# Task-Oriented Multi-Modal Mutual Learning for Vision-Language Models

Sifan Long [1,2,3 †*] Zhen Zhao[3,4 †*] Junkun Yuan[3,5 †*] Zichang Tan[3] Jiangjiang Liu[3]
Luping Zhou[4] Shengsheng Wang[1,2 ‡] Jingdong Wang[3 ‡]
[1]College of Computer Science and Technology, Jilin University, Jilin, China
[2]Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Jilin, China
[3]Baidu VIS [4]University of Sydney [5]Zhejiang University
longsf22@mails.jlu.edu.cn {zhen.zhao, luping.zhou}@sydney.edu.au yuanjk@zju.edu.cn
wss@jlu.edu.cn {tanzichang, liujiangjiang, wangjingdong}@baidu.com

## Abstract

*Prompt learning has become one of the most efficient paradigms for adapting large pre-trained vision-language models to downstream tasks. Current state-of-the-art methods, like CoOp and ProDA, tend to adopt soft prompts to learn an appropriate prompt for each specific task. Recent CoCoOp further boosts the base-to-new generalization performance via an image-conditional prompt. However, it directly fuses identical image semantics to prompts of different labels and significantly weakens the discrimination among different classes as shown in our experiments. Motivated by this observation, we first propose a class-aware text prompt (CTP) to enrich generated prompts with label-related image information. Unlike CoCoOp, CTP can effectively involve image semantics and avoid introducing extra ambiguities into different prompts. On the other hand, instead of reserving the complete image representations, we propose text-guided feature tuning (TFT) to make the image branch attend to class-related representation. A contrastive loss is employed to align such augmented text and image representations on downstream tasks. In this way, the **image-to-text** CTP and **text-to-image** TFT can be mutually promoted to enhance the adaptation of VLMs for downstream tasks. Extensive experiments demonstrate that our method outperforms the existing methods by a significant margin. Especially, compared to CoCoOp, we achieve an average improvement of 4.03% on new classes and 3.19% on harmonic-mean over eleven classification benchmarks.*

## 1. Introduction

Recently, large vision-language models (VLM), such as CLIP [33] and ALIGN [15], which employ language as supervision signal instead of discrete labels, have shown impressive generalization performance in a wide range of downstream vision tasks. Their multi-modal interaction nature delivers open-vocabulary support and achieves amazing zero-shot classification performance. Despite their impressive transferable abilities, as discussed in [25], it is essential to re-activate specific representation capabilities for optimal performance in certain downstream tasks. Considering their hundreds of millions or billions of parameters, attempting to fine-tune the entire model is impractical and even jeopardizes the well-established representation space [14]. To this end, many recent studies have centred on the efficient and effective adaptation of pre-trained and frozen large VLMs for the specific downstream tasks [28, 51, 50].

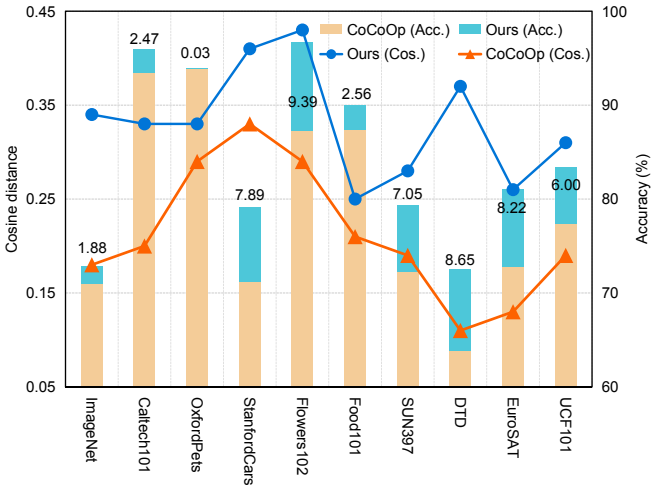


Figure 1: Comparisons between CoCoOp [50] and our method. The cosine distance between the positive and the negative prompts, which quantifies the class discrimination, and the average accuracy on benchmarks are reported.

*Equal contribution. † Interns at Baidu VIS. ‡ Corresponding authors.

Prompt, a simple, compact, and viable strategy, has become the leading solution for deploying large pre-trained VLMs into certain downstream tasks. CLIP [33] utilizes hand-crafted prompts to achieve impressive zero- and few-shot classification performance. Nevertheless, manually-designed prompts require significant domain knowledge and can be highly time-consuming and sub-optimal for specific downstream tasks. To address this problem, later studies [28, 51] adopt soft prompts to learn an appropriate text prompt via optimizing a contrastive loss on different text labels. CoCoOp [50] further highlights the limitations of such static soft prompts and proposes learning **image-dependent** prompts conditioned on individual instances rather than fixed prompts. It achieves great performance gains on unseen classes by adding high-level image embedding to text prompts. However, compared to CoOp with static prompts, CoCoOp essentially fuses identical image semantics with different text labels, leading to inevitable learning ambiguity and resulting in an average performance drop of 2.22% on base classes on 11 datasets (see Table 1). For example, it may associate the dog image semantics with a prompt that references the [class] of a cat. When using the cosine distance to measure the differences between the positive and negative text prompts, as shown in Fig. 1, CoCoOp holds low distance values, suggesting that it brings significant learning ambiguities to text prompts. Therefore, we argue that text prompts should not only condition on distinct input images for better generalization abilities, but also adapt to different classes to eliminate the potential ambiguities.

To achieve this goal, we propose Class-aware Text Prompts (CTP), which leverages label-related image information to generate finer prompts. Specifically, we first contact learnable context vectors and each class label to model the initial prompt sentences. Then we leverage these class prompt sentences to query their corresponding image regions and representations. Corresponding related image features are subsequently added to initial class prompt sentences to produce the final text prompts. In this way, generated image-dependent and class-aware prompts can better concentrate on the image information in a more precise manner. As shown in Fig. 1, our method enjoys better discrimination between positive and negative prompts and consistently outperforms CoCoOp on 11 classification datasets.

