BMIP: Bi-directional Modality Interaction Prompt Learning for VLM

Song-Lin Lv[1,3] , **Yu-Yang Chen**[1,3] , **Zhi Zhou**[3] , **Ming Yang**[2,3] and **Lan-Zhe Guo**[*,1,3]

[1]School of Intelligence Science and Technology, Nanjing University, China
[2]School of Artificial Intelligence, Nanjing University, China
[3]National Key Laboratory for Novel Software Technology, Nanjing University, China
{lvsl,chenyy,zhouz,yangm,guolz}@lamda.nju.edu.cn

Abstract

Vision-language models (VLMs) have exhibited remarkable generalization capabilities, and prompt learning for VLMs has attracted great attention for the ability to adapt pre-trained VLMs to specific downstream tasks. However, existing studies mainly focus on single-modal prompts or uni-directional modality interaction, overlooking the powerful alignment effects resulting from the interaction between the vision and language modalities. To this end, we propose a novel prompt learning method called **B**i-directional **M**odality **I**nteraction **P**rompt (BMIP), which dynamically weights bi-modal information through learning the information of the attention layer, enhancing trainability and inter-modal consistency compared to simple information aggregation methods. To evaluate the effectiveness of prompt learning methods, we propose a more realistic evaluation paradigm called *open-world generalization* complementing the widely adopted *cross-dataset transfer* and *domain generalization* tasks. Comprehensive experiments on various datasets reveal that BMIP not only outperforms current state-of-the-art methods across all three evaluation paradigms but is also flexible enough to be combined with other prompt-based methods for consistent performance enhancement.

1 Introduction

Vision-language models, such as CLIP [Radford *et al.*, 2021], pre-trained on vast web-scale text-image datasets, have demonstrated impressive zero-shot capabilities in a variety of downstream image classification tasks [Jia *et al.*, 2021; Radford *et al.*, 2021]. Concurrently, various prompt learning methods for VLMs have been proposed to further improve the performance of VLMs on specific tasks by exploiting a small amount of labeled data for the target task to finetune learnable prompts while keeping other parameters frozen [Jia *et al.*, 2022; Khattak *et al.*, 2023a; Zhou *et al.*, 2022a; Zhou *et al.*, 2022b].

---
[*]Corresponding author

Previous methods mainly learn prompts on a single modality, either language (Figure 1:a) or vision (Figure 1:b), overlooking the interaction between the two modalities, which is crucial for preserving alignment between the vision and language. Many studies [Zang *et al.*, 2022; Derakhshani *et al.*, 2023; Lee *et al.*, 2023] have reported that language prompt learning underperforms on datasets with high intra-class visual variances, while vision prompt learning struggles on datasets with small inter-class textual variances. For example, the EuroSAT dataset exhibits high intra-class visual variances, which is challenging for text-only methods to handle effectively. There are also some studies [Rao *et al.*, 2022; Kan *et al.*, 2023; Khattak *et al.*, 2023a] that attempt to transfer prompts from language to vision modality (Figure 1:c) to exploit modality interaction. However, these methods only achieve alignment provided by language prompts alone and fail to fully consider the effect of the interaction between the two modalities. To the best of our knowledge, *how to exploit the bi-directional interaction between the two modalities in VLMs remains a significant yet unsolved challenge*.

To address the lack of focus on multi-modal consistency in single-modal interaction methods, we propose a novel approach called **B**i-directional **M**odality **I**nteraction **P**rompt (BMIP), as illustrated in Figure 1:d. Given the absence of existing studies on multi-modal interaction in the VLM field and the limitations of simple aggregation functions—such as low information utilization and potential information distortion—we designed an innovative aggregation function for multi-modal interaction. This function leverages the relationship between the model's attention layer outputs and prompt importance, dynamically balancing dual-modal information through adaptive weighting to solve the above difficulties.

To evaluate the effectiveness of prompt learning methods in VLMs, inspired by open-world settings [Zareian *et al.*, 2021; Lv *et al.*, 2025], we propose a novel evaluation paradigm called *open-world generalization*. Different from the previous *base-to-new class generalization* task [Zhou *et al.*, 2022b; Khattak *et al.*, 2023a; Yao *et al.*, 2023], which evaluates base and new classes separately, the new paradigm does not predetermine whether the data belongs to base or new classes, leading to a more realistic evaluation. Experimental results on 15 benchmarks demonstrated that the BMIP architecture achieves significant performance improvement compared to state-of-the-art (SOTA) prompt learning methods, especially in
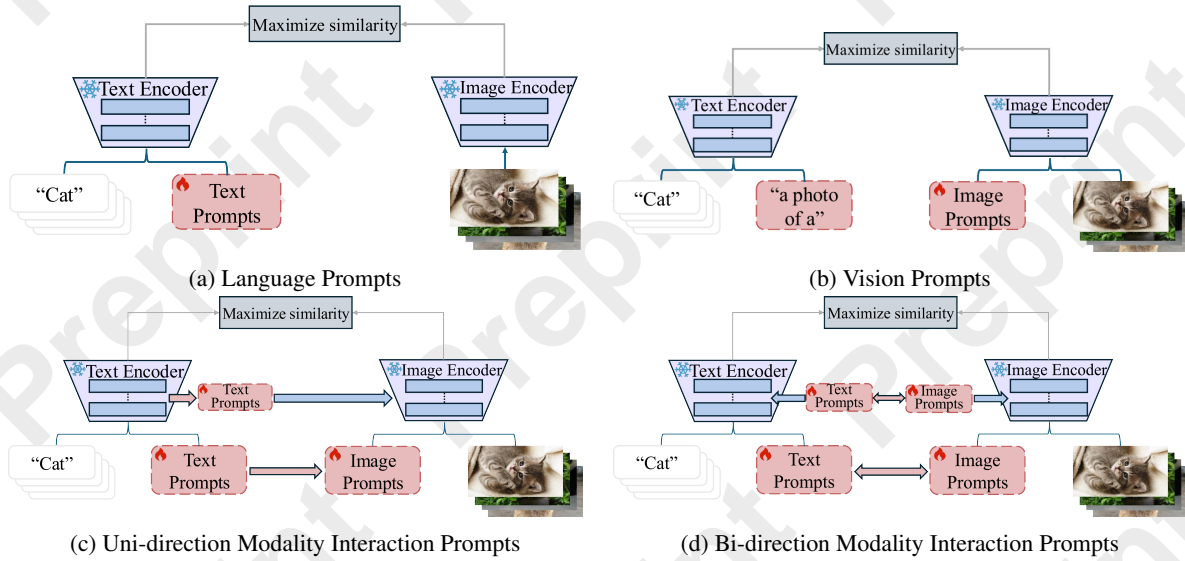
Figure 1: (a) Adapting language modality representations for downstream tasks through language prompt learning. (b) Adapting vision modality representations for downstream tasks through vision prompt learning. (c) Achieving uni-directional modality interaction by converting language prompts into vision prompts. (d) Achieving bi-directional modality interaction by aligning two modalities' representations through aggregating information between the vision and language modalities. (🔥: learnable; ❄: frozen during training)

