# MaPLe: Multi-modal Prompt Learning

**Anonymous authors**
Paper under double-blind review

## Abstract

Pre-trained vision-language (V-L) models such as CLIP have shown excellent generalization ability to downstream tasks. However, they are sensitive to the choice of input text prompts and require careful selection of prompt templates to perform well. Inspired by the Natural Language Processing (NLP) literature, recent CLIP adaptation approaches learn prompts as the textual inputs to fine-tune CLIP for downstream tasks. We note that using prompting to adapt representations in a single branch of CLIP (language or vision) is sub-optimal since it does not allow the flexibility to dynamically adjust both representation spaces on a downstream task. In this work, we propose Multi-modal Prompt Learning (MaPLe) for *both* vision and language branches to improve alignment between the vision and language representations. Our design promotes strong coupling between the vision-language prompts to ensure mutual synergy and discourages learning independent uni-modal solutions. Further, we learn separate prompts across different early stages to progressively model the stage-wise feature relationships to allow rich context learning. We evaluate the effectiveness of our approach on *three* representative tasks of generalization to novel classes, new target datasets and unseen domain shifts. Compared with the state-of-the-art method Co-CoOp, MaPLe exhibits favorable performance and achieves an absolute gain of 3.45% on novel classes and 2.72% on overall harmonic-mean, averaged over 11 diverse image recognition datasets. Our code and models will be publicly released.

## 1 Introduction

Foundational vision-language models such as CLIP (Contrastive Language-Image Pretraining) (Radford et al., 2021) have shown excellent generalization ability to downstream tasks. Such models are trained to align language and vision modalities on web-scale data *e.g.*, 400 million text-image pairs in the case of CLIP. The resulting model can reason about the open-vocabulary visual concepts, thanks to the rich supervision provided by natural language. During inference, hand-engineered text prompts are used *e.g.*, '`a photo of a <category>`' as a query for text encoder. The output text embeddings are matched with the visual embeddings from an image encoder to predict the output class. Designing high quality contextual prompts have been proven to enhance the performance of CLIP and other V-L models (Jin et al., 2021; Yao et al., 2021b).

Despite the effectiveness of CLIP towards generalization to new concepts, its massive scale and the scarcity of training data (*e.g.*, few-shot setting) makes it infeasible to fine-tune the full model for downstream tasks. Such fine-tuning can also forget the useful knowledge acquired in the large-scale pretraining phase and can pose a risk of overfitting to the downstream task. To address the above challenges, existing works propose language prompt learning to avoid manually adjusting the prompt templates and providing a mechanism to adapt the model while keeping the original weights frozen (Zhou et al., 2022a;b; Lu et al., 2022; Huang et al., 2022; Manli et al., 2022). Owing to the inspiration from Natural Language Processing (NLP) models, these approaches only explore prompt learning for the text encoder in CLIP (Fig. 1:a) while adaptation choices together with an equally important image encoder of CLIP remains an unexplored topic in the literature.

Our motivation derives from the multi-modal nature of CLIP, where a text and image encoder co-exist and *both* contribute towards properly aligning the vision-language modalities. We argue that any prompting technique should adapt the model completely and therefore, learning prompts only for the text encoder in CLIP is not sufficient to model the adaptations needed for the image encoder. To this end, we set out to achieve completeness in the prompting approach and propose **M**ulti-mod**a**l

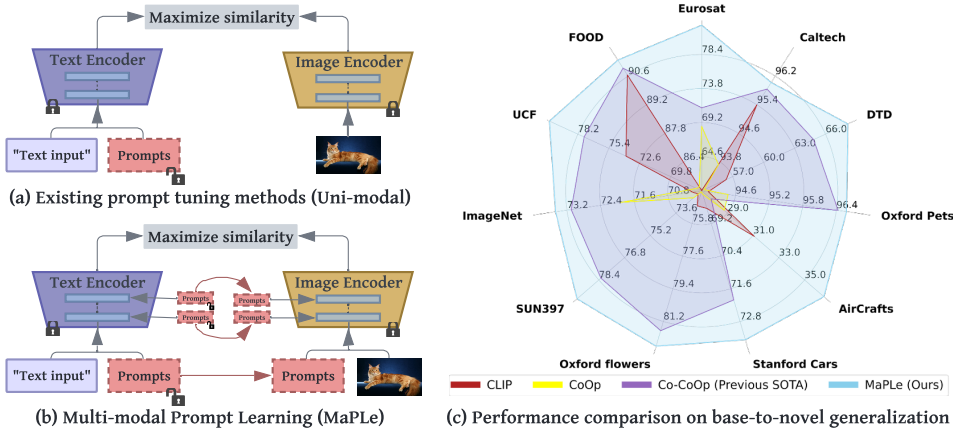


Figure 1: Comparison of MaPLe with standard prompt learning methods. **(a)** Existing methods adopt uni-modal prompting techniques to fine-tune CLIP representations as prompts are learned only in a single branch of CLIP (language or vision). **(b)** MaPLe introduces branch-aware hierarchical prompts that adapt both language and vision branches simultaneously for improved generalization. **(c)** MaPLe surpasses state-of-the-art methods on 11 diverse image recognition datasets for novel class generalization task.