On the other hand, we identify a critical problem in these text prompt-based strategies: the image branch is ignored and not adjusted to specific downstream tasks. As shown in Fig. 2 (CoCoOp), on the task of identifying birds, the output image feature, without further tuning, can be distracted to leaves of the same color. Similarly, it also wrongly highlights the beer foam that is of a similar shape to recognize golf balls. Since the final recognition is jointly inferred by both text and image branches, such an issue may degrade the classification performance. Thus it is necessary

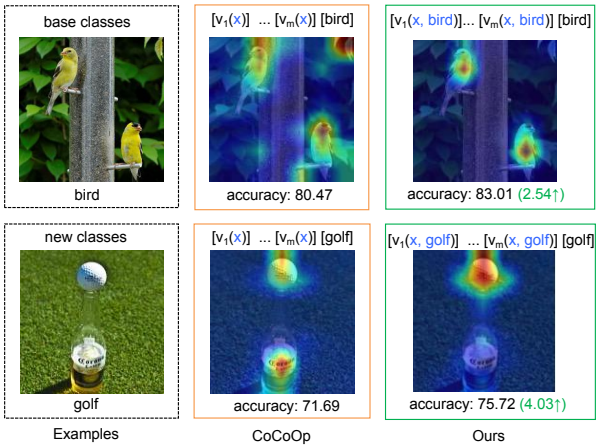


Figure 2: Comparisons of attention map visualization for CoCoOp and our method on ImageNet. Our method obtains better average accuracy of both base and new classes across 11 datasets by paying attention to task-related regions.

to tune the image features further so that the image branch can focus more on the tasks-related representation. We then propose Text-guided Feature Tuning (TFT), which leverages encoded text embedding to guide image representation more on task-related regions. As shown in Fig. 2 (ours), our method successfully focuses on task-related regions, *i.e.*, birds and golf balls. We then leverage the contrastive loss function to further align class-aware text embedding and text-guided image features on certain downstream tasks.

In summary, we propose a new task-oriented multi-modal mutual-learning method, which well-integrates our designed class-aware text prompts and text-guided feature tuning for fast adaptation of frozen VLMs on downstream tasks. Image features can help construct image-dependant class-aware text prompts, leading to more discriminative text embedding. Simultaneously, improved text embedding can further guide the image branch attending to class-related representation. In this way, these two different modality branches can be tightly coupled and mutual-beneficial across the whole training process. Our main contributions are summarized in the following.

- We propose class-aware text prompts which generate prompts based on task-relevant image semantics instead of complete visual information. In this way, we improve the classification accuracy of unseen classes without introducing extra learning ambiguities.

- We propose text-guided feature tuning which enforces image branch to pay more attention to the task-related representation. As a result, the model avoids deviating attention to the task-irrelevant regions of the image.

- Benefiting from our mutual learning strategy, our

method achieves SOTA results on four downstream tasks. Especially, ours significantly outperforms existing methods on the base-to-new generalization task.

## 2. Related work

**Vision language models (VLM).** The current VLM can be roughly divided into four categories based on the training objectives: image-text matching [3, 20, 27], contrastive loss [19, 21, 22], masked language modeling [38, 39, 46], and masked image modeling [3, 27, 38]. As a milestone, CLIP utilizes 400 million image-text pairs to train a large-scale multi-modal model and demonstrates promising performance on a wide spectrum of tasks including few-shot and zero-shot visual recognition. Motivated by this work, numerous follow-ups have been proposed to improve the effectiveness (e.g., FLIP [24], A-CLIP [45], MaskCLIP [7], and SLIP [30]) or apply it to other domains (e.g., DenseCLIP [34] and ActionCLIP [42]). The primary limitation of these methods is that hand-crafted prompts are dataset-sensitive and difficult to optimize. We design an automatic and learnable prompts method to enhance the generalization performance of pre-trained models on downstream tasks.

**Prompt learning in NLP.** As the scale and complexity of pre-trained language models continue to grow, fine-tuning for specific tasks is becoming increasingly expensive. In contrast, prompt-based approaches are an efficient and lightweight alternative that can be used to generate high-quality text with much lower computational requirements. The original prompts were manually designed prompt templates. While manually designing prompts is advantageous due to their intuitive and comprehensible nature, it also presents a significant challenge that demands extensive experimentation, experience, and language expertise, resulting in high costs. To overcome the limitations of manual prompt design, numerous studies have initiated research into automatically learn appropriate prompts. The automatic prompts can be categorized into two types: discrete prompts and continuous prompts. Discrete prompts consist of various approaches such as prompt mining [17], prompt paraphrasing [49, 10], gradient-based search [40], prompt generation [9] and prompt scoring [5]. On the other hand, continuous prompts include techniques such as prefix tuning [23], tuning initialized with discrete prompts [36] and hard-soft prompt hybrid tuning [26]. These methods have also been applied to the field of computer vision for prompt learning research. However, the task of prompt learning in computer vision is often considered more challenging than in natural language due to the relatively limited high-level semantic information present in visual data with raw pixels.