open-world generalization evaluation paradigm. Specifically, BMIP addresses the poor performance of single-modal prompt learning methods when dealing with unbalanced datasets of text and images, such as EuroSAT [Helber *et al.*, 2019] and Flowers102 [Nilsback and Zisserman, 2008], which is in line with the motivation. Additionally, as an enhanced algorithm for MaPLe, BMIP can serve as a foundational framework that integrates with any other prompt learning methods to further improve their performance.

In summary, the main contributions of this work include:

- We analyze the limitations of previous prompt learning methods and propose a novel technique that leverages bi-directional modality interaction, which enhances the alignment between vision and language modalities and paves the way for further exploration of information aggregation in other multi-modal models.
- To evaluate prompt learning methods, we propose a more realistic evaluation paradigm called *open-world generalization*. We believe this could facilitate more realistic evaluations of prompt learning methods and promote related research.
- We conduct comprehensive experiments on 15 benchmarks. The results demonstrated that BMIP achieves SOTA performance across all tasks, and is flexible enough to be combined with other prompt learning methods, consistently enhancing their performance.

2 Related Works

Vision-Language Models. VLMs [Radford *et al.*, 2021; Alayrac *et al.*, 2022; Tan *et al.*, 2025] have demonstrated outstanding performance on a wide range of image classification tasks under zero-shot and few-shot settings. By utilizing a substantial amount of text-image data, these models leverage the common embedding space of text and image representations to achieve cross-modal alignment effectively. While these pre-trained VLMs have learned generalized representations of images and texts, adapting them quickly and effectively to downstream tasks remains a challenging problem. Existing approaches exploring finetuning VLMs to downstream tasks using a small number of parameters and data can be categorized into two types: adapter tuning [Zhang *et al.*, 2021; Li *et al.*, 2023; Yu *et al.*, 2023] and prompt learning [Zhou *et al.*, 2022b; Zhou *et al.*, 2022a; Khattak *et al.*, 2023a] which is the focus of our work.

Prompt Learning in Vision Language Models. Due to the large parameter size of VLMs and the limited availability of training data for downstream tasks, it is impractical to fine-tune all parameters of the VLMs to adapt them to these tasks. Inspired by the success of prompt learning in NLP [He *et al.*, 2022; Li and Liang, 2021], many researchers have proposed to adapt VLMs by learning the prompts in end-to-end training. As the pioneering work, CoOp [Zhou *et al.*, 2022a] for the first time introduces the learnable prompts to transfer the task-specific knowledge to VLMs. To improve the generalization of the learnable language prompts in CoOp, CoCoOp [Zhou *et al.*, 2022b] and VPT [Jia *et al.*, 2022] generate a vision-conditional prompt by fusing the image feature and the learnable language prompts. KgCoOp [Yao *et al.*, 2023], and other methods [Zhou *et al.*, 2024] are other prompt-based methods for VLMs. MaPLe [Khattak *et al.*, 2023a], conducts language-vision prompt learning by jointly applying prompt learning to both the vision and text encoders, simultaneously refining the text and image representations for adapting to downstream tasks. Our proposed method focuses on addressing the imbalance in alignment between the vision and language modalities caused by the lack of interaction in existing algorithms.

3 Preliminary

Our approach BMIP is proposed based on CLIP, learning both language and vision prompts. Therefore, before introducing the proposal, we revisit the main ideas of CLIP, language prompt learning, and vision prompt learning.

CLIP. CLIP is developed to align visual and textual data in a common embedding space. CLIP consists of two encoders: an image encoder denoted as f and a text encoder denoted as g. During the training phase, the encoders extract feature representations $f(I)$ and $g(E_w(T))$ from an input image I and its corresponding text caption T, respectively. The term E_w represents the word embedding layer, tasked with transforming words into vector representations.

