptCD in the cross-dataset recognition task, we train PromptCD on the ImageNet dataset and evaluate the model on the irrelevant classification datasets. Table 3 shows a comparison between PromptCD and existing CLIP-based methods. From Table 3, the proposed PromptCD achieves the highest average result than all baseline methods, and provides a 0.54% improvement over PromptSRC. Moreover, in comparison with CoCoOP, MaPLe, and PromptSRC, PromptCD presents competitive results in most datatsets, and gets the best performance in 4 out of 10 datasets. The favorable results demonstrate the effectiveness of the proposed PtomptCD in capturing general knowledge.

5 In-depth Analysis

5.1 Ablation Study

In this section, we successively ablate different components of the proposed PromptCD model, including the bi-directional coupled-modality mechanism and mixture consistency, to comprehensively demonstrate the effectiveness of these components. Following previous methods [18, 19], the ablation experiments are implemented in the evaluation setting of base-to-novel generalization. Additionally,

Table 3. Comparison in the cross-dataset recognition of PromptCD with existing approaches on the remaining 10 datasets. The best performances are **bolded**.

	Source	Target										
	ImageNet	Average	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101
CoOP	71.51	63.88	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55
CoCoOP	71.02	65.74	**94.43**	90.14	65.32	71.88	86.06	22.94	**67.36**	45.73	45.37	68.21
MaPLe	70.72	66.30	93.53	**90.49**	65.57	**72.23**	86.20	24.74	67.01	46.49	48.06	68.69
PromptSRC	71.27	65.81	93.60	90.25	**65.70**	70.25	86.15	23.90	67.10	**46.87**	45.50	68.75
PromptCD	**72.10**	**66.35**	93.45	90.15	65.66	70.20	**86.25**	**25.30**	67.12	46.80	**49.10**	**69.50**

Table 4. Ablation study for critical components in the base-to-novel generalization task. Where "BCM" and "MC" denote the bi-directional coupled-modality mechanism and mixture consistency in the proposed PromptCD framework, repspectively.

BCM	MC	Base	Novel	HM
✓	✓	84.44	76.41	80.23
✓	×	83.95	73.36	78.30
×	✓	83.21	74.85	78.81
×	×	83.12	72.61	77.51

Table 5. Ablation study for PromptCD with different coupled-modality strategies. Where "TP" and "VP" denote the prompt projection from language to vision, and that from vision to language, respectively.

TP	VP	Base	Novel	HM
✓	✓	84.44	76.41	80.23
✓	×	84.31	76.12	80.01
×	✓	84.12	75.68	79.68
×	×	83.21	74.85	78.81

we explore the influence of the scale of the models on the generalization capability. Experimental results are presented in Table 4, Table 5 and Table 6, we can obtain the following findings:

- To understand the impact of each component, we conduct relevant experiments, as shown in Table 4. Typically, in the first row of this table, PromptCD achieves the highest performance, which provides a harmonic mean of 80.23%. Firstly, when ablating mixture consistency from PormptCD, the novel accuracy and harmonic mean are 3.05% and 1.93% drops, respectively. The reason is that mixture consistency enforces the model in learning the discrepancy between seen and unseen tasks. Such results emphasize the importance of mixture consistency in PromptCD. Subsequently, we remove the bi-directional coupled-modality mechanism from PromptCD, resulting in a 1.23% drop in the base accuracy. It highlights that the bi-directional coupled-modality mechanism can enhance the performance of the model in task-specific datasets. Finally, to illustrate the impact of the integrated components, we remove both components in the final row of Table 4. There is a serious performance degradation in the base-to-novel generalization task. This indicates that all the proposed components can effectively improve the generalization and discrimination capabilities of the model on seen and unseen tasks.
- To deeply demonstrate the effectiveness of the interaction between both vision and language branches, we perform a series of fine-grained ablation experiments on the bi-directional coupled-modality mechanism, as illustrated in Table 5. Typically, compared to the results of the final row of this table, any of the employed coupling functions can enhance the performance of the model in accuracy. Moreover, we make a comparison between "TP" and "VP". When removing "TP", the performance of this variant degrades more. The reason is that the language branch of CLIP can provide informative knowledge for VLMs on image classification. This is consistent with previous CLIP-based prompt learning methods. Notably, when ablating all coupling functions, the variant degrades as PromptCD without BCM. This result is the lowest accuracy, which is a 0.87% drop in the harmonic mean. Above all results clearly indicate that the bi-directional coupled-modality mechanism is beneficial for PromptCD on both seen and unseen tasks.