**Prompt learning in vision language models.** Prompt learning has been demonstrated to be an effective method for improving the performance of pre-trained language models on downstream tasks. Recently, prompt learning has gained increased attention in the context of vision language models. For example, CoOp [51] employs learnable vectors to model contextual words as prompts, and demonstrates that automatic prompts outperform hand-crafted prompts in downstream tasks. CoCoOp [50] extends CoOp by incorporating lightweight neural networks to dynamically generate prompts based on each image, thus mitigating sensitivity to class shifts. Different from the above methods, VP [1], VPT [43], and EVP [16] prompt with images. VP [1] directly combines learnable prompts and pixel-wise input images as new inputs to the model. EVP [43] shrinks the original image before padding the prompts around it, to avoid destroying the original image information. VPT [16] introduces a small amount learnable parameters into the input sequence of each transformer layer and learns them together with a linear head during fine-tuning. Building on the prompt learning approach of the text branch, we propose class-aware text prompt that generates image-dependent and class-aware prompts. Similarly, follow the feature tuning of image branch, we introduce text-guided tuning, which directs the image branch to focus on the task-relevant local regions rather than the global information.

## 3. Method

### 3.1. Comparisons of CLIP, CoOp, and CoCoOp

**CLIP** comprises two encoders: an image encoder and a text encoder. The image encoder, denotes by $F(x)$, converts an image $x \in \mathbb{R}^{3\times H\times W}$ with height of $H$ and width of $W$ into a $d$-dimensional image feature $f_x \in \mathbb{R}^{N\times d}$, where $N$ is the number of split patches. Meanwhile, the text encoder, denoted as $G(t)$, generates an $d$-dimensional text representation $g_t \in \mathbb{R}^{M\times d}$ from natural language text $t$, where $M$ is the number of classes. Two encoders are jointly trained using a contrastive loss function that maximizes the cosine similarity of matched pairs and minimizes that of the unmatched pairs. After training, CLIP can be directly used for zero-shot image recognition without requiring fine-tuning of the whole model. Since CLIP is pre-trained on whether an image matches a textual description, the hand-crafted prompt template is employed to convert raw labels into textual descriptions. The most common form of template in CLIP is "a photo of a [CLASS]", where the class token is replaced with specific class names such as "cat", "dog", "car", etc. We let the image features $f_x$ of an image $x$ be extracted by an image encoder and the text features $g_t$ be obtained by feeding the prompt description into the text encoder. The prediction task is defined as the classification of an image into one of $C$ categories, which are represented by the set $y \in \{1, \ldots, C\}$. Denote $y$ as the predicted category. Let $g_t^i$ be the $i$-th dimension of text features $g_t$, with image features

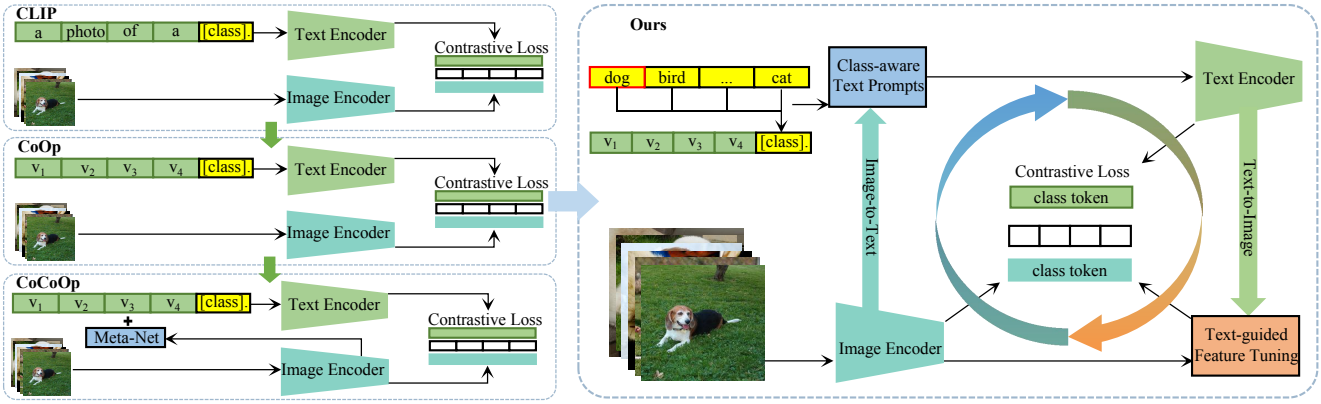


Figure 3: Comparisons of three representative prompt learning techniques and our method. The main differences lie in how the text and image branches focus on downstream tasks. CLIP artificially designs prompt templates. CoOp designs automatic prompts using learnable parameters. CoCoOp directly allows text branch to focus on images semantic through Meta-Net. We introduce Class-aware Text Prompts (CTP) and Text Feature Tuning (TFT) to the text and image branches, respectively. The CTP generates prompts based on class-related image information instead of using the identical image semantics like CoCoOp. The TFT enables the image branch to directly focus on downstream tasks. We leverage the contrastive loss function to align task-oriented text and images, making them promote each other for achieving better downstream generalization performance.

$f_x$, we have the predicted probability of the $i$-th class:

$$P(y=i \mid x)=\frac{\exp\left(\cos\left(f_x, g_t^i\right)/\tau\right)}{\sum_{j=1}^{C}\exp\left(\cos\left(f_x, g_t^j\right)/\tau\right)}, \quad (1)$$

where $\cos(\cdot,\cdot)$ denotes the cosine similarity and $\tau$ is the temperature parameter of the softmax function.

**CoOp** replaces the hand-crafted prompts with automatically generated prompts. Specifically, CoOp introduces $k$ learnable context vectors $\{v_1,\ldots,v_k\}$ to model the context words of the prompts. We define $c_i$ as the word embedding of the $i$-th class name. Then, the prompt of $i$-th class is denoted as $p_i=\{v_1,\ldots,v_k,c_i\}$. Therefore, we have the predicted probability of the $i$-th class using CoOp method:

$$P(y=i \mid x)=\frac{\exp\left(\cos\left(f_x, G\left(p_i\right)\right)/\tau\right)}{\sum_{j=1}^{C}\exp\left(\cos\left(f_x, G\left(p_j\right)\right)/\tau\right)}, \quad (2)$$

where $G(p_i)$ is the text embedding from text encoder $G$.