During the zero-shot classification phase, CLIP begins with an image I and a set of hand-designed text captions $[T_1, T_2, \ldots, T_N]$, formatted as "*a photo of a* $[CLASS_i]$", where "*a photo of a*" is a hand-designed template and $[CLASS_i]$ specifies a class from N candidate image categories. The image and captions are processed by their respective encoders to extract features, allowing for the computation of class prediction probabilities as follows:

$$p(y = i|I) = \frac{\exp(cos(f(I), g(E_w(T_i)))/\tau)}{\sum_{j=1}^{N} \exp(cos(f(I), g(E_w(T_j)))/\tau)} \quad (1)$$

In this context, τ is the temperature coefficient, and $cos(\cdot, \cdot)$ represents the cosine similarity between features.

Language Prompt Learning. To effectively adapt VLMs to downstream tasks, language prompt learning aims to generate more adaptive classifiers, without the need to finetune the text encoder g. For instance, some studies [Lu *et al.*, 2022; Zhou *et al.*, 2022a; Zhou *et al.*, 2022b] employ learnable prompts $P = [t_1, t_2, \ldots, t_b]$ to replace hand-designed language prompt templates, where t represents the prompt vector, and b specifies the prompt's length. Let $[c_1, c_2, \ldots, c_N]$ represent the word embeddings of class names. The corresponding prediction probability is calculated as follows:

$$p(y = i|I) = \frac{\exp(cos(f(I), g([P, c_i]))/\tau)}{\sum_{j=1}^{N} \exp(cos(f(I), g([P, c_j]))/\tau)} \quad (2)$$

In this context, $[\cdot, \cdot]$ denotes the operation of concatenation. For each downstream task, the learnable prompts P are optimized via cross-entropy classification loss during the few-shot learning phase. As a result, updating the language prompts P will adjust the decision boundaries accordingly, utilizing the generated classifier for the downstream tasks.

Vision Prompt Learning. Analogous to the language modality, vision prompt learning incorporates vision prompt vectors $\tilde{P}$ to extract more representative visual features. To associate images with these vectors, the image embedding layer transforms image patches $I = [I_1, I_2, \ldots, I_m]$ into image patch embeddings $E = [e_1, e_2, \ldots, e_m]$, with m represents the number of image patches. The output from the ith layer of image encoder f_i is expressed by the equation:

$$\left[CLS_i, E_i, \tilde{P}_i \right] = f_i \left(\left[CLS_{i-1}, E_{i-1}, \tilde{P}_{i-1} \right] \right) \quad (3)$$

Here, CLS represents the image class token, which will be projected by a projection head to obtain the image feature.

4 Bi-directional Modality Interaction Prompt

Many studies [Derakhshani *et al.*, 2023; Lee *et al.*, 2023; Zang *et al.*, 2022] have reported that language prompt learning underperforms on datasets with high intra-class visual variances while vision prompt learning struggles on datasets with small inter-class textual variances. Recognizing the substantial influence of the multi-modal interaction, we propose the BMIP method from three intuitions: (1) Independent vision and language prompts facilitate the collection of information from their respective modalities, while projected prompts are substitutable for original prompts; (2) prompts of suitable depth can expand the scope of prompt information and curtail overfitting; and (3) effective interaction between modalities will mitigate the drawbacks of imbalanced single-modal information and promote better alignment between vision and language modalities. Therefore, the proposed BMIP is composed of three key components: deep language prompt learning, deep vision prompt learning, and vision language modality interaction. Figure 2 illustrates the overall architecture of the BMIP framework, and we describe the details in the following. In the end, we analyze the rationale behind BMIP's modality interaction.

4.1 Deep Language Prompt Learning

In contrast to traditional language prompt learning, deep language prompt learning introduces layered prompts to expand the scope of prompt information. Specifically, we introduce layered language prompts $\{P_i \in \mathbb{R}^{1 \times b \times d_l}\}_{i=0}^{J}$, where J, b, and d_l indicates the depth, length, and dimension of the language prompts, respectively. The input to the initial layer assumes the structure $[P_0, W_0]$, where $\{W_0 \in \mathbb{R}^{N \times x \times d_l}\}$ signifies the word embedding of the text T. Here, x represents the number of words in a single caption from T, with N indicating the total count of image categories. To provide comprehensive guidance on the word embeddings W, we deploy language prompts to supersede prompts from prior layers in the first J layers of the text encoder. The inputs and outputs at the ith layer of the text encoder g are as follows:

$$[\_, W_i] = g_i \left([P_{i-1}, W_{i-1}] \right) \quad i = 1, 2, \ldots, J \quad (4)$$

Beyond the Jth layer, to prevent the model from excessive reliance on learnable prompts and overfitting, we use the prompts from the output of the preceding layer as the input to the next layer. The class feature z is obtained by projecting the class representation $[c_1, c_2, \ldots, c_N]$ corresponding to the last output of the text encoder W_K to a common embedding space via the text projection head TextProj.

$$[P_j, W_j] = g_j \left([P_{j-1}, W_{j-1}] \right) \quad j = J+1, \ldots, K \quad (5)$$