Table 6. Comparison of different prompt learning methods using various image encoders. † indicates the results produced by our re-implementation.

Method	Backbone	Base	Novel	HM
MaPLe	ViT-B/16	82.28	75.14	78.55
PromptSRC		84.26	76.10	79.97
PromptCD		84.44	76.41	80.23
MaPLe†	ViT-L/14	85.57	78.42	81.84(↑ 3.29)
PromptSRC†		87.22	81.58	84.31(↑ 4.34)
PromptCD		87.82	81.98	84.80(↑ 4.57)

- In order to demonstrate the performance of PromptCD on larger vision-language foundation models, we perform extended experiments on existing baseline methods by employing the ViT-L/14 backbone of CLIP [3], as shown in Table 6. From this table, we can obverse the following findings: (1) Diverse prompt learning methods achieve higher results by utilizing larger vision-language foundation models. These performance gains may be from the increasing numbers of backbone parameters, or tuning strategies. (2) PromptCD with the ViT-L/14 of CLIP achieves more improvements than MaPLe and PromptSRC with the same backbone. Such results further show that the proposed method can effectively improve the performance of larger vision-language models for downstream tasks.

5.2 Parameters Sensitivity

In this section, we give a sensitivity analysis of PromptCD about its critical parameters on the base-to-novel generalization task. First, we explore how the balancing factor α affects the performance of PromptCD. As shown in Figure 3, we can observe that each accuracy curve roughly presents a consistent tendency which first increases and then decreases. Specifically, PromptCD achieves the highest results when the value of α is set to 0.3. As $\alpha > 0.3$, the base accuracy has a slight fluctuation, and there is a degradation over the novel accuracy. The harmonic mean serves as a balance metric between both base and novel accuracy, which has a performance drop. Such results indicate that the balance factor α can influence the generalization of PromptCD between both seen and unseen tasks.

Next, as shown in Table 7, we investigate how the performance of PromptCD varies with different lengths and depths of prompts. Table 7(a) presents the impact of different lengths of prompts for

Table 7. Performance of PromptCD with different prompt settings.

(a) Length choices.

Length	Base	Novel	HM
1	79.21	73.23	76.10
2	82.21	73.56	77.64
3	83.56	75.41	79.28
4	84.44	76.41	80.23
5	83.17	75.32	79.05

(b) Depth choices.

Depth	Base	Novel	HM
1	81.81	74.52	78.00
3	81.91	74.62	78.10
6	82.35	76.19	79.15
9	84.44	76.41	80.23
12	84.21	76.01	79.90

Table 8. Comparison of learnable parameters for different approaches.

Method	Learnable parameters	Base	Novel	HM
CLIP	-	69.34	74.22	71.70
CoCoOp	35360	80.47	71.69	75.83
MaPLe	3.55M	82.28	75.14	78.55
PromptSRC	46080	84.26	76.10	79.97
PromptCD	412416	84.44	76.41	80.23

PromptCD, we can observe that as the length of prompts increases, the overall average performance increases. Typically, we get the highest performance when the length of prompts is set as 4. The harmonic mean does not increase for the prompt length greater than 4. Table 7(b) shows the influence of different depths of prompts for PromptCD. From this table, we find that PromptCD with 9 prompt layers achieves the best performance. Notably, differing from previous works, PromptCD improves the base accuracy while maintaining the novel accuracy, which benefits from cross-modal interactions and consistency constraints. This indicates that PromptCD has robust generalization capability to unseen classes.