**CoCoOp** extends CoOp by generating image-conditional prompts. Specifically, CoCoOp uses Meta-Net to generate the residual vector $\pi$ based on each image. Each context token is now obtained by $v_k(x)=v_k+\pi$. The prompt of the $i$-th class $c_i$ is defined as $p_i(x)=\{v_1(x),\ldots,v_k(x),c_i\}$. As a result, the prediction probability of the $i$-th class is:

$$P(y=i \mid x)=\frac{\exp\left(\cos\left(f_x, G\left(p_i(x)\right)\right)/\tau\right)}{\sum_{j=1}^{C}\exp\left(\cos\left(f_x, G\left(p_j(x)\right)/\tau\right)\right.}, \quad (3)$$

where $G\left(p_i(x)\right)$ is the the text embedding conditional on the image $x$ from the text encoder $G$.

### 3.2. Our Task-Oriented Mutual Learning Method

Our method consists of two modules, i.e., **Class-aware Text Prompts (CTP)** and **Text-guided Feature Tuning (TFT)**, as shown in Fig. 3. Compared to CoCoOp, we use CTP to generate class-aware prompts based on task-relevant local image regions instead of the global information. Besides, we use CTP to make the image branch directly pay attention to the task-related image region. We let the two modules be tightly coupled and mutual-beneficial across the training process by optimizing the contrastive loss function.

**CTP** learns image conditioned discriminative prompts for finer paying attention to semantic-related regions of the images. Specifically, in order to obtain the text semantic-related regions of the image, we leverage the prompt $p$ and the image feature $f_x$ to calculate the attention matrix $A^t$:

$$A^t = p f_x^T, \quad (4)$$

where $A^t \in \mathbb{R}^{M\times N}$ is the image-to-text attention map. $A^t_{i,j}$ represents the similarity between the $i$-th class in the text prompts and the $j$-th patch in the image. In this way, we can query the regions of the images that semantically related to the class information by the attention matrix $A^t$. That is,

$$f_x^t = \text{softmax}\left(A^t\right) f_x, \quad (5)$$

where $f_x^t$ is the regions correlated to the text of a specific class. We use it to obtain augmented class-aware prompts:

$$p^a = p + f_x^t, \quad (6)$$

where $p^a$ is the text prompts enhanced by semantically-relevant image regions. Let $p_i^a$ be the $i$-th dimension of $p^a$,

(a) **Average over 11 datasets**

| | Base | New | Hos |
|---|---|---|---|
| CLIP | 69.34 | 74.22 | 71.70 |
| CoOp | 82.69 | 63.22 | 71.66 |
| CoCoOp | 80.47 | 71.69 | 75.83 |
| ProDA | 81.56 | 72.30 | 76.65 |
| Ours | **83.01** | **75.72** | **79.02** |

(b) ImageNet

| | Base | New | Hos |
|---|---|---|---|
| CLIP | 72.43 | 68.14 | 70.22 |
| CoOp | 76.47 | 67.88 | 71.92 |
| CoCoOp | 75.98 | 70.43 | 73.10 |
| ProDA | 75.40 | 70.23 | 72.72 |
| Ours | **77.42** | **70.44** | **73.77** |

(c) Caltech101

| | Base | New | Hos |
|---|---|---|---|
| CLIP | 96.84 | 94.00 | 95.40 |
| CoOp | 98.00 | 89.81 | 93.73 |
| CoCoOp | 97.96 | 93.81 | 95.84 |
| ProDA | 98.27 | 93.23 | 95.68 |
| Ours | **98.31** | **94.75** | **96.50** |

(d) OxfordPets

| | Base | New | Hos |
|---|---|---|---|
| CLIP | 91.17 | 97.26 | 94.12 |
| CoOp | 93.67 | 95.29 | 94.47 |
| CoCoOp | 95.20 | 97.69 | 96.43 |
| ProDA | 95.43 | **97.83** | 96.62 |
| Ours | **95.86** | 97.55 | **96.70** |

(e) StanfordCars

| | Base | New | Hos |
|---|---|---|---|
| CLIP | 63.37 | **74.89** | 68.65 |
| CoOp | **78.12** | 60.40 | 68.13 |
| CoCoOp | 70.49 | 73.59 | 72.01 |
| ProDA | 74.70 | 71.20 | 72.91 |
| Ours | 76.29 | 74.17 | **75.22** |

(f) Flowers102

| | Base | New | Hos |
|---|---|---|---|
| CLIP | 72.08 | 77.80 | 74.83 |
| CoOp | 97.60 | 59.67 | 74.06 |
| CoCoOp | 94.87 | 71.75 | 81.71 |
| ProDA | **97.70** | 68.68 | 80.66 |
| Ours | 97.36 | **77.70** | **86.43** |

(g) Food101

| | Base | New | Hos |
|---|---|---|---|
| CLIP | 90.10 | 91.22 | 90.66 |
| CoOp | 88.33 | 82.26 | 85.19 |
| CoCoOp | **90.70** | 91.29 | 90.99 |
| ProDA | 90.30 | 88.57 | 89.43 |
| Ours | 90.54 | **92.31** | **91.42** |

(h) FGVCAircraft

| | Base | New | Hos |
|---|---|---|---|
| CLIP | 27.19 | **36.29** | 31.09 |
| CoOp | **40.44** | 22.30 | 28.75 |
| CoCoOp | 33.41 | 23.71 | 27.74 |
| ProDA | 36.90 | 34.13 | 35.46 |
| Ours | 39.49 | 35.37 | **37.32** |

(i) SUN397

| | Base | New | Hos |
|---|---|---|---|
| CLIP | 69.36 | 75.35 | 72.23 |
| CoOp | 80.60 | 65.89 | 72.51 |
| CoCoOp | 79.74 | 76.86 | 78.27 |
| ProDA | 78.67 | 76.93 | 77.79 |
| Ours | **82.16** | **77.49** | **79.76** |