$$z = \text{TextProj} \left([c_1, c_2, \ldots, c_N] \right) \quad (6)$$

4.2 Deep Vision Prompt Learning

To gather information about the dataset and interact with the language modality, we introduce independent visual prompts. Deep vision prompt learning employs a set of vision prompts $\{\tilde{P}_i \in \mathbb{R}^{1 \times b \times d_v}\}_{i=0}^{J}$, which match the depth and length of language branch, differing solely in the dimension of the vision prompts d_v. In the first J^{th} layers of the image encoder f, we use a learnable vision prompt to replace the output of the
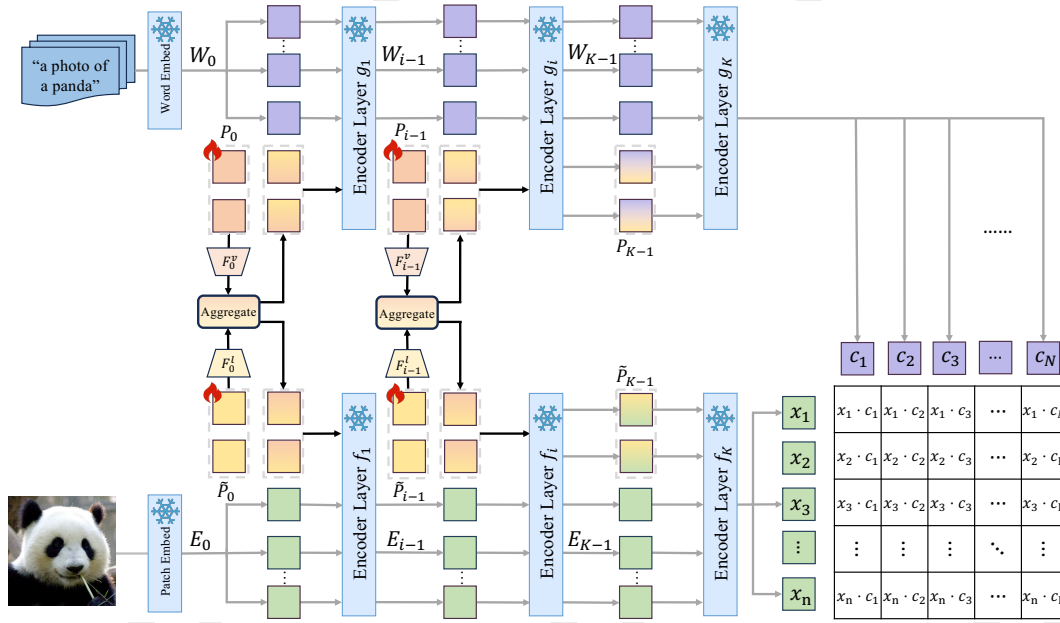
Figure 2: Overview of proposed BMIP method. BMIP finetunes the layered prompts for the vision and language branches while freezing the rest of the model parameters. It utilizes an aggregation function to modulate cross-modal prompt influences, thus enabling effective information exchange between the two modalities.

previous layer. These prompts are then concatenated with the class token CLS and the image patch embeddings E, forming the input for the next layer, as shown below:

$$[CLS_i, E_i, \_] = f_i\left(\left[CLS_{i-1}, E_{i-1}, \tilde{P}_{i-1}\right]\right) \quad (7)$$
$$i = 1, 2, \ldots, J$$

After the Jth layer, similar to the language branch, the ensuing layer's input is the immediate output of its predecessor. Upon obtaining the ultimate class token CLS_K, the image projection head, denoted as $\mathrm{ImageProj}$, is employed to map the final image feature x to the common embedding space. The formula from the $(J+1)^{th}$ layer is as follows:

$$\left[CLS_j, E_j, \tilde{P}_j\right] = f_j\left(\left[CLS_{j-1}, E_{j-1}, \tilde{P}_{j-1}\right]\right) \quad (8)$$
$$j = J+1, \ldots, K$$

$$x = \mathrm{ImageProj}\left(CLS_K\right) \quad (9)$$

4.3 Vision Language Modality Interaction

As a novel multimodal interaction approach to prompt learning, it is essential for BMIP to design with an interactive architecture that promotes effective information aggregation. Under this objective, this architecture consists of three key components: a language projection head F_l, a vision projection head F_v, and an aggregation function.

The most straightforward idea for aggregating information might include adding the aggregated information to the original prompts using unit weights or attention weights or simply connecting this information. Unfortunately, these methods driven by linear addition and modifications either fail to allocate the importance of the two modality inputs effectively

or cannot easily highlight the internal impact of the prompts, making efficient information aggregation unattainable for effective information aggregation from both modalities.

To solve this problem, we propose a learnable aggregation function using the output weight of the attention layer for effective information aggregation from both modalities. This aggregation function uses learnable modules to dynamically generate weights for different modality prompts, giving greater emphasis to the prompts that the attention layer focuses on. Specifically, the vision and language projection heads produce transformed vision and language information, represented as $\{F_v(P_i), F_l(\tilde{P}_i)\}$. The vision and language attention weights (w_v, w_l) are extracted from the output of the current attention layer, representing the degree of attention given by other inputs to the current prompt. Subsequently, we trained modality-specific 1×1 linear layers, L_l and L_v, to learn the relationship between attention weights and substitution weights (w_l, w_v), ultimately generating unique dynamic weights for each prompt. The formal expression is as follows:

$$w_v = L_v(A_v),$$
$$\tilde{P}'_i = w_v * \tilde{P}_i + (1 - w_v) * F_v(P_i). \quad (10)$$

$$w_l = L_l(A_l),$$
$$P'_i = \left[w_l * P_i + (1 - w_l) * F_l(\tilde{P}_i)\right]. \quad (11)$$

where $\{P'_i|_{i=1}^J\}$ and $\{\tilde{P}'_i|_{i=1}^J\}$ represent augmented deep language and vision prompts. After implementing the aggregation function, the first J layers of BMIP process inputs and outputs for the language and vision modalities as described subsequently:

$$[\_, W_i] = g_i\left([P'_i, W_{i-1}]\right) \quad (12)$$

$$[CLS_i, E_i, \_] = f_i\left(\left[CLS_{i-1}, E_{i-1}, \tilde{P}'_i\right]\right) \quad (13)$$

Beyond the Jth layer, subsequent layers use the output of the predecessor as input to obtain the final prediction. Using the proposed aggregation function, vision and language encoders can receive weighted prompt input, facilitating alignment of vision and language. Additionally, the learnable linear layers eliminate the need for manually tuning prompt weights, ensuring high reliability and greater flexibility. Furthermore, BMIP can enhance existing prompt fine-tuning methods by serving as a more powerful foundational model than MaPLe, leveraging its effectively balanced multimodal prompts.