Additionally, we present an analysis of the computational complexity of PromptCD compared with other methods, as depicted in Table 8. Typically, MaPLe with more learnable parameters does not achieve significant improvements compared with the novel accuracy of zero-shot CLIP. The reason may be that MaPLe utilizes prompt projection to enable the model to sufficiently learn task-specific features, and lacks the generalization to unseen tasks. This illustrates that anti-overfitting strategies should be explored to enhance the performance of the models on unseen tasks. Although PromptSRC has a few learnable parameters apart from zero-shot CLIP, its performance improvements are more than MaPLe. We think the reason why PromptSRC outperforms MaPLe is that consistency constraints can effectively learn complementary knowledge between task-specific and task-agnostic features. However, PromptSRC ignores the prompt interaction between branches, which is limited by the semantic consistency across modalities. Our proposed PromptCD integrates the bi-directional coupled-modality mechanism and mixture consistency to intensify cross-modality interaction and enhance the generalization and discrimination for adapting VLMs to image classification.

5.3 Representation Visualization

To further demonstrate the generalization and discriminative abilities of PromptCD between seen and unseen classes, we enforce t-SNE [41] to visualize the image representations of PromptCD and other traditional approaches, like MaPLe and PromptSRC, on the EuroSAT dataset. In Figure 4, in regard to seen classes, PromptCD presents a powerful discrimination between classes, which can pull consistency representations close and push those inconsistency representations apart. Such results benefit from sufficient interactions between modalities. On the other hand, unseen classes are challenging for almost all approaches. However, PromptCD still maintains a good performance compared to other baseline approaches. For example, the marked red circle in Figure 4 indicates that PromptCD can separate the subtle discrepancy between new classes. The reason is that mixture consistency can enhance the generalization of the model between seen and unseen classes. These plots clearly present that PromptCD has strong generalization and discrimination.

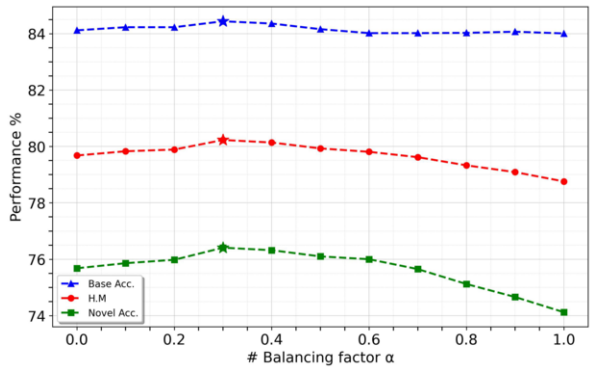

Figure 3. Performance of PromptCD for different values of the balancing factor α. Symbol $\star$ represents the best results.

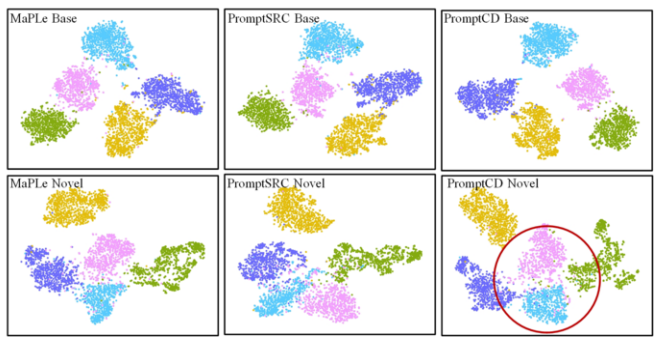

Figure 4. Visualization of the image representations obtained by MaPLe, PromptSRC, and PromptCD on the EuroSAT. The first and second rows represent the results of all methods on seen and unseen tasks, respectively.