(j) DTD

| | Base | New | Hos |
|---|---|---|---|
| CLIP | 53.24 | 59.90 | 56.37 |
| CoOp | 79.44 | 41.18 | 54.24 |
| CoCoOp | 77.01 | 56.00 | 64.85 |
| ProDA | **80.67** | 56.48 | 66.44 |
| Ours | 79.47 | **61.53** | **69.36** |

(k) EuroSAT

| | Base | New | Hos |
|---|---|---|---|
| CLIP | 56.48 | 64.05 | 60.03 |
| CoOp | **92.19** | 54.74 | 68.69 |
| CoCoOp | 87.49 | 60.04 | 71.21 |
| ProDA | 83.90 | 66.00 | 73.88 |
| Ours | 92.14 | **73.87** | **82.00** |

(l) UCF101

| | Base | New | Hos |
|---|---|---|---|
| CLIP | 70.53 | 77.50 | 73.85 |
| CoOp | 84.69 | 56.05 | 67.46 |
| CoCoOp | 82.33 | 73.45 | 77.64 |
| ProDA | **85.23** | 71.97 | 78.04 |
| Ours | 84.12 | **77.74** | **80.80** |

Table 1: Results (%) of the **base-to-new generalization task** on 11 benchmark datasets. We report the accuracy with CLIP ViT-B/16 model on the base classes (Base), the unseen classes (New), and the harmonic mean of both of them (Hos).

we then have the predicted probability of the $i$-th class:

$$P(y=i \mid x)=\frac{\exp\left(\cos\left(f_x, G(p_i^a)\right)/\tau\right)}{\sum_{j=1}^{C}\exp\left(\cos\left(f_x, G(p_j^a)/\tau\right)\right.}. \quad (7)$$

We generate class-aware prompts instead of fusing identical image semantics with prompts of different classes, bringing category discrimination to the specific downstream tasks.

**TFT** leverages text features to guide images to focus on task-related regions. Specifically, using the embeddings $g^a$ of the augmented prompts $p^a$ as input, we have attention:

$$A^x = f_x(g^a)^T, \quad (8)$$

where $A^x \in \mathbb{R}^{N\times M}$ denotes text-to-image attention map. $A^x_{i,j}$ represents the similarity between the $i$-th patch in the image and the $j$-th class in the text representation. Similar to image-to-text, we use it to query the class-related part of the text correlated to the image, augmenting image features:

$$f^a = \mathrm{softmax}(A^x)g^a + f_x, \quad (9)$$

where $f^a$ is the augmented image embeddings. We thus let image branch focus on the tasks-related representation.

**Augmented contrastive loss function** is then employed to further align class-aware text embedding and text-guided image features on specific downstream tasks. The predicted probability of the $i$-th class, which is used to calculate the contrastive loss, after mutual augmentation is:

$$P(y=i \mid x)=\frac{\exp\left(\cos\left(f^a, g_i^a\right)/\tau\right)}{\sum_{j=1}^{C}\exp\left(\cos\left(f^a, g_i^a/\tau\right)\right.}. \quad (10)$$

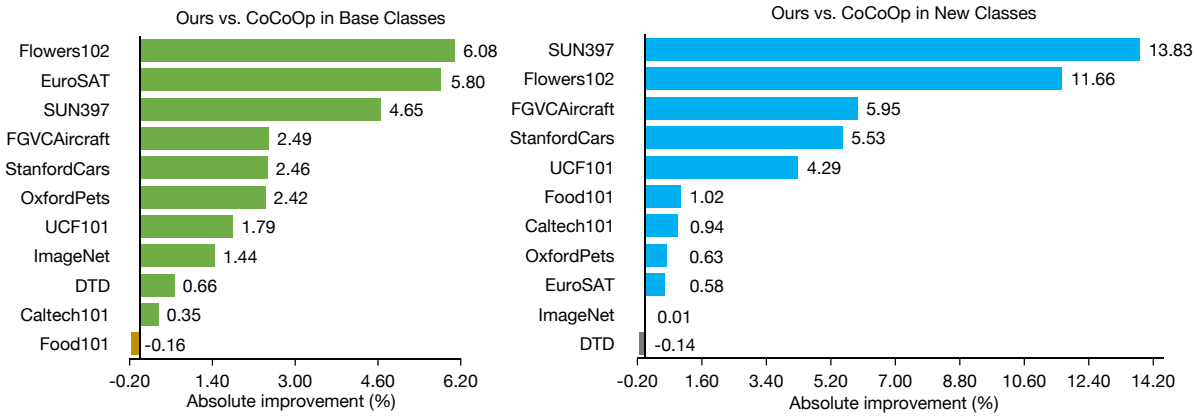


Figure 4: Absolute improvement over CoCoOp in the base-to-new generalization task. Compared to CoCoOp, Our method achieves improvement on both base (left sub-figure) and new (right sub-figure) classes on most of the datasets.

| | ImageNet | Caltech101 | OxfordPets | StanfordCars | Flowers102 | Food101 | FGVCAircraft | SUN397 | DTD | EuroSAT | UCF101 | *Average* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CLIP | 68.63 | 89.36 | 88.99 | 65.67 | 70.49 | 89.23 | 27.12 | 65.29 | 46.02 | 54.17 | 69.83 | 66.80 |
| CoOp | 71.51 | 95.53 | 93.31 | 74.25 | 95.70 | 87.23 | 34.18 | 74.82 | 68.46 | 77.82 | 77.29 | 77.28 |
| CoCoOp | 71.02 | 93.43 | 93.93 | 71.21 | 87.34 | 87.39 | 32.03 | 72.32 | 63.84 | 72.78 | 77.40 | 74.79 |
| Ours | **72.90** | **95.90** | **93.96** | **79.10** | **96.73** | **89.95** | **38.72** | **79.37** | **72.49** | **81.00** | **83.45** | **80.32** |

Table 2: Results (%) of **16-shot learning task** on 11 datasets.