4.4 Analysis

CoCoOp and subsequent prompt learning methods demonstrate that for well-aligned pre-trained VLMs, fine-tuned prompts are interchangeable. For instance, CoCoOp incorporates image information into text prompts, while MaPLe uses transformed language prompts as vision prompts, which reflects that vision and language prompts can be simultaneously optimized and mutually substituted, as demonstrated in the Ablation Studies.

In practice, especially when approaching the convergence point, the magnitude of $\frac{\partial L}{\partial w}$ is usually very close to zero, indicating that the probability of staying around w is large. In other words, when the attention weight is equal to zero, this prompt will almost become redundant during the later training process. Therefore, using w to combine the original prompt with aggregated information will enhance the trainability of the model and facilitate modal alignment. We immediately have the following corollary,

Corollary 1. *If the minimal of weights implies $w = 0$, then the prompt combining will only decrease the training loss, i.e. $\min_{P'} L \leq \min_P L$, given the sufficiently expressive $\min_{P'} L$ and $\min_P L$ which denote the cases with and without prompt combining, respectively.*

These analysis suggests that replacement prompts enhance BMIP's trainability, maintaining image text alignment while adapting to downstream tasks.

5 Experiments

5.1 Experimental Setup

Evaluation Paradigms. There are three widely adopted tasks to evaluate prompt learning methods: *generalization from base to novel classes, cross-dataset transfer, and domain generalization* [Zhou *et al.*, 2022a; Khattak *et al.*, 2023a; Lee *et al.*, 2023; Yao *et al.*, 2023]. Generalization from base to novel classes task evaluates the performance of models on base and new classes separately. Although this evaluation paradigm can comprehensively evaluate the performance of both base and new classes, it lacks practicality for real-world applications since downstream tasks cannot determine whether the data belongs to base or new classes in advance, which is a major metric of a model's ability to generalize. Therefore, we propose a more realistic evaluation paradigm, termed *open-world generalization*, which introduces a metric used in open-world settings [Zareian *et al.*, 2021; Lv *et al.*, 2025; Guo *et al.*, 2025; Li *et al.*, 2019], the simultaneous evaluation

of base classes and new classes, to the generalization from base to novel classes task. This new paradigm assesses the model's performance on an unknown distribution composed of both base and novel classes. The *cross-dataset transfer* task demonstrates the model's zero-shot generalization to novel datasets, while the *domain generalization* task assesses the model's robustness to out-of-distribution data.

Datasets. We assessed 11 different image classification datasets in the open-world generalization and cross-dataset transfer tasks, encompassing ImageNet [Deng *et al.*, 2009], Caltech101 [Fei-Fei *et al.*, 2004], OxfordPets [Parkhi *et al.*, 2012], StanfordCars [Krause *et al.*, 2013], Flowers102 [Nilsback and Zisserman, 2008], Food101 [Bossard *et al.*, 2014], FGVC-Aircraft [Maji *et al.*, 2023], SUN397 [Xiao *et al.*, 2010], UCF101 [Soomro *et al.*, 2012], DTD [Cimpoi *et al.*, 2014], and EuroSAT [Helber *et al.*, 2019]. For domain generalization task, ImageNet serves as our source dataset, while 4 variants ImageNetV2 [Recht *et al.*, 2019], ImageNet-Sketch [Wang *et al.*, 2019], ImageNet-A [Hendrycks *et al.*, 2021b], and ImageNet-R [Hendrycks *et al.*, 2021a] serve as the target datasets.

Compared Methods. There are several methods similar to BMIP that use prompts exclusively to finetune VLMs, such as CoOp, ProDA, and MaPLe; therefore, we focus on benchmarking against four current representative methods: CLIP, CoOp, CoCoOp, and MaPLe, the SOTA method. It is worth noting that BMIP is orthogonal to studies that employ regularization techniques and can be integrated with these methods to enhance prompt learning optimization. We report the performance improvements brought by BMIP in ablation studies. The **implementation details** are in Appendix A.

5.2 Open-World Generalization Task

To assess the open-world robustness of the BMIP approach, the average performance across all datasets, as well as the detailed performance on each dataset measured by two metrics, i.e., the harmonic mean (HM) of the accuracies for the base and new classes and Accuracy, are reported. The average performance and variance under three runs are presented in Table 1. Compared to the SOTA method MaPLe, BMIP demonstrates performance improvements on both base and novel classes across most datasets. In the remaining three datasets, the BMIP method achieves competitive performance compared to the MaPLe method and delivers the best average performance on both the HM and Accuracy metrics. BMIP shows an absolute average performance gain of 0.82% over MaPLe, when considering both base and novel classes simultaneously. Additionally, BMIP performs exceptionally well on datasets with imbalanced text and image information [Nilsback and Zisserman, 2008; Helber *et al.*, 2019; Zang *et al.*, 2022], such as those with high intra-class visual variance (e.g., EuroSAT, from 81.43% to 86.10%) and low inter-class text variance (e.g., Flowers102, from 82.78% to 83.86%), reflecting our motivation that BMIP can overcome the shortcomings of single-modal prompt learning methods. The above results highlight the crucial role of the independent prompts of each modality and the strong alignment between vision and language in enhancing generalization capacity.

