International Journal of Computer Vision
https://doi.org/10.1007/s11263-024-02243-z

Mutual Prompt Leaning for Vision Language Models

Sifan Long[1,2,3] · **Zhen Zhao**[4] · **Junkun Yuan**[3,5] · **Zichang Tan**[3] · **Jiangjiang Liu**[3] · **Jingyuan Feng**[1,2] · **Shengsheng Wang**[1,2] · **Jingdong Wang**[3]

Received: 21 September 2023 / Accepted: 12 September 2024

Abstract

Large pre-trained vision language models (VLMs) have demonstrated impressive representation learning capabilities, but their transferability across various downstream tasks heavily relies on prompt learning. Since VLMs consist of text and visual sub-branches, existing prompt approaches are mainly divided into text and visual prompts. Recent text prompt methods have achieved great performance by designing input-condition prompts that encompass both text and image domain knowledge. However, roughly incorporating the same image feature into each learnable text token may be unjustifiable, as it could result in learnable text prompts being concentrated on one or a subset of characteristics. In light of this, we propose a fine-grained text prompt (FTP) that decomposes the single global image features into several finer-grained semantics and incorporates them into corresponding text prompt tokens. On the other hand, current methods neglect valuable text semantic information when building the visual prompt. Furthermore, text information contains redundant and negative category semantics. To address this, we propose a text-reorganized visual prompt (TVP) that reorganizes the text descriptions of the current image to construct the visual prompt, guiding the image branch to attend to class-related representations. By leveraging both FTP and TVP, we enable mutual prompting between the text and visual modalities, unleashing their potential to tap into the representation capabilities of VLMs. Extensive experiments on 11 classification benchmarks show that our method surpasses existing methods by a large margin. In particular, our approach improves recent state-of-the-art CoCoOp by 4.79% on new classes and 3.88% on harmonic mean over eleven classification benchmarks.

Keywords Vision language models · Prompt learning · Visual prompt · Mutual learning · Visual recognition

1 Introduction

Task-specific supervised models (He et al., 2016; Dosovitskiy et al., 2020) achieve excellent performance on various

Communicated by Wanli Ouyang.

Zhen Zhen, Junkun Yuan, and Zichang Tan have contributed equally to this work.

✉ Shengsheng Wang
wss@jlu.edu.cn

Sifan Long
longsf22@mails.jlu.edu.cn

Zhen Zhao
zhen.zhao@sydney.edu.au

Junkun Yuan
yuanjk@zju.edu.cn

Zichang Tan
tanzichang@baidu.com

Jiangjiang Liu
liujiangjiang@baidu.com

Jingyuan Feng
jingyuan22@mails.jlu.edu.cn

Jingdong Wang
wangjingdong@baidu.com

[1] College of Computer Science and Technology, Jilin University, 2699 Qianjin Street, Changchun 130012, Jilin, China

[2] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, 2699 Qianjin Street, Changchun 130012, Jilin, China

[3] Department of Computer Vision Technology (VIS), Baidu Inc, Beijing, China

[4] School of Electrical and Information Engineering, University of Sydney, Sydney, Australia

[5] College of Computer Science and Technology, Zhejiang University, Hangzhou, China

Published online: 26 September 2024

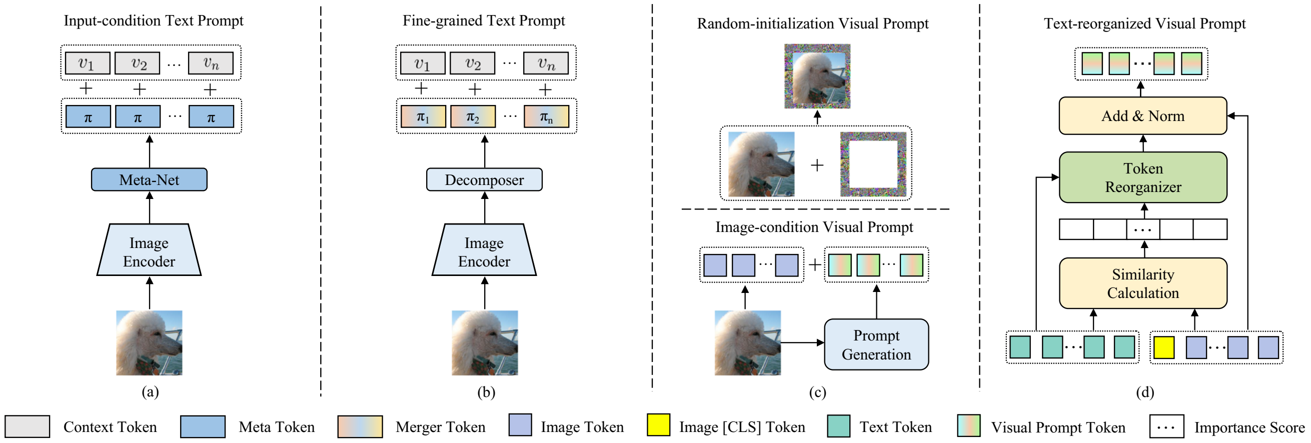

Fig. 1 **a** Input-condition text prompt directly focuses on text and image domain knowledge via incorporating the same image feature into each learnable text token. **b** Our fine-grained text prompt decomposes the single global image features into several finer-grained semantics and incorporates them into corresponding text prompt tokens. **c** Existing visual prompt learning is based on random initialization or image condition, neglecting textual semantic information. **d** Our text-reorganized visual prompt simultaneously explores image and text knowledge and reorganizes the text descriptions of the current image to construct the visual prompt for each image

vision tasks such as classification (Yuan et al., 2021a; Wu et al., 2021; Touvron et al., 2021; Graham et al., 2021), semantic segmentation (Xie et al., 2021; Li et al., 2022b; Liu et al., 2021b; Wang et al., 2021b), and object detection (Carion et al., 2020; Dai et al., 2021). Nevertheless, constraining these models to predict within a fixed set of classes limits their ability to generalize and be versatile. The latest visual-language models (VLMs), exemplified by CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021), have surpassed these limitations, showcasing remarkable representation and generalization capabilities by effectively incorporating language as a supervisory signal. Despite their impressive transferable abilities (Liu et al., 2023a), directly applying pre-trained VLMs to downstream tasks may not achieve satisfied performance due to a potential gap between the pre-training and the task-specific objectives (Liu et al., 2023b). A straightforward strategy for bridging this gap involves fine-tuning, a process that adapts all parameters of the VLMs to suit the requirements of specific downstream tasks, utilizing labelled data (Devlin et al., 2018). However, fine-tuning comes with a set of drawbacks: entailing substantial computational costs, prone to overfitting, and susceptible to catastrophic forgetting (Houlsby et al., 2019).