**P**rompt **Le**arning (MaPLe) to adequately fine-tune the text and image encoder representations such that their optimal alignment can be achieved on the downstream tasks (Fig. 1:b). Our extensive experiments on three key representative settings including base-to-novel class generalization, cross-dataset evaluation, and domain generalization demonstrate the strength of MaPLe. On base-to-novel class generalization, our proposed MaPLe outperforms existing prompt learning approaches across 11 diverse image recognition datasets (Fig. 1:c) and achieves absolute average gain of 3.45% on novel classes and 2.72% on harmonic-mean over the state-of-the-art method Co-CoOp (Zhou et al., 2022a). Further, MaPLe demonstrates favorable generalization ability and robustness in cross-dataset transfer and domain generalization settings, leading to consistent improvements compared to existing approaches. Owing to its streamlined architectural design, MaPLe exhibits improved efficiency during both training and inference without much overhead, as compared to Co-CoOp which lacks efficiency due to its image instance conditioned design.

In summary, the main contributions of this work include:

- We propose *multi-modal* prompt learning to favourably align the vision-language representations in CLIP. To the best of our knowledge, this is the first multi-modal prompting approach developed for fine-tuning CLIP.
- In order to link prompts learned in text and image encoders, we propose a *coupling function* to explicitly condition vision prompts on their language counterparts. It acts as a bridge between the two modalities and allows mutual propagation of gradients to promote synergy.
- Our multi-modal prompts are learned across multiple transformer blocks in both vision and language branches to *progressively* learn the synergistic behaviour of both modalities. This deep prompting strategy allows modeling the contextual relationships independently, thus providing more flexibility to align the vision-language representations.

## 2 Related Work

**Vision Language Models:** The combined use of language supervision with natural images is found to be of great interest in the computer vision community. In contrast to models learned with only image supervision, these vision-language (V-L) models encode rich multimodal representations. Recently, V-L models like CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), LiT (Zhai et al., 2022), FILIP (Yao et al., 2021a) and Florence Yuan et al. (2021) have demonstrated exceptional performance on a wide spectrum of tasks including few-shot and zero-shot visual recognition. These models learn joint image-language representations in a self-supervised manner using abundantly available data from the web. For example, CLIP and ALIGN respectively use $\sim$400M and $\sim$1B image-text pairs to train a multi-modal network. Although these pre-trained V-L models learn generalized representations, efficiently adapting them to downstream tasks is still a challenging problem.

Many works have demonstrated better performance on downstream tasks by using tailored methods to adapt V-L models for few-shot image-recognition (Gao et al., 2021; Zhang et al., 2022a; Kim et al., 2022), object detection (Rasheed et al., 2022; Maaz et al., 2022; Zhou et al., 2022c; Gu et al., 2021; Zang et al., 2022; Feng et al., 2022), and segmentation (Li et al., 2022; Rao et al., 2022; Ding et al., 2022; Lüddecke & Ecker, 2022). In this work, we propose a novel multi-modal prompt learning technique to effectively adapt CLIP for few-shot and zero-shot visual recognition tasks.

**Prompt Learning:** The instructions in the form of a sentence, known as text prompt, are usually given to the language branch of a V-L model, allowing it to better understand the task. The prompt can be handcrafted for a downstream task or learned automatically during fine-tuning stage. The latter is referred to as 'Prompt Learning' which was first used in NLP (Li & Liang, 2021; Lester et al., 2021; Liu et al., 2021) followed by the adaptation in V-L (Zhou et al., 2022b;a; Zhu et al., 2022) and vision-only (Jia et al., 2022; Zhang et al., 2022b; Wang et al., 2022a;b) models.

**Prompt Learning in Vision Language models:** Full fine-tuning and linear probing (Gao et al., 2021) are two typical approaches to adapt a V-L model (*i.e.* CLIP) to the downstream tasks. The complete fine-tuning results in degrading the previously learned joint V-L representation while linear probing limits the zero-shot capability of CLIP. To this end, inspired from prompt learning in NLP, many works have proposed to adapt V-L models by learning the prompt tokens in an end-to-end training. CoOp (Zhou et al., 2022b) fine-tunes CLIP for few-shot image classification by optimizing continuous set of prompt vectors at its language branch. Co-CoOp (Zhou et al., 2022a) highlights the inferior performance of CoOp on novel classes and solves the generalization issue by explicitly conditioning prompts on image instances. Lu et al. (2022) proposes to optimize multiple set of prompts by learning the distribution of prompts. Ju et al. (2021) adapt CLIP by learning prompts for video understanding tasks. Bahng et al. (2022) perform visual prompt tuning on CLIP by prompting on the vision branch. We note that the existing methods follow independent *uni-modal* solutions and learn prompts either in the language or in the vision branch of CLIP, thus adapting CLIP partially. In this paper, we explore an important question: given the multimodal nature of CLIP, is complete prompting (*i.e.*, in both language and vision branches) better suited to adapt CLIP? Our work is the first to answer this question by investigating the effectiveness of multi-modal prompt learning in order to improve alignment between vision and language representations.

## 3 METHOD

Our approach concerns with fine-tuning a pre-trained multi-modal CLIP for better generalization to downstream tasks through context optimization via prompting. Fig. 2 shows the overall architecture of our proposed MaPLe (**M**ulti-mod**a**l **P**rompt **Le**arning) framework. Unlike previous approaches (Zhou et al., 2022b;a) which learn context prompts only at the language branch, MaPLe proposes a joint prompting approach where the context prompts are learned in both vision and language branches. Specifically, we append learnable context tokens in the language branch and explicitly condition the vision prompts on the language prompts via a coupling function to establish interaction between them. In order to learn hierarchical contextual representations, we introduce deep prompting in both branches through separate learnable context prompts across different transformer blocks. During the fine-tuning stage, only the context prompts along with their coupling function are learned while the rest of the model is frozen. Below, we first outline the pre-trained CLIP architecture and then present our proposed fine-tuning approach.