Task-targeted semantic information is transferred between the two branches by minimizing the augmented contrastive loss. We merge probability before and after augmentation:

$$P(y=i \mid x)=\frac{\exp((\cos(f, g_i)+\lambda(\cos(f^a, g_i^a))/\tau)}{\sum_{j=1}^{C}\exp((\cos(f, g_i)+\lambda(\cos(f^a, g_i^a))/\tau)}. \tag{11}$$

where $\lambda$ is the balance hyper-parameter, which is analyzed in our experiments. We let the two different modalities tightly coupled and mutual beneficial across the whole training process by performing the contrastive optimization.

## 4. Experiments

We evaluate the performance of our method on four generalization tasks, including 1) generalization from base classes to new classes; 2) few-shot classification; 3) cross-dataset transfer; 4) domain generalization. After that, we provide extensive ablation studies and in-depth analyses.

**Datasets.** Following [33, 51], we use 11 image recognition datasets for the tasks of base-to-new generalization, few-shot classification and cross-dataset transfer. It contains generic image classification datasets (ImageNet [6] and Caltech101 [8]), fine-grained classification datasets (Oxford Pets [32] , StanfordCars [18], Flowers102 [31], Food101 [2] and FGVCAircraft [29]), scene recognition ( SUN397 [44]), action recognition (UCF101 [37]), texture classification (DTD [4]), and satellite imagery recognition (EuroSAT [11]). For the domain generalization task, we use ImageNet as the source dataset and select ImageNetV2 [35], ImageNet-Sketch [41], ImageNet-A [13], and ImageNet-R [12], which are the ImageNet variants, as the target.

**Training Details.** By following [50, 51], we use the best visual backbone available in CLIP, i.e., ViT-B/16, throughout the experiments. We train 10 epochs using SGD optimizer with base learning rate of 0.002 and cosine decay schedule. We set the hyper-parameter $\lambda$ in Eq. (11) to 0.2 for all experiments, and provide sensitivity analyses in Fig. 5. We run all the experiments three times with different random seeds and report the average classification accuracy.

**Baselines.** We compare our method with 4 baselines. (1) Zero-shot CLIP [33] with hand-crafted prompts. (2) CoOp [51], using automatically generated prompts from few data. (3) CoCoOp [50], dynamically generating prompts conditioned on the images. (4) ProDA [28], which learns prompts from few data samples and mitigates the domain gap.

### 4.1. Generalization From Base to New Classes

Following the previous works, we split the classes equally into two groups for each dataset: one as base and

| | Source | Target | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ImageNet | Caltech101 | OxfordPets | StanfordCars | Flowers102 | Food101 | FGVCAircraft | 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 |
| 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 |
| Ours | **72.90** | **95.73** | **90.22** | 65.14 | 69.89 | **86.38** | **23.32** | 66.49 | **46.47** | **47.24** | 67.43 | **66.47** |

Table 3: Results of **cross-dataset transfer task**. Each method is trained on the source dataset and evaluated on the target.

| | 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 |
| CoCoOp | 71.02 | 64.07 | 48.75 | 50.63 | 76.18 |
| Ours | **72.90** | **64.57** | **49.11** | **50.94** | **76.68** |

Table 4: Results of **domain generalization task**. Each method is trained on ImageNet and evaluated on ImageNet variants.

the other as new. The learnable modules are trained exclusively on the base classes, while evaluation is carried out separately on both the base and new classes to testify generalization ability. We report the results on 11 benchmarks in Table 1. Although compared to CoOp, CoCoOp significantly narrows generalization gap in unseen classed, but it decreases the accuracy in seen classes from 82.69% to 80.47%. We attribute it to the homogeneous prompts of CoCoOp, which weakens the discriminative semantics of different categories. In comparison, our method improves the accuracy in seen classes from 80.47% to 83.01% by prompting each text label with corresponding image information. Benefit from the mutual learning of our CTP and TFT modules, our method further improves the accuracy in unseen classes from 71.69% to 75.72 %, even surpasses the accuracy of CLIP hand-crafted prompts. We provide a detailed comparisons of CoCoOp and our method of per-dataset improvement in Fig. 4. Our method gains significant improvements over CoCoOp in both seen and unseen classes on 10 out of 11 recognition datasets. Surprisingly, our method significantly improves CoCoOp by more than 10% in unseen classes on SUN397 and Flowers102 datasets.

### 4.2. Few-Shot Classification

We report few-shot classification results in Table 2. Our method surpasses baseline methods on all datasets in the few-shot setting. Especially, our method outperforms CoCoOp by 9.39%, 8.65%, and 8.22% on Flowers102, DTD, and EuroSAT, respectively, and the average improvement over 11 datasets is 5.53%. Our method also achieve 2% on the challenging dataset of ImageNet. The above experiments shows the great discriminative ability of our method.

### 4.3. Cross-Dataset Transfer

We then evaluate the generalization ability of our method on more challenging cross-dataset tasks. In this setting, we learn multi-modal prompts on ImageNet of 1000 classes. The effectiveness of the learned prompts is then tested on 10 datasets containing generic and fine-grained image classification, scene recognition, and texture classification. The results are reported in Table 3. Our method achieves the best average accuracy on the 11 datasets, especially ImageNet. It demonstrates the great transfer ability of our method.