Recently, prompt tuning, a simple, compact, and viable strategy, has become the leading solution for deploying large pre-trained VLMs into specific downstream tasks. Initially, CLIP (Radford et al., 2021) directly utilizes hand-crafted prompts to achieve impressive zero-shot classification performance. But designing appropriate prompts for each specific task is a non-trivial task. It commonly requires a considerable amount of time and domain knowledge to carefully refine the choice of words. This is because even a slight modification in wording can lead to considerable differences in performance (Zhou et al., 2022b). Inspired by tuning studies in large language models (LLMs) (Li & Liang, 2021; Lester et al., 2021), latter studies like CoOp (Zhou et al., 2022b) and ProDA (Lu et al., 2022), train automatic and learnable text prompts to alleviate such reliance on hard-prompt designs, mainly enhancing the flexibility and effectiveness of tuning process. Recent CoCoOp (Zhou et al., 2022a) points out the limitations of these static uni-modal prompts and introduces a lightweight neural network to encourage text prompts towards a more comprehensive consideration of both textual and visual semantics. Its image-dependent design yields great performance improvements by effectively exploring the few-shot domain knowledge of both modalities.

However, as shown in Fig. 1a, CoCoOp ignores the fact that incorporating the same image feature, denoted as π, into learnable text tokens, denoted as $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$, may be unjustifiable, where such setting may lead to the learnable text prompts focusing on one or a subset of characteristics (Chen et al., 2023). To address this, as shown Fig. 1b, we propose a fine-grained text prompt (FTP) that decomposes the single global image features into several finer-grained semantics, denoted as $\{\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_n\}$, and incorporates them into different text prompt tokens. In this way, our approach achieves the alignment of different text prompt tokens with distinct visual local features rather than the single global features. To be specific, we first leverage the density peak clustering (DPC) algorithm (Rodriguez & Laio, 2014) to cluster similar tokens of image features into the same group. Some clustering examples can be found in Fig. 2. Then, we combine the tokens from the same group into a new one, named as the synthetic token. In this way, we decomposes the single

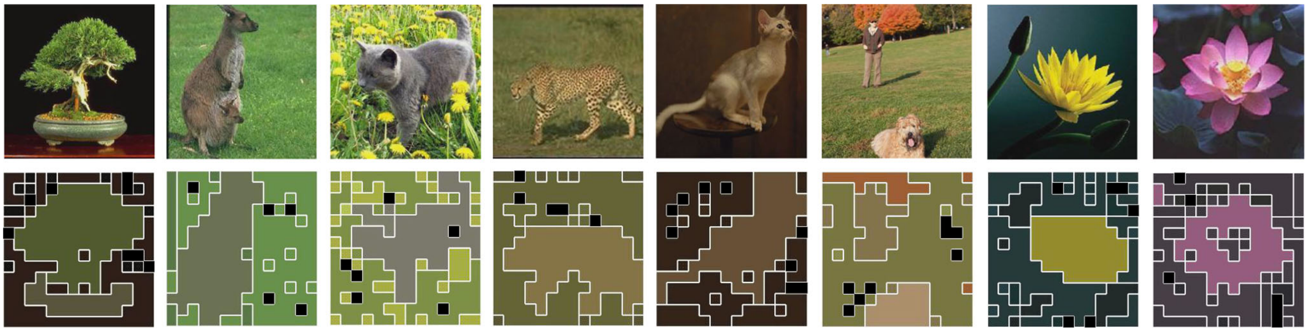

Fig. 2 Visualization of dividing the single global image features into several finer-grained semantics. Tokens are grouped into distinct categories, marked using different colours

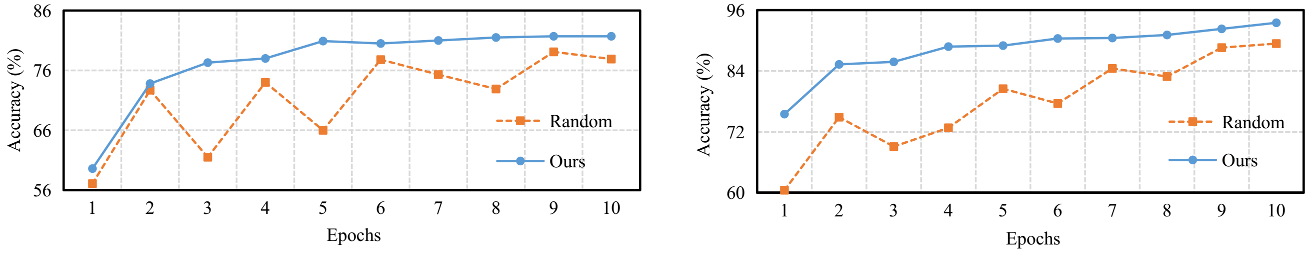

Fig. 3 Comparison of the accuracy with text-reorganized visual prompt and random-initialization visual prompt during training. Left and right sub-figures correspond to the results of the base-to-new generalization task on DTD and EuroSAT datasets, respectively

global image features into several finer-grained semantics. Subsequently, each synthetic token is incorporated into each corresponding text token, thereby generating finer-grained text prompts. This approach enables a more precise alignment between individual text and image tokens.

Diverging from text prompts, which consider both textual and visual domain knowledge (Zhou et al., 2022a; Zang et al., 2022; Khattak et al., 2023; Long et al., 2023b), visual prompts are usually based on random initialization or image condition, as illustrated in Fig. 1c. However, random-initialization visual prompts (Bahng et al., 2022; Wu et al., 2022; Jia et al., 2022) introduce excessive perturbation for the pixel-wise image input, which would lead to unstable training process, thereby resulting in lower classification accuracy (Fig. 3). Moreover, image-condition visual prompts (Loedeman et al., 2022) may overlook valuable textual semantic information when constructing visual prompts. In view of this, we aim to synergistically leverage both text and image domain knowledge to construct the visual prompt. However, directly incorporating text information into visual prompts is not feasible due to the redundant and negative category semantics contained in the text information. Therefore, we propose a text-reorganized visual prompt (TVP) that reorganizes the text descriptions of the current image to construct the visual prompt for each image. As shown in Fig. 1d, the text features consist of M text tokens, where M represents the number of categories. We firstly calculate the cosine similarity between each text token and the `[CLS]` token of image feature as the importance score. In this way, we obtain the importance of each category semantic for the current image in text knowledge. Then, we introduce a token reorganizer that leverages the importance score of each text token as its weight to combine different text tokens into a new text token. Finally, we incorporate the merged text token into each token of image features as the visual prompt. Benefiting from the text-reorganized visual prompt, our approach enables guide the image encoder to focus on class-related representations.

In summary, we propose a new mutual prompt learning framework, dubbed MPL, which seamlessly integrates our designed fine-grained text prompt (FTP) and text-reorganized visual prompt (TVP) for fast adaptation of frozen VLMs on downstream tasks. In specific, FTP employs finer-grained image semantics to enable image-dependant text embeddings, while TVP effectively leverages the text information to encourage the image branch to attend to class-related representations. As shown in Fig. 4, image-dependant text prompts built from image features generate finer-grained text embeddings. Meanwhile, text embeddings are used to construct the visual prompt for text reorganization, resulting in more representative image features. In this way, the image-to-text FTP and text-to-image TVP can be tightly coupled and mutually promoted to enhance the adaptation of VLMs for downstream tasks. Our main contributions are summarized in the following.