### 3.1 REVISITING CLIP

We build our approach on a pre-trained vision-language (V-L) model, CLIP, which consists of a text and vision encoder. Consistent with existing prompting methods (Zhou et al., 2022b;a), we use a vision transformer (ViT) (Dosovitskiy et al., 2021) based CLIP model. CLIP encodes an image $I \in \mathbb{R}^{H \times W \times 3}$ and a corresponding text description as explained below.

**Encoding Image:** Image encoder $\mathcal{V}$ with $K$ transformer layers $\{\mathcal{V}_i\}_{i=1}^{K}$, splits the image $I$ into $M$ fixed-size patches which are projected into patch embeddings $E_0 \in \mathbb{R}^{M \times d_v}$. Patch embeddings $E_i$ are then input to the $(i+1)^{\text{th}}$ transformer block $(\mathcal{V}_{i+1})$ along with an appended learnable class (CLS) token $\mathrm{c}_i$ and sequentially processed through $K$ transformer blocks,

$$[\mathrm{c}_i, E_i] = \mathcal{V}_i([\mathrm{c}_{i-1}, E_{i-1}]) \qquad i = 1, 2, \cdots, K.$$

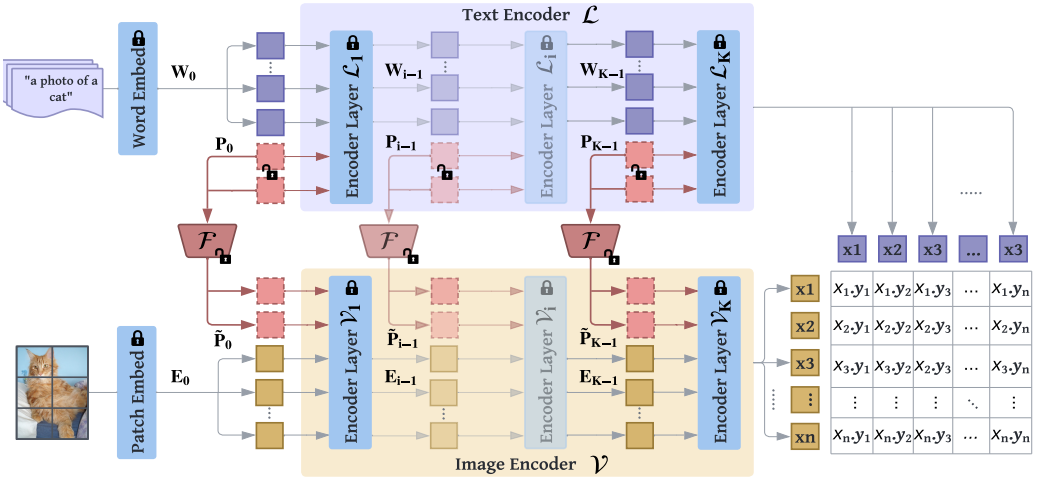


Figure 2: Overview of our proposed MaPLe (**M**ulti-moda**l** **P**rompt **Le**arning) framework for prompt learning in V-L models. MaPLe tunes both vision and language branches where only the context prompts are learned, while the rest of the model is frozen. MaPLe conditions the vision prompts on language prompts via a V-L coupling function $\mathcal{F}$ to induce mutual synergy between the two modalities. Our framework uses deep contextual prompting where separate context prompts are learned across multiple transformer blocks.

To obtain the final image representation $x$, the class token $\mathtt{c}_K$ of last transformer layer $(\mathcal{V}_K)$ is projected to a common V-L latent embedding space via $\mathtt{ImageProj}$,

$$x = \mathtt{ImageProj}(\mathtt{c}_K) \qquad x \in \mathbb{R}^{d_{vl}}.$$

**Encoding Text:** CLIP text encoder generates feature representations for text description by tokenizing the words and projecting them to word embeddings $W_0 = [w_0^1, w_0^2, \cdots, w_0^N] \in \mathbb{R}^{N \times d_l}$. At each stage, $W_i$ is input to the $(i+1)^{\text{th}}$ transformer layer of text encoding branch $(\mathcal{L}_{i+1})$,

$$[W_i] = \mathcal{L}_i(W_{i-1}) \qquad i = 1, 2, \cdots, K.$$

The final text representation $z$ is obtained by projecting the text embeddings corresponding to the last token of the last transformer block $\mathcal{L}_K$ to a common V-L latent embedding space via $\mathtt{TextProj}$,

$$z = \mathtt{TextProj}(w_K^N) \qquad z \in \mathbb{R}^{d_{vl}}.$$

**Zero-shot Classification:** For zero-shot classification, text prompts are hand-crafted with class labels $y \in \{1, 2, \ldots C\}$ (*e.g.*, '`a photo of a <category>`') having $C$ classes. Prediction $\hat{y}$ corresponding to the image $I$ having the highest cosine similarity score $(sim(\cdot))$ is formulated as,

$$p(\hat{y}|x) = \frac{\exp(sim(x, z_{\hat{y}})/\tau)}{\sum_{i=1}^{C} \exp(sim(x, z_i))}.$$

Here, $\tau$ is a temperature parameter.

### 3.2 MaPLe: Multi-modal Prompt Learning

To efficiently fine-tune CLIP for downstream image recognition tasks, we explore the potential of *multi-modal* prompt tuning. We reason that prior works that have predominantly explored uni-modal approaches are less suitable as they do not offer the flexibility to dynamically adapt both language and vision representation spaces. Thus to achieve completeness in prompting, we underline the importance of multi-modal prompting approach. In Fig. 3, we visualize and compare the image embeddings of MaPLe with recent state-of-the-art work, Co-CoOp. Note that the image embeddings of CLIP, CoOp and Co-CoOp will be identical as they do not learn prompts in the vision branch. The visualization shows that image embeddings of MaPLe are more separable indicating that learning vision prompts in addition to language prompts leads to better adaptation of CLIP.