	Average		ImageNet		Caltech101		OxfordPets	
	HM	Acc.	HM	Acc.	HM	Acc.	HM	Acc.
CLIP	70.84	63.92	70.20 ± 0.00	66.73 ± 0.00	95.41 ± 0.00	92.90 ± 0.00	92.93 ± 0.00	88.03 ± 0.00
CoOp	72.14	65.57	64.95 ± 1.11	61.79 ± 1.09	95.96 ± 0.39	93.24 ± 0.68	95.38 ± 0.33	89.61 ± 0.34
CoCoOp	74.72	67.67	72.71 ± 0.33	69.41 ± 0.36	95.55 ± 0.24	93.43 ± 0.37	95.71 ± 0.76	90.24 ± 1.32
MaPLe	78.22	71.76	$\mathbf{73.60 \pm 0.12}$	$\mathbf{70.30 \pm 0.16}$	96.47 ± 0.41	94.67 ± 0.33	$\mathbf{96.68 \pm 0.12}$	$\mathbf{92.60 \pm 0.49}$
BMIP	$\mathbf{79.04}$	$\mathbf{72.17}$	73.47 ± 0.11	70.23 ± 0.06	$\mathbf{96.54 \pm 0.25}$	$\mathbf{94.73 \pm 0.23}$	96.40 ± 0.19	92.53 ± 0.32

	StandfordCars		Flowers102		Food101		FGVCAircraft	
	HM	Acc.	HM	Acc.	HM	Acc.	HM	Acc.
CLIP	68.75 ± 0.00	65.39 ± 0.00	72.74 ± 0.00	67.28 ± 0.00	90.18 ± 0.00	85.40 ± 0.00	30.25 ± 0.00	23.94 ± 0.00
CoOp	68.22 ± 0.49	63.81 ± 0.44	78.33 ± 2.26	72.11 ± 2.36	86.65 ± 1.38	80.84 ± 1.50	29.38 ± 1.78	24.80 ± 1.23
CoCoOp	71.49 ± 0.62	67.75 ± 0.68	80.04 ± 1.46	71.95 ± 1.24	90.41 ± 0.24	85.61 ± 0.43	27.87 ± 11.36	21.46 ± 7.42
MaPLe	73.57 ± 0.77	69.97 ± 0.87	82.78 ± 0.69	77.32 ± 0.89	$\mathbf{91.46 \pm 0.14}$	$\mathbf{87.10 \pm 0.22}$	35.29 ± 0.58	27.63 ± 1.10
BMIP	$\mathbf{74.48 \pm 0.85}$	$\mathbf{71.03 \pm 0.68}$	$\mathbf{83.86 \pm 1.70}$	$\mathbf{78.90 \pm 1.57}$	90.86 ± 0.13	86.43 ± 0.30	$\mathbf{37.25 \pm 0.93}$	$\mathbf{29.93 \pm 2.75}$

	SUN397		DTD		EuroSAT		UCF101	
	HM	Acc.	HM	Acc.	HM	Acc.	HM	Acc.
CLIP	72.26 ± 0.00	62.57 ± 0.00	57.32 ± 0.00	44.56 ± 0.00	58.16 ± 0.00	41.40 ± 0.00	71.00 ± 0.00	64.97 ± 0.00
CoOp	71.37 ± 1.21	61.82 ± 1.11	57.22 ± 2.37	48.18 ± 1.78	74.33 ± 4.35	59.65 ± 5.07	71.68 ± 2.84	65.41 ± 2.18
CoCoOp	77.17 ± 0.27	68.17 ± 0.33	60.59 ± 1.51	47.90 ± 1.43	73.77 ± 3.58	58.08 ± 1.49	76.59 ± 0.79	70.39 ± 1.25
MaPLe	$\mathbf{79.58 \pm 0.13}$	$\mathbf{70.90 \pm 0.22}$	64.49 ± 3.73	54.13 ± 2.19	81.43 ± 0.53	70.33 ± 2.74	80.69 ± 0.26	73.53 ± 0.45
BMIP	79.02 ± 0.24	70.57 ± 0.47	$\mathbf{67.02 \pm 0.90}$	$\mathbf{54.90 \pm 1.28}$	$\mathbf{86.10 \pm 1.58}$	$\mathbf{73.77 \pm 1.10}$	$\mathbf{82.29 \pm 0.96}$	$\mathbf{75.00 \pm 0.80}$

Table 1: Performance comparison on 11 datasets using ViT-B/16 architecture. The best performance is in bold.

	Source	**Target**										**Average**
	ImageNet	Caltech	Pets	Cars	Flowers	Food	Aircraft	SUN	DTD	EuroSAT	UCF	
CLIP	66.70	93.30	89.10	65.70	70.70	85.90	**24.80**	62.60	44.30	48.30	67.60	65.24
CoOp	**71.51**	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55	63.88
CoCoOp	71.02	**94.43**	90.14	65.32	71.88	86.06	22.94	**67.36**	45.73	45.37	68.21	65.74
MaPLe	70.72	93.53	**90.49**	65.57	**72.23**	86.20	24.74	67.01	**46.49**	48.06	68.69	66.30
BMIP	70.86	94.13	90.13	**66.03**	72.13	**86.23**	24.10	67.30	45.97	**49.63**	**68.96**	**66.86**

Table 2: Comparison of BMIP with existing approaches on cross-dataset transfer task.