- To solve the issues in the current text and visual prompts, we propose a mutual prompt learning (MPL) approach consisting of a fine-grained text prompt and a text-reorganized vision prompt as a strong baseline for reactivating the task-related representation abilities of VLMs.
- We propose a fine-grained text prompt that decomposes the single global image features into several finer-grained semantics and incorporates them into corresponding text prompt tokens. In this way, we achieves the alignment of different text prompt tokens with distinct visual local features rather than the single global features.
- We construct a text-reorganized visual prompt that leverages text information to construct the visual prompt to guide the image branch to attend to class-related representations. In TVP, we reorganize the text descriptions of the current image to remove the redundant and negative category semantics contained in the text information.
- Benefiting from the mutual learning strategy, our method achieves new state-of-the-art (SOTA) results on four downstream tasks. For example, MPL significantly outperforms the current SOTA CoCoOp by 4.79% on new classes and 3.88% on harmonic-mean over eleven classification benchmarks.

This work represents an extension of our conference paper (Long et al., 2023b) published on ICCV 2023. We have substantially expanded upon the initial conference version of our research in the following ways. First, we introduce a new fine-grained text prompt approach to align different text prompt tokens with distinct visual local features rather than the single global features, which is parameter-free, efficient and effective compared to the text prompt used in the conference version (Long et al., 2023b). Second, we extend the original text-guided feature tuning module as a text-reorganized visual prompt, which constructs a text-guided visual prompt to encourage the image branch to attend to class-related representations. Third, we provide more insights into the prompt designs and conduct extensive ablation studies to investigate the influences of textual and visual prompt hyperparameters on model performance, including text prompt initialization, text prompt context length, and visual prompt location. Fourth, we further compare MPL with two vanilla structures and two task-oriented strategies (Long et al., 2023b) to demonstrate the effectiveness and efficiency of our method. All these improvements help us set new SOTA results on eleven benchmark datasets. We hope that our proposed method and various explorations can further inspire future research on fast adaptation of large-scale VLMs and advance the development of this frontier.

2 Related Work

2.1 Vision Language Models (VLMs)

Existing VLMs can be roughly grouped into one-stream and dual-stream model structures. The one-stream architecture (Li et al., 2019; Su et al., 2019; Chen et al., 2020b; Kim et al., 2021) refers to concatenating textual and visual features as input to a single transformer-based encoder. Benefiting from the unordered representation nature of the transformer, the single-stream architecture can handle different input formats in different vision language tasks in a unified framework. For example, VisualBERT (Li et al., 2019) treats the captured image features as unordered input tokens and feeds them together with the text into multiple transformer layers for joint processing. While single-stream architecture parameters that take the same set of parameters for both modalities are more efficient, they may ignore interactions within a single modality. Therefore, another part of the work leverages a dual-stream architecture to separately model the vision and language modalities.

The dual-stream architecture (Zhang et al., 2020; Dou et al., 2022) refers to independently feeding textual and visual features to two different transformer-based encoders. Then, the two encoders project the image features and text embeddings to the same semantic space by using a contrastive loss function (Radford et al., 2021; Jia et al., 2021). For example, CLIP leverages 400 million image-text pairs to train a large-scale multi-modal model, achieving promising representation learning capabilities. Motivated by this work, numerous follow-ups have been proposed to improve the effectiveness (e.g., FLIP (Li et al., 2022a), A-CLIP (Yang et al., 2022), MaskCLIP (Dong et al., 2022), and SLIP (Mu et al., 2022)) or apply it to other domains (e.g., DenseCLIP (Rao et al., 2022) and ActionCLIP (Wang et al., 2021a)). However, their transfer ability relies heavily on prompt learning. To tackle this issue, we design text-reorganized visual prompt (TVP) and fine-grained text prompt (FTP) according to dual-stream architecture. In this way, an automatic, learnable mutual prompt approach enhances the generalization performance of pre-trained models to downstream tasks.

2.2 Prompt Learning in Large Language Models

Recent large language models (LLMs) like T5 (Raffel et al., 2020), GPT-3 (Brown et al., 2020), Bloom (Scao et al., 2022) and LLaMA (Touvron et al., 2023) have shown promising performance in natural language processing tasks. It pushes language tasks to a higher level and has attracted extensive attention in academia and industry.

Although these LLMs can capture rich knowledge from massive corpora, they still need fine-tuning to apply to downstream tasks. As the scale of LLMs continues to increase, it becomes prohibitive to fine-tune all the parameters of the model. Lightweight and efficient prompt learning becomes a new paradigm for adapting LLMs to downstream tasks.

Existing prompt learning approaches are mainly divided into hand-craft and automatic prompts. The initial prompts are intuitive templates manually designed based on human perception. For example, the LAMA dataset (Petroni et al., 2019) manually constructs cloze templates from multiple datasets to explore knowledge in LLMs. However, hand-craft prompt often demands extensive experimentation, experience, and language expertise, resulting in time-consuming and difficult-to-optimize results. To solve these problems, later studies train an automatic and learnable prompt through few-shot corpora knowledge. Automatic prompts can be divided into discrete prompts and continuous prompts. Discrete prompts mainly include prompt mining (Jiang et al., 2020), prompt paraphrasing (Yuan et al., 2021b; Haviv et al., 2021), gradient-based search (Wallace et al., 2019), prompt generation (Gao et al., 2020) and prompt scoring (Davison et al., 2019). Continuous prompts consist of various approaches such as prefix tuning (Li & Liang, 2021), tuning initialized with discrete prompts (Shin et al., 2020) and hard-soft prompt hybrid tuning (Liu et al., 2021a). It inspires research methods for prompt learning in computer vision. Vision contains much less high-level semantic information than language, making the task more challenging.

2.3 Prompt Learning in Vision Language Models

Prompt learning has received increasing attention in adapting vision language models (VLMs) to downstream tasks. Since VLMs consist of text and visual branches, existing prompt learning approaches are mainly divided into text and visual prompt learning. In text prompt learning, CoOp (Zhou et al., 2022b) trains automatically learnable text prompts on the downstream task for the first time. CoCoOp (Zhou et al., 2022a) extends CoOp via designing an input-condition prompt that directly focuses on text and image domain knowledge. However, existing input conditional text prompts that incorporate the same image features into each learnable text token may cause learnable text prompts to focus on one or a subset of characteristics. We propose a fine-grained text prompt (FTP) that decomposes the single global image features into several finer-grained semantics and incorporates them into corresponding text prompt tokens. Among existing methods, the most related to fine-grain prompt tuning (FTP) is the PLOT (Chen et al., 2023). Both FTP and PLOT methods argue that roughly incorporating the same image feature into text prompts may be problematic, as it could result in learnable text prompts being concentrated on one or a subset of characteristics. Differently, PLOT applies optimal transport to align the local visual features and text features indirectly optimizing multiple text prompts. FTP decomposes the single global image features into several finer-grained semantics and directly merges them into the text prompt tokens. Consequently, our method achieves 80.35% on the average of 11 datasets, against 76.20% of PLOT by a large margin.