In addition to multi-modal prompting, we find that it is essential to learn prompts in the deeper transformer layers to progressively model stage-wise feature representations. To this end, we propose to introduce learnable tokens in the first $J$ (where $J < K$) layers of both vision and language branches. These multi-modal hierarchical prompts utilize the knowledge embedded in CLIP model to effectively learn task relevant contextual representations (see Fig. 4).

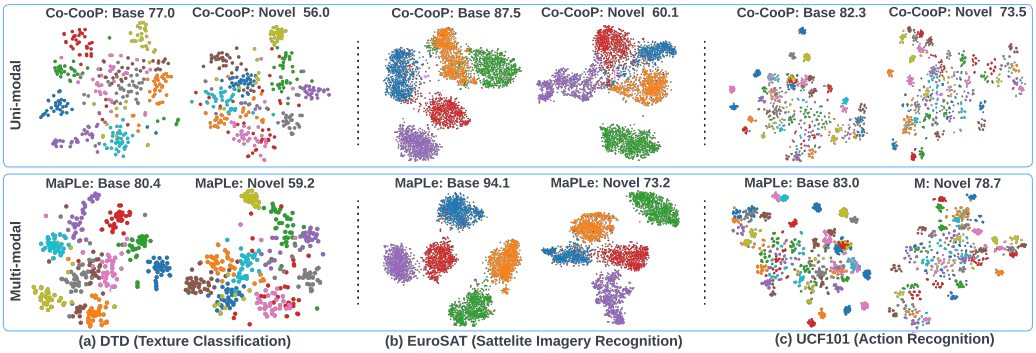


Figure 3: t-SNE plots of image embeddings in uni-modal prompting method Co-CoOp, and MaPLe on 3 diverse image recognition datasets. MaPLe shows better separability in both base and novel classes.

#### 3.2.1 Deep Language Prompting

To learn the language context prompts, we introduce $b$ learnable tokens $\{P^i \in \mathbb{R}^{d_l}\}_{i=1}^{b}$, in the language branch of CLIP. The input embeddings now follow the form $[P^1, P^2, \cdots, P^b, W]$, where $W = [w^1, w^2, \cdots, w^N]$ corresponds to fixed input tokens. New learnable tokens are further introduced in each transformer block of the language encoder ($\mathcal{L}_i$) up to a specific depth $J$,

$$[\_,\ W_i] = \mathcal{L}_i([P_{i-1}, W_{i-1}]) \qquad i = 1, 2, \cdots, J. \tag{1}$$

Here $[\cdot,\ \cdot]$ refers to the concatenation operation. After $J^{\text{th}}$ transformer layer, the subsequent layers process previous layer prompts and final text representation $z$ is computed,

$$[P_j,\ W_i] = \mathcal{L}_j([P_{j-1}, W_{j-1}]) \qquad j = J+1, \cdots, K, \tag{2}$$

$$z = \texttt{TextProj}(w_K^N). \tag{3}$$

When $J = 1$, the learnable tokens $P$ are only applied at the input of first transformer layer, and this deep language prompting technique degenerates to CoOp (Zhou et al., 2022b).

#### 3.2.2 Deep Vision Prompting

Similar to deep language prompting, we introduce $b$ learnable tokens $\{\tilde{P}^i \in \mathbb{R}^{d_v}\}_{i=1}^{b}$, in the vision branch of CLIP alongside the input image tokens. New learnable tokens are further introduced in deeper transformer layers of the image encoder ($\mathcal{V}$) up to depth $J$.

$$\begin{aligned}
[c_i, E_i,\ _\,] &= \mathcal{V}_i([c_{i-1}, E_{i-1}, \tilde{P}_{i-1}]) && i = 1, 2, \cdots, J, \\
[c_j, E_j, \tilde{P}_j] &= \mathcal{V}_j([c_{j-1}, E_{j-1}, \tilde{P}_{j-1}]) && j = J+1, \cdots, K, \\
x &= \texttt{ImageProj}(c_K).
\end{aligned}$$

Our deep prompting provides the flexibility to learn prompts across different feature hierarchies within the ViT architecture. We find that sharing prompts across stages is better compared to independent prompts as features are more correlated due to successive transformer block processing. Thus, the later stages do not provide independently-learned complimentary prompts as compared to the early stages.

#### 3.2.3 Vision Language Prompt Coupling

We reason that in prompt tuning it is essential to take a multi-modal approach and *simultaneously* adapt both the vision and language branch of CLIP in order to achieve completeness in context optimization. A simple approach would be to naively combine deep vision and language prompting, where both the language prompts $P$, and the vision prompts $\tilde{P}$, will be learned during the same training schedule. We name this design as *'Independent V-L Prompting'*. Although this approach satisfies the requirement of completeness in prompting, this design lacks synergy between vision and language branch as both branches do not interact while learning the task relevant context prompts.