6 Conclusion

In this work, we propose a novel coupled and decoupled prompt learning framework for vision-language foundation models, which aims to improve the performance of the models for both seen and unseen tasks. PromptCD is a well-designed approach with two critical components that establish vision-language interaction and enhance generalization for various downstream tasks. Extensive experiments are conducted across three evaluation settings, such as base-to-novel and domain generalization, and cross-dataset recognition, demonstrate the effectiveness of our proposed approach. Moreover, ablation analysis comprehensively confirms the impact of each component for adapting vision-language models to image classification.

In the future, we would like to extend our proposed prompt learning method to fine-tune other vision-language models for a range of multimodal applications. The reason is that the alignment between modalities or features is a common challenge in these applications.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62076173), the High-level Entrepreneurship and Innovation Plan of Jiangsu Province (No. JSSCRC2021524), and the Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

[1] H. Bao, W. Wang, L. Dong, Q. Liu, O. K. Mohammed, K. Aggarwal, S. Som, S. Piao, and F. Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. In *NIPS*, pages 32897–32912, 2022.

[2] L. Bossard, M. Guillaumin, and L. Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, pages 446–461, 2014.

[3] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, pages 2818–2829, 2023.

[4] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. Describing textures in the wild. In *CVPR*, pages 3606–3613, 2014.

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[6] J. Ding, N. Xue, G.-S. Xia, and D. Dai. Decoupling zero-shot semantic segmentation. In *CVPR*, pages 11583–11592, 2022.

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

[8] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR*, pages 178–178, 2004.

[9] H. Gao, H. Zhang, X. Yang, W. Li, F. Gao, and Q. Wen. Generating natural adversarial examples with universal perturbations for text classification. *Neurocomputing*, 471:175–182, 2022.

[10] P. Gao, J. Han, R. Zhang, Z. Lin, S. Geng, A. Zhou, W. Zhang, P. Lu, C. He, X. Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.

[11] C. Han, Q. Wang, Y. Cui, Z. Cao, W. Wang, S. Qi, and D. Liu. E2vpt: An effective and efficient approach for visual prompt tuning. In *ICCV*, pages 17445–17456, 2023.

[12] P. Helber, B. Bischke, A. Dengel, and D. 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.

[13] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, pages 8340–8349, 2021.

[14] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2021.

[15] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916, 2021.

[16] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim. Visual prompt tuning. In *ECCV*, pages 709–727, 2022.

[17] R. Karimi Mahabadi, J. Henderson, and S. Ruder. Compacter: Efficient low-rank hypercomplex adapter layers. In *NIPS*, pages 1022–1035, 2021.

[18] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan. Maple: Multi-modal prompt learning. In *CVPR*, pages 19113–19122, 2023.

[19] M. U. Khattak, S. T. Wasim, M. Naseer, S. Khan, M.-H. Yang, and F. S. Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *ICCV*, pages 15190–15200, 2023.

[20] G. Kim, S. Kim, and S. Lee. Aapl: Adding attributes to prompt learning for vision-language models. In *CVPR*, pages 1572–1582, 2024.

[21] W. Kim, B. Son, and I. Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, pages 5583–5594, 2021.

[22] J. Krause, M. Stark, J. Deng, and L. Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV*, pages 554–561, 2013.

[23] D. Lee, S. Song, J. Suh, J. Choi, S. Lee, and H. J. Kim. Read-only prompt optimization for vision-language few-shot learning. In *ICCV*, pages 1401–1411, 2023.

[24] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, pages 121–137, 2020.

[25] X. L. Li and P. Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *ACL-IJCNLP*, pages 4582–4597, 2021.

[26] Y. Li, F. Liang, L. Zhao, Y. Cui, W. Ouyang, J. Shao, F. Yu, and J. Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *ICLR*, 2021.

[27] Z. Li, X. Li, X. Fu, X. Zhang, W. Wang, S. Chen, and J. Yang. Promptkd: Unsupervised prompt distillation for vision-language models. In *CVPR*, pages 26617–26626, 2024.