Unlike the above methods, visual prompt learning provides a new perspective on adapting pre-trained VLMs in vision. VPT (Jia et al., 2022) introduces a small number of learnable parameters into the vision sequence of each transformer layer. VP (Bahng et al., 2022) pads the region around the visual image with learnable parameters as a vision prompt. EVP (Wu et al., 2022) shrinks the original image before padding the prompts around it to avoid destroying the original image information. PGN (Loedeman et al., 2022) generates prompts conditioned on the input images. However, building the visual prompt based on random initialization or image condition both neglect textual semantic information. We propose a text-reorganized visual prompt (TVP) to simultaneously explore both modality knowledge while reorganizing the text descriptions of the current image.

3 Preliminaries

In this section, we first introduce a representative VLM, CLIP, which utilizes hand-crafted prompts in a zero-shot manner for downstream tasks. Next, we discuss existing prompt learning methods, which are mainly divided into text and visual prompt learning. In the text prompt learning, we introduce CoOp (Zhou et al., 2022b), the first soft-prompt approach, and CoCoOp (Zhou et al., 2022a), an input-condition prompt method. On the other hand, we introduce the random initialization manner VP (Bahng et al., 2022) and the image-condition way PGN (Loedeman et al., 2022) in visual prompt learning.

3.1 Vision Language Models

The most popular visual language model, CLIP, utilizes 400 million image-text pairs to train large-scale multimodal models and demonstrates impressive performance on various tasks, including zero-shot image recognition. We analyze the VLMs adaptation problem by taking CLIP as an example. But note that our method can be easily applied to the VLMs.

CLIP consists of two branches: an image encoder $\phi(\cdot)$ and a text encoder $\theta(\cdot)$. The image encoder typically leverages a vision transformer (ViT) as the backbone, which is used to convert an image $\mathbf{x} \in \mathbb{R}^{3\times H\times W}$ into a d-dimensional image feature $\mathbf{f} \in \mathbb{R}^{N\times d}$, where N is the number of split patches, and $d = 512$ in the ViT backbone. Meanwhile, the text encoder adopts a 12-layer transformer (Vaswani et al., 2017), which takes a sequence of word tokens as text input and generates an d-dimensional text representation $\mathbf{t} \in \mathbb{R}^{M \times d}$, where M is the number of classes. The two encoders are jointly trained using a contrastive loss function to align image and text feature spaces, which enables the model to transfer to downstream tasks in a zero-shot manner. After training, the entire parameters of the CLIP model are kept frozen to downstream tasks. Due to the gap between the pre-training target's textual description and the downstream task's discrete label, CLIP employs hand-crafted prompt templates to transform raw labels into textual descriptions. In the classification task, the prediction objective is defined as the classification of an image into one of C categories, which are represented by the set $y \in \{1, \ldots, C\}$. The common form of prompt template is "`a photo of a [CLASS]`", where the `[CLASS]` token is filled with the i-th class name such as "`dog`", "`fish`", "`bird`", etc. Then we construct M textual descriptions based on class labels. In this way, the text features $\mathbf{t}$ can be obtained by feeding the textual descriptions into the text encoder, and $\mathbf{t}_i$ is the i-th class token of text features. We let the image features $\mathbf{f}$ of an image $\mathbf{x}$ be extracted by an image encoder, and then we have the predicted probability of the i-th class:

$$P(y = i \mid \mathbf{x}) = \frac{\exp\left(\cos\left(\mathbf{f}_{\text{cls}}, \mathbf{t}_i\right)/\tau\right)}{\sum_{j=1}^{C} \exp\left(\cos\left(\mathbf{f}_{\text{cls}}, \mathbf{t}_i\right)/\tau\right)}, \tag{1}$$

where $\mathbf{f}_{\text{cls}}$ denotes the `[CLS]` token of the image feature $\mathbf{f}$, $\cos(\cdot, \cdot)$ denotes the cosine similarity, and τ is the temperature parameter of the softmax function.

3.2 Prompt Learning

Since VLMs consist of text and visual branches, existing prompt learning approaches are mainly divided into text and visual prompt learning.

Text prompt learning. CoOp (Zhou et al., 2022b) for the first time shows that replacing the hand-crafted prompts with automatic prompts yields a considerable performance gain. Specifically, CoOp introduces k learnable vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ to model the context words of the text prompts. We define $\mathbf{c}_i$ as the word embedding of the i-th class name. Then, the text prompt of the i-th class is denoted as $\mathbf{p}_i = \{\mathbf{v}_1, \ldots, \mathbf{v}_k, \mathbf{c}_i\}$. Therefore, we have the predicted probability of the i-th class:

$$P(y = i \mid \mathbf{x}) = \frac{\exp\left(\cos\left(\mathbf{f}_{\text{cls}}, \mathbf{t}_i\right)/\tau\right)}{\sum_{j=1}^{C} \exp\left(\cos\left(\mathbf{f}_{\text{cls}}, \mathbf{t}_j\right)/\tau\right)}, \tag{2}$$

where $\mathbf{t}_j$ is the text embedding from the text encoder.

CoCoOp (Zhou et al., 2022a) extends CoOp by designing an input-condition prompt that learns both text and image domain knowledge. As shown in Fig. 1a, CoCoOp introduces a lightweight Meta-Net network $\psi(\cdot)$ to dynamically generate prompts $\boldsymbol{\pi} = \psi(\phi(\mathbf{x}))$ for each image $\mathbf{x}$. Each text prompt token is now acquired by $\mathbf{v}_i(\mathbf{x}) = \mathbf{v}_i + \boldsymbol{\pi}$. The prompt of the i-th class $\mathbf{c}_i$ is defined as $\mathbf{p}_i = \{\mathbf{v}_1(\mathbf{x}), \ldots, \mathbf{v}_k(\mathbf{x}), \mathbf{c}_i\}$. Based on the updated prompts $\mathbf{p}_i$, the predicted probability of the i-th class is similar to Eq. 2.

Visual prompt learning. The text prompt learning only adapts to the frozen VLMs by modifying the text data space. Since both text and visual branches jointly infer the final recognition of VLMs. VP (Bahng et al., 2022) introduces a small number of learnable parameters padded around the input image to fit the frozen VLMs by modifying the image pixel space. As shown in Fig. 1c, the random-initialization method, VP adds the visual prompt parameters $\boldsymbol{w}$ to the input image $\mathbf{x}$ to form a prompted input image $\mathbf{x} + \boldsymbol{w}$. We have the predicted probability of the i-th class:

$$P(y=i \mid \mathbf{x}) = \frac{\exp(\cos(\phi(\mathbf{x}+\boldsymbol{w}), \mathbf{t}_i)/\tau)}{\sum_{j=1}^{C} \exp\left(\cos\left(\phi(\mathbf{x}+\boldsymbol{w}), \mathbf{t}_j\right)/\tau\right)}, \tag{3}$$

where $\phi(\mathbf{x} + \boldsymbol{w})$ is the image feature from the image encoder.