### 4.4. Domain Generalization

The domain generalization setting evaluates the generalization ability of the model on the target domain that is similar to but different from the source domain [48, 47]. Zero-shot CLIP introduces no additional training parameters and exhibits great robustness to naturally distribution shifts. Other methods use few samples to train learnable parameters, there is a risk of overfitting the source distribution. Therefore, we conduct experiments using ImageNet as the source domain and evaluate the ability of generalizing to unknown on four ImageNet variants. The results are shown in Table 4. Our method achieves significant performance on the 4 ImageNet variant datasets. It verifies that our method improves the classification ability of the source domain dataset while maintaining the generalization on the target domain.

### 4.5. Ablation Analysis

**Effectiveness of each module.** To evaluate the effectiveness of Class-aware Text Prompts (CTP) and Text-guided Feature Tuning (TFT) of our method, we conduct ablation

| Method | Average | | | ImageNet | | | Caltech101 | | | OxfordPets | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Base | New | Hos | Base | New | Hos | Base | New | Hos | Base | New | Hos |
| A | 82.69 | 63.22 | 71.66 | 76.47 | 67.88 | 71.92 | 98.00 | 89.81 | 93.73 | 93.67 | 95.29 | 94.47 |
| B | 82.38 | 72.44 | 76.64 | 76.96 | 69.62 | 73.11 | 98.23 | 94.24 | 96.19 | 95.61 | **97.97** | **96.78** |
| C | 82.93 | 72.98 | 77.24 | 77.21 | 69.86 | 73.35 | **98.44** | 92.72 | 95.49 | 95.49 | 97.81 | 96.64 |
| Ours | **83.01** | **75.72** | **79.02** | **77.42** | **70.44** | **73.77** | 98.31 | **94.75** | **96.50** | **95.86** | 97.55 | 96.70 |
| Method | StanfordCars | | | Flowers102 | | | Food101 | | | FGVCAircraft | | |
| | Base | New | Hos | Base | New | Hos | Base | New | Hos | Base | New | Hos |
| A | **78.12** | 60.40 | 68.13 | **97.60** | 59.67 | 74.06 | 88.33 | 82.26 | 85.19 | **40.44** | 22.30 | 28.75 |
| B | 74.62 | 73.68 | 74.15 | 96.72 | 66.42 | 78.76 | 90.30 | 91.47 | 90.88 | 36.41 | 34.39 | 35.37 |
| C | 75.84 | **74.53** | 75.18 | 97.32 | 74.86 | 84.63 | **90.56** | 91.65 | 91.10 | 37.82 | 33.17 | 35.34 |
| Ours | 76.29 | 74.17 | **75.22** | 97.36 | **77.70** | **86.43** | 90.54 | **92.31** | **91.42** | 39.49 | **35.37** | **37.32** |
| Method | SUN397 | | | DTD | | | EuroSAT | | | UCF101 | | |
| | Base | New | Hos | Base | New | Hos | Base | New | Hos | Base | New | Hos |
| A | 80.60 | 65.89 | 72.51 | 79.44 | 41.18 | 54.24 | **92.19** | 54.74 | 68.69 | 84.69 | 56.05 | 67.46 |
| B | 81.73 | 76.89 | 79.24 | 80.18 | 51.79 | 62.93 | 91.70 | 67.62 | 77.84 | 83.72 | 72.72 | 77.83 |
| C | **82.29** | 76.24 | 79.15 | **81.71** | 54.74 | 65.56 | 90.10 | 60.52 | 72.41 | **85.40** | 76.68 | **80.81** |
| Ours | 82.16 | **77.49** | **79.76** | 79.47 | **61.53** | **69.36** | 92.14 | **73.87** | **82.00** | 84.12 | **77.74** | 80.80 |

Table 5: **Ablation studies** of our method on 11 datasets. Three ablation cases are considered: **A**: Ours w/o TVP w/o FTP. **B**: Ours w/o TVP. **C**: Ours w/o FTP. TVP is the text-reorganized vision prompt, and FTP is the fine-grained text prompt.

| Prompt Learning | | Feature Tuning | | Accuracy (%) |
|---|---|---|---|---|
| MLP-PL | CTP | MLP-FT | TFT | |
| | | | | 71.66 (CoOp) |
| ✓ | | | | 75.83 (+4.17) |
| | ✓ | | | 76.64 (+4.98) |
| | | ✓ | | 75.94 (+4.28) |
| | | | ✓ | 77.24 (+5.58) |
| ✓ | | ✓ | | 77.05 (+5.39) |
| | ✓ | | ✓ | **79.02** (+7.36) |

Table 6: Comparison of different structures for prompt learning and feature tuning. The average results of harmonic mean of from-base-to-new generalization task on 11 datasets are reported. In compared to our attention design in CTP and TFT modules, MLP-PL and MLP-FT are designed using the Linear-ReLU-Linear block setting of [50]. Improvements over the baseline of CoOp, are marked in green.

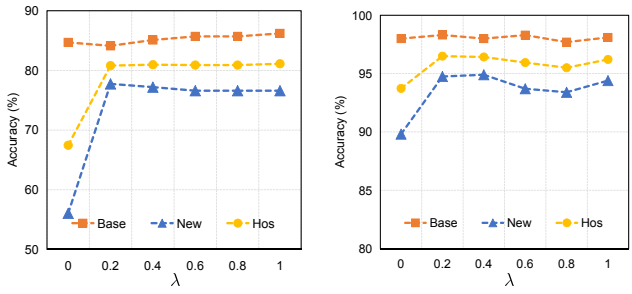


Figure 5: **Sensitivity analysis of** $\lambda$, with base, new, and hos metrics, on UCF101 (left) and Caltech101 (right) datasets.

experiments on 11 datasets, as reported in Table 5. In most cases, each module significantly improves the performance of the model. For average results, CTP and TFT improves the results by 5.58% and 4.98%, respectively, and the combination of them improves the results by 7.36%. It show the effectiveness of the two branches of text-to-image and image-to-text, the mutual learning of the two modules further improves the performance on downstream tasks.