To this end, we propose a branch-aware multi-modal prompting which tunes vision and language branch of CLIP together by sharing prompts across both modalities. Language prompt tokens are

introduced in the language branch up to $J^{\text{th}}$ transformer block similar to deep language prompting as illustrated in Eqs. 1-3. To ensure mutual synergy between V-L prompts, vision prompts $\tilde{P}$, are obtained by projecting language prompts $P$ via vision-to-language projection which we refer to as *V-L coupling function* $\mathcal{F}(\cdot)$, such that $\tilde{P}_k = \mathcal{F}_k(P_k)$. The coupling function is implemented as a linear layer which maps $d_l$ dimensional inputs to $d_v$. This acts as a bridge between the two modalities, thus encouraging mutual propagation of gradients.

$$
\begin{aligned}
[c_i, E_i, \_\_] &= \mathcal{V}_i([c_{i-1}, E_{i-1}, \boldsymbol{\mathcal{F}_{i-1}(P_{i-1})}]) && i = 1, 2, \cdots, J \\
[c_j, E_j, \tilde{P}_j] &= \mathcal{V}_j([c_{j-1}, E_{j-1}, \tilde{P}_{j-1}]) && j = J+1, \cdots, K \\
x &= \texttt{ImageProj}(c_K)
\end{aligned}
$$

Unlike independent V-L prompting, the explicit conditioning of $\tilde{P}$ on $P$ helps to learn prompts in a shared embedding space between language and vision branch, thus improving the mutual synergy.

# 4 EXPERIMENTS

## 4.1 BENCHMARK SETTING

We evaluate and conduct experiments in three benchmark settings for image recognition.

**Generalization from Base-to-Novel Classes:** To evaluate the generalizability of our approach, we follow a zero-shot setting where the datasets are split into base and novel classes. The model is trained only on the base classes in a few-shot setting and evaluated on both base and novel categories.

**Cross-dataset Evaluation:** To validate the potential of our approach in cross-dataset transfer, we evaluate our ImageNet trained model directly on other datasets. Consistent with Co-CoOp, our model is trained on all 1000 ImageNet classes in a few-shot manner.

**Domain Generalization:** We evaluate the robustness of our method on out-of-distribution datasets. Similar to cross-dataset evaluation, we test our ImageNet trained model directly on four other ImageNet datasets that contain various types of domain shifts.

**Datasets:** For generalization from base-to-novel classes and cross-dataset evaluation, we follow Zhou et al. (2022b;a) and evaluate the performance of our method on 11 image classification datasets which covers a wide range of recognition tasks. This includes two generic-objects datasets, ImageNet (Deng et al., 2009) and Caltech101 (Fei-Fei et al., 2004); five fine-grained classifcation datasets, OxfordPets (Parkhi et al., 2012), StanfordCars (Krause et al., 2013), Flowers102 (Nilsback & Zisserman, 2008), Food101 (Bossard et al., 2014), and FGVCAircraft (Maji et al., 2013); a scene recognition dataset SUN397 (Xiao et al., 2010); an action recognition dataset UCF101 (Soomro et al., 2012); a texture dataset DTD (Cimpoi et al., 2014) and a satellite-image dataset EuroSAT (Helber et al., 2019). For domain generalization experiments, we use ImageNet as source dataset and its four variants as target datasets including ImageNetV2 (Recht et al., 2019), ImageNet-Sketch (Wang et al., 2019), ImageNet-A (Hendrycks et al., 2021b) and ImageNet-R (Hendrycks et al., 2021a).

## 4.2 IMPLEMENTATION DETAILS

We use a few-shot training strategy in all experiments where the number of shots are set to 16 which are randomly sampled for each class. We apply prompt tuning on a pre-trained ViT-B/16 CLIP model where $d_l = 512$, $d_v = 768$ and $d_{vl} = 512$. For MaPLe, we set prompt depth $J$ to 9 and the language and vision prompt lengths to 2. All models are trained for 5 epochs with a batch-size of 4 and a learning rate of 0.0035 via SGD optimizer on a single NVIDIA A100 GPU. We report base and novel class accuracies and their harmonic mean (HM) averaged over 3 runs. We initialize the language prompts of the first layer $P_0$ with the pre-trained CLIP word embeddings of the template '`a photo of a <category>`', while for the subsequent layers they are randomly initialized from a normal distribution. For training MaPLe on all 1000 classes of ImageNet as a source model, prompt depth $J$ is set to 3 and the model trained for 2 epochs with learning rate of 0.0026. Hyper-parameters for deep language prompting, deep vision prompting, and independent V-L prompting are detailed in Appendix A. The corresponding hyper-parameters are fixed across all datasets. Fruther, we provide comparison of prompt complexity in Appendix D.

Table 1: Comparison of MaPLe with different deep prompting designs in base-to-novel generalization. Results are averaged over 11 datasets. Co-CoOp ( Zhou et al. (2022a)) is the previous state-of-the-art. HM refers to harmonic mean.

| Method | Base Acc. | Novel Acc. | HM |
|---|---|---|---|
| Co-CoOp (CVPR'22) | 80.47 | 71.69 | 75.83 |
| 1: Deep vision prompting | 80.24 | 73.43 | 76.68 |
| 2: Deep language prompting | 81.72 | 73.81 | 77.56 |
| 3: Independent V-L prompting | 82.15 | 74.07 | 77.90 |
| 4: MaPLe (Ours) | **82.28** | **75.14** | **78.55** |

### 4.3 Prompting CLIP via Vision-Language Prompts

**Prompting Variants:** We first evaluate the performance of different possible prompting design choices as an ablation for our proposed branch-aware multi-modal prompting, MaPLe. These variants include deep language prompting, deep vision prompting and independent V-L prompting. In Table 1, we present the results averaged over the 11 image recognition datasets. Deep language prompting (row-2) shows improvements over deep vision prompting (row-1), indicating that prompts learned at the language branch provide better adaptation of CLIP. Although separately combining the above two approaches (row-3) further improves the performance, it struggles to achieve comprehensive benefits from the language and vision branches. We hypothesize that this is due to the lack of synergy between the learned vision and language prompts as they do not interact with each other during training. Meanwhile, MaPLe (row-4) combines the benefits of prompting in both branches by enforcing interactions through explicit conditioning of vision prompts on the language prompts. It provides improvements on novel and base class accuracies which leads to the best HM of 78.55%. We explore other possible design choices and present the ablations in Appendix C.