PGN (Loedeman et al., 2022) generates the visual prompt conditioned on the input images instead of random initialization. As shown in Fig. 1c, the image-condition method PGN introduces an extra neural network $\sigma(\cdot)$ to learn the dependency $\boldsymbol{\epsilon} = \sigma(\mathbf{x})$. In this way, the input images $\mathbf{x}$ can be directly transformed into the prompt vectors $\boldsymbol{\epsilon}$. The predicted probability of the i-th class is similar to Eq. 3.

4 Mutual Prompt Learning

Our method consists of two modules, i.e., fine-grained text prompt (FTP) and text-reorganized visual prompt (TVP), as shown in Fig. 4. In the FTP module, we design a fine-grained text prompt to align different text prompt tokens with distinct visual local features rather than the single global features. On the other hand, compared to existing visual prompt approaches, the TVP module leverages text information to construct the visual prompt to guide the image branch to attend to class-related representations. The text and image prompt modules are tightly coupled and mutually beneficial throughout the training process.

4.1 Fine-grained Text Prompt

In this section, we introduce a new fine-grained text prompt (FTP) approach to align different text prompt tokens with distinct visual local features rather than the single global features, which is parameter-free, efficient and effective compared to our previous version CTP Long et al. (2023b). In the proposed FTP, the core idea is to decouple the single global image features into several finer-grained semantic and add the

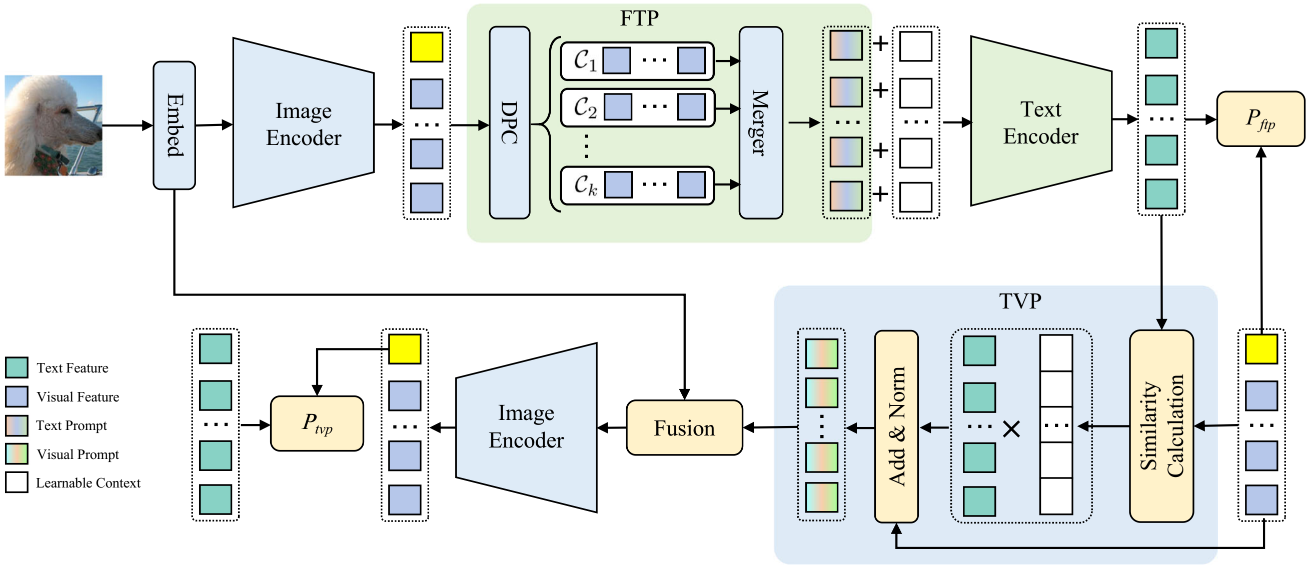

Fig. 4 Illustration of our approach. We introduce FTP and TVP to text and visual sub-branches, respectively. The FTP generates prompts based on finer-grained image information instead of using identical image semantics like CoCoOp. The TVP simultaneously explores image and text knowledge to construct the visual prompt rather than random initialization or image-condition. We leverage TVP and FTP to mutually prompt and fully unleash the potential representation capabilities of both modalities, achieving better downstream generalization performance

each merged token to different text prompt tokens. Specifically, we apply a density peak clustering algorithm to cluster tokens in image features, then merge the tokens from the same cluster into a new token by a weighted sum. In this way, we can obtain k merged tokens $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$, corresponding to k learnable tokens $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ in the text prompt. We further fuse the merged image tokens and learnable text prompt tokens to generate image-aware text prompts, which is $\{\mathbf{u}_1 + \mathbf{v}_1, \ldots, \mathbf{u}_k + \mathbf{v}_k, \mathbf{c}\}$. In the following, we first introduce the clustering algorithm, and then present how to generate image-aware text prompts based on the image token clusters.

Common clustering algorithms such as K-means and Hierarchical Clustering (Hartigan & Wong, 1979; Hastie et al., 2009) require multiple iterations and extra parameters to acquire good cluster results. However, we need an effective and efficient cluster approach for parameter-efficient fine-tuning. After extensive research, we find the density peak clustering algorithm (DPC) (Rodriguez & Laio, 2014) achieves remarkable performance without requiring an iterative process and extra parameters. The basis of the DPC algorithm is the assumption that the local density of the cluster center is higher than that of its neighbors, and the distance between it and any point with a higher density is relatively large. In order to define the local range of each point, a hard threshold cut-off distance is introduced in the DPC algorithm. Considering only a few dozen tokens for clustering in image features, the Gaussian kernel function is more effective for small-scale clustering (Du et al., 2016; Chen et al., 2020a). Therefore, we simplify the density as the inverse measure of the distance. For the image features $\mathbf{f}$, it contains N tokens in total. Let the i-th token be $\mathbf{f}_i$, and its corresponding density ρ_i is calculated as following:

$$\rho_i = \exp\left(-\sum_{\mathbf{f}_j \in \mathbf{f}} \left\|\mathbf{f}_i - \mathbf{f}_j\right\|_2^2\right), \tag{4}$$

where $\mathbf{f}_i$ and $\mathbf{f}_j$ are the i-th and j-th tokens, respectively.

Then, the minimum distance between the i-th token and any other tokens with higher density, denoted by δ_i, is defined as:

$$\delta_i = \begin{cases} \min_{j:\rho_j>\rho_i} \left\|\mathbf{f}_i - \mathbf{f}_j\right\|_2, \text{ if } \exists j \text{ s.t. } \rho_j > \rho_i \\ \max_j \left\|\mathbf{f}_i - \mathbf{f}_j\right\|_2, \text{ otherwise} \end{cases}, \tag{5}$$

where $\max_j \left\|\mathbf{f}_i - \mathbf{f}_j\right\|_2$ denotes the maximum distance between the i-th token and any other tokens. Since there is no higher density token for the highest density token, we define its minimum distance as the maximum distance between it and any other tokens.

According to the DPC algorithm, only points with relatively high ρ_i and δ_i are considered as cluster centers. To this end, we denote the cluster center score η_i of the i-th token as

$$\eta_i = \rho_i \times \delta_i, \tag{6}$$

where ρ_i and δ_i are the density and distance of the i-th token, respectively.