5.3 Cross-Dataset Transfer Task

To evaluate the transfer ability of our method, we finetune multi-modal prompts using the ImageNet dataset and directly transfer them to the other 10 datasets. Table 2 presents a performance comparison between CoOp, CoCoOp, MaPLe, and BMIP. BMIP exhibits competitive performance across 10 datasets, achieving the highest average accuracy of 66.86%, with a particularly notable accuracy of 49.63% on EuroSAT. Although CoOp performs the best on the source dataset, its performance degrades on other datasets. These results suggest that using bi-directional modality interaction within BMIP aids in better cross-dataset generalization, exceeding MaPLe.

5.4 Domain Generalization Task

We assess the robustness of the mentioned methods trained on ImageNet to various out-of-distribution (OOD) datasets. BMIP outperforms MaPLe on three of four out-of-distribution

datasets, enhancing the OOD average classification accuracy to 60.40% and achieving an overall average accuracy of 62.50% on both training and testing dataset, as shown in Table 3. The results demonstrate that BMIP effectively addresses open-world challenges, including open-world generalization and domain generalization, highlighting the model's strong performance in these areas.

5.5 Flexibility of BMIP

BMIP focuses on prompt learning for modality interaction and can be combined with any prompt-based model, using MaPLe as the foundational model. Therefore, to verify the flexibility of BMIP, we combine BMIP with two representative methods, PromptSRC [Khattak *et al.*, 2023b] and Co-Prompt [Roy and Etemad, 2024], which represent the latest approaches utilizing additional knowledge, such as regularized information, to enhance model training. Table 4 indicates

Method	Source	Target				Average	OOD Average
	ImageNet	ImagenetV2	ImageNet-S	ImageNet-A	ImageNet-R		
CLIP	66.73	60.83	46.15	47.77	73.86	59.07	57.15
CoOp	**71.51**	64.20	47.99	49.71	75.21	61.72	59.28
CoCoOp	71.02	64.07	48.75	50.63	76.18	62.13	59.91
MaPLe	70.72	64.07	**49.15**	50.90	76.98	62.36	60.28
BMIP	70.86	**64.23**	49.13	**51.06**	**77.20**	**62.50**	**60.40**

Table 3: Comparison of BMIP with existing approaches in domain generalization setting.

Method	HM	Acc.
CLIP	71.70 ± 0.00	63.92 ± 0.00
MaPLe	78.22 ± 0.26	71.76 ± 0.28
+BMIP	$\mathbf{79.04 \pm 0.13}$	$\mathbf{72.17 \pm 0.16}$
PromptSRC	79.67 ± 0.46	73.43 ± 0.35
+BMIP	$\mathbf{80.03 \pm 0.22}$	$\mathbf{73.97 \pm 0.43}$
CoPrompt	78.99 ± 0.25	71.48 ± 0.73
+BMIP	$\mathbf{79.54 \pm 0.28}$	$\mathbf{72.35 \pm 0.32}$

Table 4: Performance improvements from combining three baselines
with BMIP (average across 11 datasets)

Aggregation Method	Open-World Generalization	
	HM	Acc.
IVLP	77.51	71.18
CoCoOp	75.83	67.67
MaPLe	78.18	71.72
MaPLe†	77.19	71.66
Addition	78.40	71.42
Attention	77.74	71.56
Joint	78.79	72.24
BMIP	**79.04**	**72.17**

Table 5: Comparison of BMIP with different methods in open-world
generalization tasks. IVLP represents independently trained vision
and language prompts.

that BMIP brings a significant performance improvement to
these methods, demonstrating BMIP's superior capability as a
foundational model compared to MaPLe.

5.6 Ablation Studies

In our ablation studies, we explore various aggregation functions to determine the individual contributions of the BMIP components to their overall performance, thus validating the intuition derived from our corollary. To ensure that the performance improvement of BMIP is not attributed to the number of parameters, we conduct comparative experiments that equalize the number of parameters between MaPLe and BMIP.

Compare with Other Aggregation Functions. We describe and compare different aggregation functions in more detail in Appendices B and C to verify the benefits of the learnable weighted aggregation function over direct prompt modification functions, such as Addition, Attention, and Joint. As we can see from Table 5, the comparison with Addition highlights the importance of our learnable weights, while the contrast with Attention demonstrates the reliability of learnable weights over similarity-based prompt calculations. Furthermore, the comparison with concatenation-based prompt aggregation underscores the effectiveness of our replacement strategy in reducing prompt redundancy. The comparisons above validate the intuition presented in our analysis: the prompts in VLMs can be mutually replaced and optimized simultaneously and BMIP improves the trainability of prompts. We provide a detailed report on the performance of different aggregation functions across all settings in Appendix C.

Number of Parameters. To ascertain that BMIP's enhanced performance is not due to an increase in parameter count relative to MaPLe, we equalize the number of parameters between

MaPLe (denoted MaPLe†) and BMIP. Table 5 demonstrates that MaPLe† tends to overfit to base classes when parameters are increased, whereas BMIP maintains robust performance despite a higher parameter count. This indicates the efficacy of BMIP's bi-directional modality interaction framework in harmonizing prompt information across modalities. We also assess the impact of varying prompt depths and lengths in Appendices D and E.