### 4.4 Base-to-Novel Generalization

**Generalization to Unseen Classes:** Table 2 presents the performance of MaPLe in base-to-novel generalization setting on 11 recognition datasets. We compare its performance with CLIP zero-shot, and recent prompt learning works including CoOp (Zhou et al., 2022b) and Co-CoOp (Zhou et al., 2022a). In case of CLIP, we use hand-crafted prompts that are specifically designed for each dataset.

In comparison with the state-of-the-art work Co-CoOp, our method shows improved performance on both base and novel categories on all 11 datasets with an exception of marginal reduction on only the base class performance of Caltech101. With mutual synergy from the branch-aware multi-modal prompting, MaPLe better generalizes to novel categories on all 11 datasets in comparison with Co-CoOp, where we obtain an overall gain from 71.69% to 75.14%. When taking into account both the base and novel classes, MaPLe shows an absolute average gain of 2.72% over Co-CoOp.

In comparison with CLIP on novel classes, Co-CoOp improves only on 4/11 datasets dropping the average novel accuracy from 74.22% to 71.69%. MaPLe is a strong competitor which improves accuracy over CLIP on novel classes on 6/11 datasets, with an average gain from 74.22% to 75.14%.

**Generalization and Performance on Base Classes:** Co-CoOp solves the poor generalization problem in CoOp by explicitly conditioning prompts on image instances and shows significant gains in novel categories. However on base classes, it improves over CoOp only on 3/11 datasets with an average drop in performance from 82.69% to 80.47%. Meanwhile, the completeness in prompting helps MaPLe improve over CoOp on base classes in 6/11 datasets maintaining the average base accuracy to around 82.28%, in addition to its improvement in generalization to novel classes.

We find that the training strategies of Co-CoOp can be used to substantially boost the generalization performance of vanilla CoOp (6.8% gain in novel classes). We therefore compare our method with CoOp$^{\dagger}$, which trains CoOp in Co-CoOp setting. In comparison with CoOp$^{\dagger}$, the vanilla CoOp model seems to overfit on base classes. When compared to CoOp$^{\dagger}$ which attains an average base accuracy of 80.85%, MaPLe shows an improvement of 1.43% with the average base accuracy of 82.28% (Table 3).

Table 3: Generalization comparison of MaPLe with CoOp†.

| | Base | Novel | HM |
|---|---|---|---|
| CoOp | **82.69** | 63.22 | 71.66 |
| Co-CoOp | 80.47 | 71.69 | 75.83 |
| CoOp† | 80.85 | 70.02 | 75.04 |
| MaPLe | 82.28 | **75.14** | **78.55** |

### 4.5 Cross-Dataset Evaluation

We test the cross-dataset generalization ability of MaPLe by learning multi-modal prompts on all the 1000 ImageNet classes and then transferring it directly on the remaining 10 datasets. Ta-

ble 4 shows the performance comparison between MaPLe, CoOp and Co-CoOp. On the ImageNet source dataset, MaPLe achieves performance comparable to competing approaches but demonstrates a much stronger generalization performance by surpassing CoOp in 9/10 and Co-CoOp in 8/10 datasets. Overall, MaPLe shows competitive performance leading to the highest accuracy of 66.30% averaged over all datasets. This suggests that the use of branch-aware V-L prompting in MaPLe facilitates better generalization for cross-dataset transfer.

### 4.6 Domain Generalization

We show that MaPLe generalizes favourably on out-of-distribution datasets as compared to CoOp and Co-CoOp. We evaluate the direct transferability of ImageNet trained model to various out-of-domain datasets, and observe that it consistently improves against all the existing approaches as indicated in Table 5. This indicates that utilizing multi-modal branch-aware prompting helps MaPLe in enhancing the generalization and robustness of V-L models like CLIP.

### 4.7 Ablation Experiments

**Prompt Depth:** In Fig. 4 (left), we illustrate the effect of prompt depth $J$ for MaPLe and ablate on the depth of language and vision branch *individually*. In general, the performance improves as

Table 2: **Comparison with state-of-the-art methods on base-to-novel generalization**. MaPLe learns multi-modal prompts and demonstrates strong generalization performance over existing methods on 11 different recognition datasets. Absolute gains over Co-CoOp are indicated in blue.