Higher scores η mean a higher potential to be cluster centers. Since introducing k learnable tokens $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ in the text prompts, we select the k highest scoring tokens as cluster centers. Finally, we construct k clusters by assigning each remaining token to its nearest center token. For simplicity, we denote such k clusters as $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_k\}$.

Although tokens in the same cluster have similar semantic information, the semantic importance of each token is not necessarily the same. Instead of blindly averaging the tokens in the same cluster, we combine these tokens by a weighted sum. Referring to (Liang et al., 2022), we calculate the importance score of each token by

$$\mathbf{A}_{cls} = \text{softmax}\left(\frac{\mathbf{f}_{\text{cls}} \cdot \mathbf{f}^{\top}}{\sqrt{d}}\right), \tag{7}$$

where $\mathbf{f}_{\text{cls}}$ denotes the `[CLS]` token of the image feature. Since the `[CLS]` token is taken out for classification in the last layer of the encoder. It is naturally to assume that the class attention value $\mathbf{A}_{cls}$ indicates the importance score of each token. By introducing a class attention value $\mathbf{A}_{cls}$ to represent the importance score, we combine the tokens of the i-th cluster $\mathcal{C}_i$ into a new token $\mathbf{u}_i$ by

$$\mathbf{u}_i = \sum_{\mathbf{f}_j \in \mathcal{C}_i} s_j \mathbf{f}_j, \tag{8}$$

where $s_j \in \mathbf{A}_{cls}$ denotes the importance score of token $\mathbf{f}_j$, and $\mathcal{C}_i$ denotes the i-th cluster.

In this way, we can obtain k merged tokens $\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$. Note that the k merged tokens are fixedly arranged in descending order of cluster center scores. Then we add the k merged tokens to the learnable text tokens in descending order of their cluster center scores. Considering that the learnable text tokens are randomly initialized in our method without any prior information, so it is reasonable for us to add merged tokens to the text tokens in a certain order. In this way, the i-th token of text prompt is now denoted by $\mathbf{v}_i(\mathbf{x}) = \mathbf{v}_i + \mathbf{u}_i$. The prompt of the i-th class $\mathbf{c}_i$ is defined as $\mathbf{p}_i^a = \{\mathbf{v}_1(\mathbf{x}), \ldots, \mathbf{v}_k(\mathbf{x}), \mathbf{c}_i\}$ and then input it into the text encoder to generate enhanced text features $\mathbf{t}^a$. As a result, the prediction probability of the i-th class is:

$$P_{ftp}(y = i \mid \mathbf{x}) = \frac{\exp\left(\cos\left(\mathbf{f}_{\text{cls}}, \mathbf{t}^a\right)/\tau\right)}{\sum_{j=1}^{C} \exp\left(\cos\left(\mathbf{f}_{\text{cls}}, \mathbf{t}^a\right)/\tau\right)}, \tag{9}$$

where $\mathbf{t}^a$ is the enhanced text embedding from the text encoder, incorporating text and image knowledge.

4.2 Text-reorganized Visual Prompt

In this section, we simultaneously leverages image and text features to construct visual prompts that guide the image branch to attend to class-related representations. In previous version TFT Long et al. (2023b), we propose a text-guided feature tuning strategy that utilizes global text information to guide images to focus on task-relevant regions. However, not all class prompts in text features contribute positively to the final classification task. To this end, we further propose a text-reorganized visual prompt (TVP) that reorganizes the text label descriptions of the current image according to the importance scores of different categories. Specifically, as shown in Fig. 1d, we calculate the cosine similarity between each text class token and the image `[CLS]` token as the importance score. Mathematically, the similarity scores $\mathbf{A}_{sim}$ between the image `[CLS]` token and text tokens can be calculated by

$$\mathbf{A}_{sim} = \text{softmax}\left(\frac{\mathbf{f}_{\text{cls}} \cdot (\mathbf{t}^a)^{\top}}{\sqrt{d}}\right), \tag{10}$$

where $\mathbf{f}_{\text{cls}}$ denotes the `[CLS]` token of the image feature. Then, we merge different text tokens into a new text token $\mathbf{t}_{\text{cls}}^a$ according to the similarity scores $\mathbf{A}_{sim}$ by

$$\mathbf{t}_{\text{cls}}^a = \mathbf{A}_{sim} \cdot \mathbf{t}^a, \tag{11}$$

where $\mathbf{t}_{\text{cls}}^a$ denotes the reorganized token of the text feature. Finally, we add the merged text token to the image feature by

$$\boldsymbol{\Delta} = \text{Norm}\left(\mathbf{f} + \mathbf{t}_{\text{cls}}^a\right), \tag{12}$$

where $\boldsymbol{\Delta}$ is the visual prompt vectors and Norm$(\cdot)$ refers to layer normalization (Ba et al., 2016). Then we add it to d-dimensional token embeddings to obtain prompted image tokens $\mathbf{e} = \text{Embed}(\mathbf{x}) + \boldsymbol{\Delta}$. Let $\mathbf{f}^a = \phi(\mathbf{e})$ denotes the enhanced image feature from the image encoder, then the predicted probability of the i-th class is:

$$P_{tvp}(y = i \mid \mathbf{x}) = \frac{\exp\left(\cos\left(\mathbf{f}_{\text{cls}}^a, \mathbf{t}_i^a\right)/\tau\right)}{\sum_{j=1}^{C} \exp\left(\cos\left(\mathbf{f}_{\text{cls}}^a, \mathbf{t}_j^a\right)/\tau\right)}, \tag{13}$$

where $\mathbf{f}_{\text{cls}}^a$ denotes the `[CLS]` token of the image feature $\mathbf{f}^a$ and $\mathbf{t}_j^a$ is the enhanced text embedding from the text encoder.

4.3 Overall Contrastive Loss Function

The contrastive loss function is employed to further align prompted text and image features on specific downstream tasks. Text and image semantic information is transferred between the two branches by minimizing the overall contrastive loss. We merge the prediction probability of P_{ftp} (Eq. 9), and P_{tvp} (Eq. 13),

$$P_{all}(y = i \mid \mathbf{x}) = (\alpha P_{ftp} + \beta P_{tvp})/2, \tag{14}$$

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)

Method	Average over 11 datasets			ImageNet			Caltech101			OxfordPets		
	Base	New	Hos	Base	New	Hos	Base	New	Hos	Base	New	Hos
CLIP	69.34	74.22	71.70	72.43	68.14	70.22	96.84	94.00	95.40	91.17	97.26	94.12
CoOp	82.69	63.22	71.66	76.47	67.88	71.92	98.00	89.81	93.73	93.67	95.29	94.47
CoCoOp	80.47	71.69	75.83	75.98	70.43	73.10	97.96	93.81	95.84	95.20	97.69	96.43
ProDA	81.56	72.30	76.65	75.40	70.23	72.72	98.27	93.23	95.68	95.43	97.83	96.62
VarPT	80.10	74.94	77.43	76.00	**70.93**	73.37	98.00	94.93	96.44	95.67	**98.00**	**96.82**
LASP	81.42	74.17	77.62	76.23	70.40	73.20	97.80	94.25	96.00	95.43	97.70	96.55
MaPLe	82.28	75.14	78.55	76.66	70.54	73.47	97.74	94.36	96.02	95.43	97.76	96.58
CTP+TFT	83.01	75.72	79.02	**77.42**	70.44	**73.77**	98.31	94.75	96.50	**95.86**	97.55	96.70
Ours	**83.67**	**76.48**	**79.71**	76.81	70.85	73.71	**98.64**	**95.31**	**96.95**	95.48	97.65	96.55