[28] X. Liu, W. Tang, J. Lu, R. Zhao, Z. Guo, and F. Tan. Deeply coupled cross-modal prompt learning. In *ACL*, pages 7957–7970, 2023.

[29] J. Lu, D. Batra, D. Parikh, and S. Lee. Vilbert: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NIPS*, pages 13–23, 2019.

[30] M. Maaz, H. Rasheed, S. Khan, F. S. Khan, R. M. Anwer, and M.-H. Yang. Class-agnostic object detection with multi-modal transformer. In *ECCV*, pages 512–531, 2022.

[31] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[32] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, pages 722–729, 2008.

[33] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar. Cats and dogs. In *CVPR*, pages 3498–3505, 2012.

[34] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021.

[35] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, pages 5389–5400, 2019.

[36] S. Roy and A. Etemad. Consistency-guided prompt learning for vision-language models. In *ICLR*, 2024.

[37] K. Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *NIPS*, volume 29, 2016.

[38] K. Soomro, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[39] S. Subramanian, W. Merrill, T. Darrell, M. Gardner, S. Singh, and A. Rohrbach. Reclip: A strong zero-shot baseline for referring expression comprehension. In *ACL*, pages 5198–5215, 2022.

[40] H. Sun, X. He, J. Zhou, and Y. Peng. Fine-grained visual prompt learning of vision-language models for image recognition. In *ACM MM*, pages 5828–5836, 2023.

[41] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, Ł. Gomez, and I. Polosukhin. Attention is all you need. In *NIPS*, 2017.

[43] A. Wang, H. Chen, Z. Lin, Z. Ding, P. Liu, Y. Bao, W. Yan, and G. Ding. Hierarchical prompt learning using clip for multi-label classification with single positive labels. In *ACM MM*, pages 5594–5604, 2023.

[44] F. Wang, M. Li, X. Lin, H. Lv, A. Schwing, and H. Ji. Learning to decompose visual features with latent textual prompts. In *ICLR*, 2022.

[45] H. Wang, S. Ge, Z. Lipton, and E. P. Xing. Learning robust global representations by penalizing local predictive power. In *NIPS*, pages 10506–10518, 2019.

[46] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492, 2010.

[47] C. Xu, Y. Zhu, G. Zhang, H. Shen, Y. Liao, X. Chen, G. Wu, and L. Wang. Dpl: Decoupled prompt learning for vision-language models. *arXiv preprint arXiv:2308.10061*, 2023.

[48] C. Yang, F. Meng, S. Chen, M. Liu, and R. Zhang. Instance-wise adaptive tuning and caching for vision-language models. In *ECAI*, pages 2834–2841. 2023.

[49] H. Yao, R. Zhang, and C. Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *CVPR*, pages 6757–6767, 2023.

[50] L. Yao, R. Huang, L. Hou, G. Lu, M. Niu, H. Xu, X. Liang, Z. Li, X. Jiang, and C. Xu. Filip: Fine-grained interactive language-image pre-training. In *ICLR*, 2021.

[51] F. Yu, J. Tang, W. Yin, Y. Sun, H. Tian, H. Wu, and H. Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *AAAI*, pages 3208–3216, 2021.

[52] J. Zhang, S. Wu, L. Gao, H. T. Shen, and J. Song. Dept: Decoupled prompt tuning. In *CVPR*, pages 12924–12933, 2024.

[53] R. Zhang, Z. Jiang, Z. Guo, S. Yan, J. Pan, H. Dong, P. Gao, and H. Li. Personalize segment anything model with one shot. In *ICLR*, 2023.

[54] K. Zhou, J. Yang, C. C. Loy, and Z. Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16816–16825, 2022.

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

[56] B. Zhu, Y. Niu, Y. Han, Y. Wu, and H. Zhang. Prompt-aligned gradient for prompt tuning. In *ICCV*, pages 15659–15669, 2023.