(a) **Average over 11 datasets**

| | Base | Novel | HM |
|---|---|---|---|
| CLIP | 69.34 | 74.22 | 71.70 |
| CoOp | **82.69** | 63.22 | 71.66 |
| Co-CoOp | 80.47 | 71.69 | 75.83 |
| MaPLe | 82.28 | **75.14** | **78.55** |
| | +1.81 | +3.45 | +2.72 |

(b) ImageNet.

| | Base | Novel | HM |
|---|---|---|---|
| CLIP | 72.43 | 68.14 | 70.22 |
| CoOp | 76.47 | 67.88 | 71.92 |
| Co-CoOp | 75.98 | 70.43 | 73.10 |
| MaPLe | **76.66** | **70.54** | **73.47** |
| | +0.68 | +0.11 | +0.37 |

(c) Caltech101

| | Base | Novel | HM |
|---|---|---|---|
| CLIP | 96.84 | 94.00 | 95.40 |
| CoOp | **98.00** | 89.81 | 93.73 |
| Co-CoOp | 97.96 | 93.81 | 95.84 |
| MaPLe | 97.74 | **94.36** | **96.02** |
| | -0.22 | +0.55 | +0.18 |

(d) OxfordPets

| | Base | Novel | HM |
|---|---|---|---|
| CLIP | 91.17 | 97.26 | 94.12 |
| CoOp | 93.67 | 95.29 | 94.47 |
| Co-CoOp | 95.20 | 97.69 | 96.43 |
| MaPLe | **95.43** | **97.76** | **96.58** |
| | +0.23 | +0.07 | +0.15 |

(e) StanfordCars

| | Base | Novel | HM |
|---|---|---|---|
| CLIP | 63.37 | **74.89** | 68.65 |
| CoOp | **78.12** | 60.40 | 68.13 |
| Co-CoOp | 70.49 | 73.59 | 72.01 |
| MaPLe | 72.94 | 74.00 | **73.47** |
| | +2.45 | +0.41 | +1.46 |

(f) Flowers102

| | Base | Novel | HM |
|---|---|---|---|
| CLIP | 72.08 | **77.80** | 74.83 |
| CoOp | **97.60** | 59.67 | 74.06 |
| Co-CoOp | 94.87 | 71.75 | 81.71 |
| MaPLe | 95.92 | 72.46 | **82.56** |
| | +1.05 | +0.71 | +0.85 |

(g) Food101

| | Base | Novel | HM |
|---|---|---|---|
| CLIP | 90.10 | 91.22 | 90.66 |
| CoOp | 88.33 | 82.26 | 85.19 |
| Co-CoOp | 90.70 | 91.29 | 90.99 |
| MaPLe | **90.71** | **92.05** | **91.38** |
| | +0.01 | +0.76 | +0.39 |

(h) FGVCAircraft

| | Base | Novel | HM |
|---|---|---|---|
| CLIP | 27.19 | **36.29** | 31.09 |
| CoOp | **40.44** | 22.30 | 28.75 |
| Co-CoOp | 33.41 | 23.71 | 27.74 |
| MaPLe | 37.44 | 35.61 | **36.50** |
| | +4.03 | +11.90 | +8.76 |

(i) SUN397

| | Base | Novel | HM |
|---|---|---|---|
| CLIP | 69.36 | 75.35 | 72.23 |
| CoOp | 80.60 | 65.89 | 72.51 |
| Co-CoOp | 79.74 | 76.86 | 78.27 |
| MaPLe | **80.82** | **78.70** | **79.75** |
| | +1.08 | +1.84 | +1.48 |

(j) DTD

| | Base | Novel | HM |
|---|---|---|---|
| CLIP | 53.24 | **59.90** | 56.37 |
| CoOp | 79.44 | 41.18 | 54.24 |
| Co-CoOp | 77.01 | 56.00 | 64.85 |
| MaPLe | **80.36** | 59.18 | **68.16** |
| | +3.35 | +3.18 | +3.31 |

(k) EuroSAT

| | Base | Novel | HM |
|---|---|---|---|
| CLIP | 56.48 | 64.05 | 60.03 |
| CoOp | 92.19 | 54.74 | 68.69 |
| Co-CoOp | 87.49 | 60.04 | 71.21 |
| MaPLe | **94.07** | **73.23** | **82.35** |
| | +6.58 | +13.19 | +11.14 |

(l) UCF101

| | Base | Novel | HM |
|---|---|---|---|
| CLIP | 70.53 | 77.50 | 73.85 |
| CoOp | **84.69** | 56.05 | 67.46 |
| Co-CoOp | 82.33 | 73.45 | 77.64 |
| MaPLe | 83.00 | **78.66** | **80.77** |
| | +0.67 | +5.21 | +3.13 |

Table 4: Comparison of MaPLe with existing approaches on cross-dataset evaluation. Overall, MaPLe achieves competitive performance providing highest average accuracy, indicating better generalization.

| | **Source** | **Target** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *ImageNet* | *Caltech101* | *OxfordPets* | *StanfordCars* | *Flowers102* | *Food101* | *Aircraft* | *SUN397* | *DTD* | *EuroSAT* | *UCF101* | *Average* |
| 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 |
| Co-CoOp | 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** |

Table 5: Comparison of MaPLe with existing approaches in domain generalization setting. MaPLe shows highest performance on all target datasets.

| | **Source** | **Target** | | | |
|---|---|---|---|---|---|
| | ImageNet | ImageNetV2 | ImageNet-Sketch | ImageNet-A | ImageNet-R |
| CLIP | 66.73 | 60.83 | 46.15 | 47.77 | 73.96 |
| CoOp | **71.51** | 64.20 | 47.99 | 49.71 | 75.21 |
| Co-CoOp | 71.02 | **64.07** | 48.75 | 50.63 | 76.18 |
| MaPLe | 70.72 | **64.07** | **49.15** | **50.90** | **76.98** |

prompt depth increases. As earlier methods (CoOp and Co-CoOp) utilize shallow language prompting ($J = 1$), we find it interesting to compare our method with deep language prompting. Overall, MaPLe achieves better performance than deep language prompting. We observe that MaPLe achieves maximum performance on validation set at a depth of 9 .