Method	StanfordCars			Flowers102			Food101			FGVCAircraft		
	Base	New	Hos	Base	New	Hos	Base	New	Hos	Base	New	Hos
CLIP	63.37	74.89	68.65	72.08	**77.80**	74.83	90.10	91.22	90.66	27.19	**36.29**	31.09
CoOp	**78.12**	60.40	68.13	97.60	59.67	74.06	88.33	82.26	85.19	40.44	22.30	28.75
CoCoOp	70.49	73.59	72.01	94.87	71.75	81.71	90.70	91.29	90.99	33.41	23.71	27.74
ProDA	74.70	71.20	72.91	97.70	68.68	80.66	90.30	88.57	89.43	36.90	34.13	35.46
VarPT	72.93	73.23	73.07	95.70	70.40	81.12	**91.03**	92.13	**91.57**	34.40	35.00	34.69
LASP	72.73	71.74	72.23	96.20	73.93	83.61	90.70	91.36	91.02	33.03	32.30	32.66
MaPLe	72.94	74.00	73.47	95.92	72.46	82.56	90.71	92.05	91.38	37.44	35.61	36.50
CTP+TFT	76.29	74.17	75.22	97.36	77.70	**86.43**	90.48	91.89	91.18	39.49	35.37	37.32
Ours	76.64	**74.89**	**75.75**	**98.29**	75.89	85.65	90.54	**92.31**	91.42	**41.48**	34.37	**37.59**

Method	SUN397			DTD			EuroSAT			UCF101		
	Base	New	Hos	Base	New	Hos	Base	New	Hos	Base	New	Hos
CLIP	69.36	75.35	72.23	53.24	59.90	56.37	56.48	64.05	60.03	70.53	77.50	73.85
CoOp	80.60	65.89	72.51	79.44	41.18	54.24	92.19	54.74	68.69	84.69	56.05	67.46
CoCoOp	79.74	76.86	78.27	77.01	56.00	64.85	87.49	60.04	71.21	82.33	73.45	77.64
ProDA	78.67	76.93	77.79	80.67	56.48	66.44	83.90	66.00	73.88	85.23	71.97	78.04
VarPT	79.17	77.87	78.51	75.30	60.80	67.27	80.30	75.30	77.71	82.53	75.77	79.00
LASP	80.33	77.93	79.12	79.57	59.47	68.06	90.26	69.23	78.46	83.43	77.60	80.40
MaPLe	80.82	**78.70**	79.75	80.36	59.18	68.16	**94.07**	73.23	82.35	83.00	78.66	80.77
CTP+TFT	**82.16**	77.49	**79.76**	79.47	**61.53**	69.36	92.14	73.87	82.00	84.12	77.74	80.80
Ours	81.66	77.47	79.51	**81.37**	61.11	**69.80**	93.48	**82.08**	**87.41**	**85.99**	**79.72**	**82.74**

where α and β are the balance hyper-parameters, which are analyzed in our experiments. We let the two different modalities be tightly coupled and mutual beneficial across the whole training process by performing the contrastive optimization.

5 Experiments

We evaluate the performance of our MPL on four tasks, including (1) base-to-new classes generalization (Sect. 5.2); (2) few-shot classification (Sect. 5.3); (3) cross-dataset transfer (Sect. 5.4); (4) domain generalization (Sect. 5.5). Additionally, we provide extensive ablation studies and further analysis (Sect. 5.6).

5.1 Setup

Datasets. To evaluate the effectiveness of our approach, we conduct experiments on 11 image recognition datasets for the first three tasks, namely base-to-new classes generalization, few-shot classification and cross-dataset transfer. Following (Radford et al., 2021; Zhou et al., 2022b), these include generic image classification datasets (ImageNet by (Deng et al., 2009) and Caltech101 by (Fei-Fei et al., 2004)), fine-grained classification datasets (Oxford Pets by (Parkhi et al.,

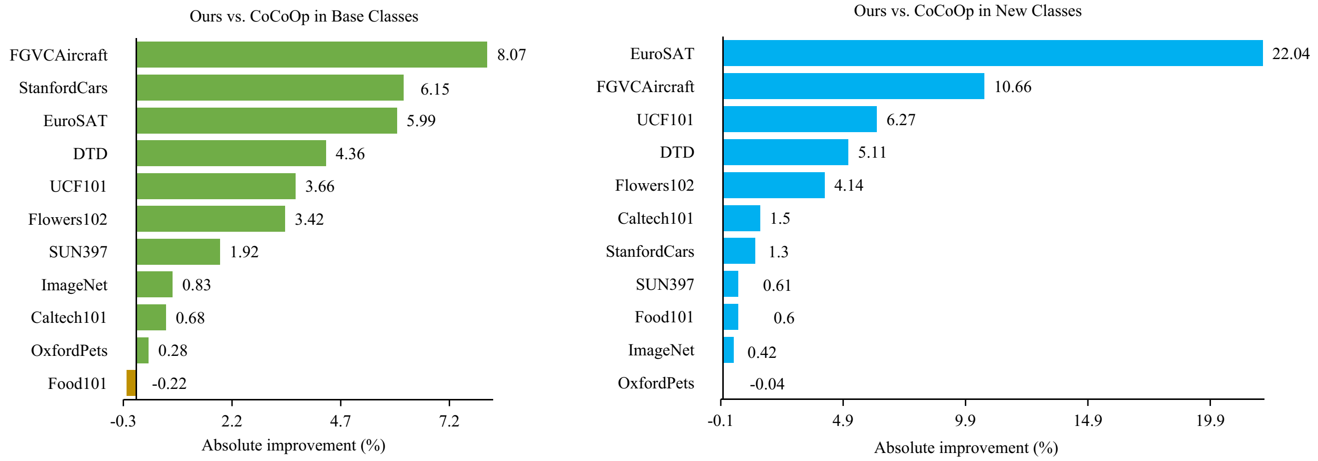

Fig. 5 Absolute improvement over CoCoOp in the base-to-new generalization task. Compared to CoCoOp, our method achieves improvement on both base (left sub-figure) and new (right sub-figure) classes on most of the datasets

2012), StanfordCars by (Krause et al., 2013), Flowers102 by (Nilsback & Zisserman, 2008), Food101 by (Bossard et al., 2014) and FGVCAircraft by (Maji et al., 2013)), scene recognition (SUN397 by (Xiao et al., 2010)), action recognition (UCF101 by (Soomro et al., 2012)), texture classification (DTD by (Cimpoi et al., 2014)), and satellite imagery recognition (EuroSAT by (Helber et al., 2019)). For the domain generalization task, we choose ImageNet as the source domain dataset and report our evaluation result on four target domain datasets which are the ImageNet variants, namely ImageNetV2 (Recht et al., 2019), ImageNet-Sketch (Wang et al., 2019), ImageNet-A (Hendrycks et al., 2021b), and ImageNet-R (Hendrycks et al., 2021a).