6 Conclusion

Prompt learning is regarded as a pivotal technique for adapting pre-trained VLMs to specific downstream tasks. However, existing methods mainly focus on either single-modal prompts or uni-directional modality interaction, which neglects the potent alignment effects that arise from the modalities interaction, causing performance degradation on various datasets. This paper proposes a novel bi-directional modality interaction prompt learning method (BMIP) that includes advanced mechanisms for deep language prompt learning, deep vision prompt learning, and most notably, the interaction between vision and language modalities. Furthermore, we propose a new evaluation paradigm termed *open-world generalization* which offers a more realistic evaluation of prompt learning. Extensive experimental results demonstrate the effectiveness of the proposed BMIP method across various evaluation tasks, particularly in handling datasets with unbalanced text and image variances, such as EuroSAT, and the proposal is flexible enough to be applied to further improve the performance of other methods consistently.

Acknowledgments

References

[Alayrac *et al.*, 2022] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing*, pages 23716–23736, 2022.

[Bossard *et al.*, 2014] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 - mining discriminative components with random forests. In *Proceedings of the 13th European Conference on Computer Vision*, pages 446–461, 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 2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2014.

[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 *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[Derakhshani *et al.*, 2023] Mohammad Mahdi Derakhshani, Enrique Sanchez, Adrian Bulat, Victor Guilherme Turrisi da Costa, Cees G. M. Snoek, Georgios Tzimiropoulos, and Brais Martínez. Variational prompt tuning improves generalization of vision-language models, 2023.

[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 *the 2024 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, page 178, 2004.

[Guo *et al.*, 2025] Lan-Zhe Guo, Lin-Han Jia, Jie-Jing Shao, and Yu-Feng Li. Robust semi-supervised learning in open environments. *Frontiers of Computer Science*, 19(8):198345, 2025.

[He *et al.*, 2022] Yun He, Huaixiu Steven Zheng, Yi Tay, Jai Prakash Gupta, Yu Du, Vamsi Aribandi, Zhe Zhao, YaGuang Li, Zhao Chen, Donald Metzler, Heng-Tze Cheng, and Ed H. Chi. Hyperprompt: Prompt-based task-conditioning of transformers. In *Proceedings of the 39th International Conference on Machine Learning*, pages 8678–8690, 2022.

[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*, pages 2217–2226, 2019.

[Hendrycks *et al.*, 2021a] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the 2021 IEEE International Conference on Computer Vision*, pages 8340–8349, 2021.

[Hendrycks *et al.*, 2021b] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021.

[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 *Proceedings of the 38th International Conference on Machine Learning*, pages 4904–4916, 2021.

[Jia *et al.*, 2022] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Proceedings of the 17th European Conference on Computer Vision*, pages 709–727, 2022.

[Kan *et al.*, 2023] Baoshuo Kan, Teng Wang, Wenpeng Lu, Xiantong Zhen, Weili Guan, and Feng Zheng. Knowledge-aware prompt tuning for generalizable vision-language models. In *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision*, pages 15670–15680, 2023.

[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 2023 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 *The 2023 IEEE/CVF International Conference on Computer Vision*, pages 15144–15154, 2023.

[Krause *et al.*, 2013] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *the 2013 IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013.

[Lee *et al.*, 2023] Dongjun Lee, Seokwon Song, Jihee Suh, Joonmyeong Choi, Sanghyeok Lee, and Hyunwoo J Kim. Read-only prompt optimization for vision-language few-shot learning. In *Proceedings of the 2023 IEEE Interna-*

tional Conference on Computer Vision, pages 1401–1411, 2023.

[Li and Liang, 2021] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 4582–4597, 2021.

[Li *et al.*, 2019] Yu-Feng Li, Lan-Zhe Guo, and Zhi-Hua Zhou. Towards safe weakly supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):334–346, 2019.

[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. In *Advances in Neural Information Processing Systems*, pages 55785–55801, 2023.

[Lu *et al.*, 2022] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5206–5215, 2022.

[Lv *et al.*, 2025] Song-Lin Lv, Rui Zhu, Yu-Feng Li, and Lan-Zhe Guo. Unlabeled data or pre-trained model: Rethinking semi-supervised learning and pretrain-finetuning. 2025.

[Maji *et al.*, 2023] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft, 2023.

[Nilsback and Zisserman, 2008] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the 6th Indian Conference on Computer Vision*, pages 722–729, 2008.

[Parkhi *et al.*, 2012] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3498–3505, 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, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763, 2021.

[Rao *et al.*, 2022] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. DenseCLIP: Language-guided dense prediction with context-aware prompting. In *Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18082–18091, 2022.

[Recht *et al.*, 2019] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, pages 5389–5400, 2019.

[Roy and Etemad, 2024] Shuvendu Roy and Ali Etemad. Consistency-guided prompt learning for vision-language models. In *The 12th International Conference on Learning Representations,*, 2024.

[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, 2012.

[Tan *et al.*, 2025] Hao-Zhe Tan, Zhi Zhou, Yu-Feng Li, and Lan-Zhe Guo. Vision-language model selection and reuse for downstream adaptation. 2025.

[Wang *et al.*, 2019] Haohan Wang, Songwei Ge, Zachary C. Lipton, and Eric P. Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019.

[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 *Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3485–3492, 2010.

[Yao *et al.*, 2023] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6757–6767, 2023.

[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 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10899–10909, 2023.

[Zang *et al.*, 2022] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Unified vision and language prompt learning, 2022.

[Zareian *et al.*, 2021] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021.

[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, 2021.

[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 2022 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*, pages 2337–2348, 2022.

[Zhou *et al.*, 2024] Zhi Zhou, Ming Yang, Jiang-Xin Shi, Lan-Zhe Guo, and Yufeng Li. Decoop: Robust prompt tuning with out-of-distribution detection. In *Proceedings of the 40th International Conference on Machine Learning*, 2024.