**Comparison of different structure design of multi-modal mutual learning.** To further provide in-depth analysis about our mutual learning, we further explore two vanilla structures: (1) MLP-PL: The image features are forwarded to a block of Linear-ReLU-Linear, borrowed from [50], and then added to the text for augmenting it (the same to us). (2) MLP-FT: The text prompts are forwarded to the Linear-ReLU-Linear block, and then added to the image for augmenting it (the same to us). In comparison, our class-aware text prompts (CTP) module and text-guided feature tuning (TFT) module, adopt text-image attention to learn the augmented features instead of the Linear-ReLU-Linear block. We report the results of the different designs in Table 6. First, we find that combining MLP-PL & MLP-FT and CTP & TFT can both improve the results compared with using either of them. It indicates that both prompt learning and feature tuning are important to achieve better results. Second, compared with the design of Linear-ReLU-Linear block, our design of text-image attention further improves performance by 0.81% and 1.3% for prompt learning and feature tuning, respectively. It demonstrates the effectiveness of our design of attention, which helps the model to focus on class-aware and task-related semantics. Third, com-

pared with CoOp, both of the designs could improve the final results by large margins. The key factor of our mutual learning to achieve significant performance is the task-related alignment of vision and language in latent space.

**Sensitivity Analysis of $\lambda$.** We evaluate the parameter sensitivity of $\lambda$ of Eq. (11) in Fig. 5. The results suggest that the performance of our method is generally robust to $\lambda$, indicating a wide range of $\lambda$ works well in downstream tasks.

## 5. Conclusion

In this paper, we introduce task-oriented multi-modal mutual learning for adapting large vision-language models to downstream vision tasks. We propose class-aware text prompt and text-guided feature tuning to unleash the potential of the vision-language model by re-activating its task-related representation abilities. Our method yields impressive generalization performance on a wide range of vision tasks and datasets. We hope the presented findings and insights in this paper could benefit the following works in designing more efficient and effective adaptation methods. For the future work, we think it is interesting to extend the adaptation of vision language models to more vision tasks, such as semantic segmentation, object detection, etc.

**Limitations.** Similar to CoCoOp, we learn image-conditioned representations, thus the batch-size of training need to be set to 1, which is not efficient enough for learning. We aim to solve this efficiency issue in the future work.

## 6. Acknowledgements.

This work is supported by the Innovation Capacity Construction Project of Jilin Province Development and Reform Commission (2021FGWCXNLJSSZ10) and the Fundamental Research Funds for the Central Universities, JLU. Zichang Tan is the project lead.

## References

[1] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 1(3):4, 2022. 3

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pages 446–461. Springer, 2014. 6

[3] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, pages 104–120. Springer, 2020. 3

[4] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 6

[5] Joe Davison, Joshua Feldman, and Alexander M Rush. Commonsense knowledge mining from pretrained models. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 1173–1178, 2019. 3

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[7] Xiaoyi Dong, Yinglin Zheng, Jianmin Bao, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. *arXiv preprint arXiv:2208.12262*, 2022. 3

[8] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. 6

[9] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*, 2020. 3

[10] Adi Haviv, Jonathan Berant, and Amir Globerson. Bertese: Learning to speak to bert. *arXiv preprint arXiv:2103.05327*, 2021. 3

[11] 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. 6

[12] 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*, pages 8340–8349, 2021. 6

[13] 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*, pages 15262–15271, 2021. 6

[14] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019. 1

[15] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 1

[16] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Computer Vision–ECCV 2022: 17th*

*European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, pages 709–727. Springer, 2022. 3

[17] Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438, 2020. 3

[18] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 6

[19] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 3

[20] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 3

[21] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*, 2020. 3

[22] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020. 3

[23] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021. 3

[24] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. *arXiv preprint arXiv:2212.00794*, 2022. 3

[25] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023. 1

[26] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. Gpt understands, too. *arXiv preprint arXiv:2103.10385*, 2021. 3

[27] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 3

[28] 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*, pages 5206–5215, 2022. 1, 2, 6

[29] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 6

[30] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, pages 529–544. Springer, 2022. 3

[31] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008. 6

[32] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 6

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 6

[34] 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*, pages 18082–18091, 2022. 3

[35] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019. 6

[36] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*, 2020. 3

[37] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. A dataset of 101 human action classes from videos in the wild. *Center for Research in Computer Vision*, 2(11), 2012. 6

[38] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 3

[39] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 3

[40] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. Universal adversarial triggers for attacking and analyzing nlp. *arXiv preprint arXiv:1908.07125*, 2019. 3

[41] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019. 6

[42] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 3

[43] Junyang Wu, Xianhang Li, Chen Wei, Huiyu Wang, Alan Yuille, Yuyin Zhou, and Cihang Xie. Unleashing the power

of visual prompting at the pixel level. *arXiv preprint arXiv:2212.10556*, 2022. 3

[44] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 6

[45] Yifan Yang, Weiquan Huang, Yixuan Wei, Houwen Peng, Xinyang Jiang, Huiqiang Jiang, Fangyun Wei, Yin Wang, Han Hu, Lili Qiu, et al. Attentive mask clip. *arXiv preprint arXiv:2212.08653*, 2022. 3

[46] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021. 3

[47] Junkun Yuan, Xu Ma, Defang Chen, Kun Kuang, Fei Wu, and Lanfen Lin. Label-efficient domain generalization via collaborative exploration and generalization. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2361–2370, 2022. 7

[48] Junkun Yuan, Xu Ma, Defang Chen, Kun Kuang, Fei Wu, and Lanfen Lin. Domain-specific bias filtering for single labeled domain generalization. *International Journal of Computer Vision*, 131(2):552–571, 2023. 7

[49] Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277, 2021. 3

[50] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 1, 2, 3, 6, 8

[51] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 1, 2, 3, 6