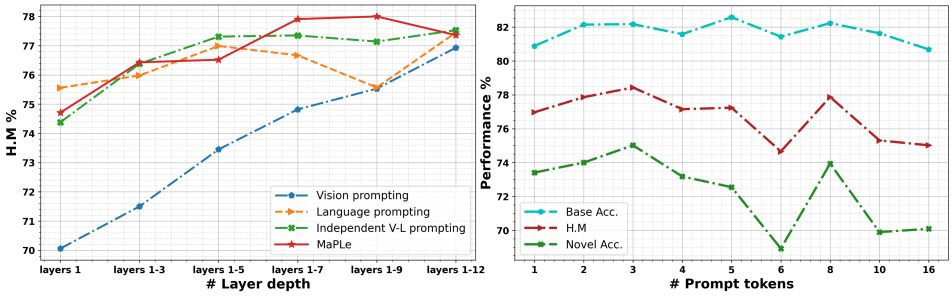


Figure 4: Ablation on prompt depth ***(left)*** and prompt length ***(right)*** in MaPLe. We report average results on the held-out validation sets of all datasets.

**Prompt Length:** Fig. 4 (right) shows the effect of prompt length for MaPLe. As the prompt length increases, the performance on base classes is generally maintained, while the novel class accuracy decreases. This indicates over-fitting which inherently hurts the generalization to novel classes.

**Effectiveness of Multi-modal Prompting:** Fig. 5 shows the analysis of per class accuracy for selected datasets in the order of increasing diversity. It indicates that the performance gains of MaPLe in comparison to Co-CoOp varies across different datasets. MaPLe provides significant gains over Co-CoOp for datasets that have large distribution shifts from the pretraining dataset of CLIP, and vision concepts that are usually rare and less generic. Further analysis is provided in Appendix B.

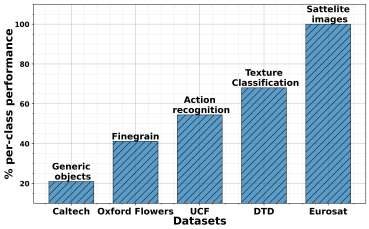


Figure 5: Percentage of classes where MaPLe > Co-CoOp increases as diversity of dataset increases (left to right).

## 5 Conclusion

Adaptation of large-scale V-L models, *e.g.*, CLIP (Radford et al., 2021) to downstream tasks is a challenging problem due to the large number of tunable parameters and limited size of downstream datasets. Prompt learning is an efficient and scalable technique to tailor V-L models to novel downstream tasks. To this end, the current prompt learning approaches either consider only the vision or language side prompting. Our work shows that it is critical to perform prompting for both vision and language branches to appropriately adapt V-L models to downstream tasks. Further, we propose a strategy to ensure synergy between vision-language modalities by explicitly conditioning the vision prompts on textual prompt across different transformer stages. Our approach improves the generalization towards novel categories, cross-dataset transfer and datasets with domain shifts.

## REFERENCES


Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Visual prompting: Modifying pixel space to adapt pre-trained models. *arXiv preprint arXiv:2203.17274*, 2022.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *The European Conference on Computer Vision*, pp. 446–461. Springer, 2014.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2014.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 248–255. Ieee, 2009.

Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11583–11592, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

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

Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. Promptdet: Towards open-vocabulary detection using uncurated images. In *The European Conference on Computer Vision*, 2022.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.

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

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

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

Tony Huang, Jack Chu, and Fangyun Wei. Unsupervised prompt learning for vision-language models. *arXiv preprint arXiv:2204.03649*, 2022.

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

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *The European Conference on Computer Vision*, 2022.

Woojeong Jin, Yu Cheng, Yelong Shen, Weizhu Chen, and Xiang Ren. A good prompt is worth millions of parameters? low-resource prompt-based learning for vision-language models. *arXiv preprint arXiv:2110.08484*, 2021.

Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *The European Conference on Computer Vision*, 2021.

Konwoo Kim, Michael Laskin, Igor Mordatch, and Deepak Pathak. How to adapt your large-scale vision-and-language model, 2022. URL https://openreview.net/forum?id=EhwEUb2ynIa.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 554–561, 2013.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Conference on Empirical Methods in Natural Language Processing*, 2021.

Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022.

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*, 2021.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021.

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

Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7086–7096, 2022.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, Fahad Shahbaz Khan, Rao Muhammad Anwer, and Ming-Hsuan Yang. Class-agnostic object detection with multi-modal transformer. In *The European Conference on Computer Vision*. Springer, 2022.

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

Shu Manli, Nie Weili, Huang De-An, Yu Zhiding, Goldstein Tom, Anandkumar Anima, and Xiao Chaowei. Test-time prompt tuning for zero-shot generalization in vision-language models. 2022.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722–729. IEEE, 2008.

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

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

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18082–18091, 2022.

Hanoona Rasheed, Muhammad Maaz, Muhammad Uzair Khattak, Salman Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. In *Advances in Neural Information Processing Systems*, 2022.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pp. 5389–5400. PMLR, 2019.

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

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *The European Conference on Computer Vision*, 2022a.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 139–149, 2022b.

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

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

Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Cpt: Colorful prompt tuning for pre-trained vision-language models. *arXiv preprint arXiv:2109.11797*, 2021b.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In *The European Conference on Computer Vision*, 2022.

Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18123–18133, 2022.

Renrui Zhang, Rongyao Fang, Peng Gao, Wei Zhang, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. In *The European Conference on Computer Vision*, 2022a.

Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. Neural prompt search. *arXiv preprint arXiv:2206.04673*, 2022b.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16816–16825, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, pp. 1–12, 2022b.

Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *The European Conference on Computer Vision*, 2022c.

Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. *arXiv preprint arXiv:2205.14865*, 2022.