Training details. By following (Zhou et al., 2022a, b), we leverage a few-shot training strategy in all experiments, randomly sampling 16 samples from each class. We leverage ViT-B/16 as the visual encoder and a 12-layer transformer as the text encoder throughout the experiments. For the text prompt, we fix the context length to 6 and randomly initialize the prompt vector by drawing from a zero-mean Gaussian distribution with a standard deviation equal to 0.02. For each task and dataset, we train 10 epochs using the SGD optimizer with a base learning rate of 0.002 and a cosine decay schedule. We set the hyper-parameters α and β in Equation (14) to 1 for all experiments and provide sensitivity analyses in Fig. 10. We ran all the experiments three times with different random seeds and reported the average classification accuracy. The implementation code will be released.

Baselines. We compare our method with text prompt, visual prompt, universal prompt, and adapter fine-tuning approaches. For text prompt learning, we compare MPL with CoOp (Zhou et al., 2022b), CoCoOp (Zhou et al., 2022a), ProDA (Lu et al., 2022), VarPT (Derakhshani et al., 2022), LASP (Bulat & Tzimiropoulos, 2022), and ProDA (Lu et al., 2022). ProDA (Lu et al., 2022) proposes prompt distribution learning, which not only learns from presented few samples but also captures the distribution of diverse prompts. VarPT (Derakhshani et al., 2022) propose probabilistic modelling of the underlying prompt distribution, providing better generalization capabilities for downstream tasks. LASP (Bulat & Tzimiropoulos, 2022) adds a cross-entropy loss to minimize the distance between the learned and hand-crafted prompts. For visual prompt learning, we compare MPL with VPT (Derakhshani et al., 2022), VP (Bahng et al., 2022) and EVP (Wu et al., 2022).

For universal prompt learning, we compare MPL with UPT (Zang et al., 2022), MaPLe (Khattak et al., 2023) and CTP+TFT (Long et al., 2023b). UPT (Zang et al., 2022) proposes a self-attention network to generate both textual and visual prompts, which can preserve the benefits of a single modality. MaPLe (Khattak et al., 2023) proposes multimodal prompt learning to improve alignment between the vision and language representations. CTP+TFT (Long et al., 2023b) proposes class-aware text prompt and text-guided feature tuning to realize task-oriented multi-modal mutual learning. For adapter fine-tuning, we compare MPL with CLIP-Adapter (Gao et al., 2021), Tip-Adapter-F (Zhang et al., 2021) and TaskRes (Yu et al., 2023). CLIP-Adapter (Gao et al., 2021) adopts an additional bottleneck layer on the vision or language branch to learn new features and perform residual style feature blending with the original pre-trained features. Tip-Adapter-F (Zhang et al., 2021) builds a key-value cache model from a few samples, which can obtain adapter weights with good performance without any training. TaskRes (Yu et al., 2023) decouples the prior knowledge of the pre-trained model and the new knowledge about the target task and directly performs on the text classifier. In addition, we compare MPL with zero-shot CLIP, which leverages hand-crafted prompts designed specifically for each dataset. In order to fully demonstrate our performance, we compare

Table 2 Results (%) of **16-shot learning task** on 11 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
CLIP-Adapter	63.59	92.49	87.84	74.01	93.90	78.25	32.10	69.55	65.96	84.43	76.76	74.44
Tip-Adapter-F	65.44	92.63	88.18	75.75	94.23	78.11	35.86	71.00	66.94	84.94	79.03	75.65
TaskRes	65.73	93.43	87.83	76.83	96.03	77.60	36.30	70.67	67.13	84.03	77.97	75.78
PLOT	66.17	93.18	87.54	76.00	96.10	78.36	36.21	71.64	67.79	**85.75**	79.51	76.20
CTP+TFT	72.90	95.90	93.96	79.10	96.73	**89.95**	38.72	**79.37**	**72.49**	81.00	83.45	80.32
Ours	**72.91**	**96.12**	**93.97**	**80.93**	**97.25**	87.92	**40.47**	75.89	71.68	82.91	**83.86**	**80.35**

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

	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
VarPT	70.70	93.67	90.63	65.00	70.90	86.30	**24.93**	67.47	46.10	45.87	68.67	65.95
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
CTP+TFT	72.90	95.73	90.22	65.14	69.89	86.38	23.32	66.49	46.47	47.24	67.43	66.47
Ours	**72.91**	**95.89**	**90.97**	**66.86**	71.28	**87.13**	23.60	**67.51**	46.37	**48.34**	67.72	**67.14**

experimental results on MPL not only with the approved papers but also with the latest papers.

5.2 Generalization from Base to New Classes

To verify the generalization ability from base to new classes, we split the classes of each dataset equally into two groups: base classes (Base) and new classes (New). The learnable parameters introduced by all methods are trained only on the base classes, and the accuracy is evaluated separately on the base classes and the new classes. We leverage the harmonic mean to evaluate the average accuracy of base classes and new classes. Table 1 shows the comparison of our MPL approach with recent prompt learning works on 11 benchmarks. We observe that the proposed MPL obtains the best average performance in terms of all metrics. Compared with the classical method CoCoOp, MPL improves the accuracy in base classes from 80.47% to 83.67% by adding fine-grained image semantics to the single text prompt token. Benefiting from the mutual prompt of our FTP and TVP modules, MPL enhances the generalization performance on new classes and achieves an average gain from 71.69% to 76.48% on 11 datasets. When taking into account both the base and novel classes, MPL shows an absolute average gain of 3.88% over CoCoOp.

We provide a detailed comparison of CoCoOp and our method of per-dataset improvement in Fig. 5. Our approach 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 EuroSAT and FGVCAircraft datasets. Table 6 presents the extra parameters and computation cost of different prompt learning approaches. Compared with the latest method MaPLe, MPL achieves higher performance but introduces fewer GFLOPs and parameters. It is because we only use a single class token as input in the attention mechanism, while MaPLe increases the number of tokens in each layer of the original model. Last, we achieve new state-of-the-art results on the base-to-new classes generalization task compared to the conference version CTP+TFT (Long et al., 2023b).

5.3 Few-Shot Classification

Although CoCoOp solves CoOp's poor generalization ability on new classes, its average performance on base classes drops from 82.69% to 80.47%. We further conduct the few-shot classification experiment and report results in Table 2. It is obvious that CoCoOp decreases CoOp by 2.49% on few-shot classification. The above experiments indicate that although CoCoOp enhances the inter-class generalization, it sacrifices the intra-class discrimination ability. In comparison, our method surpasses baseline methods on all datasets on few-shot classification. Significantly our method outperforms CoCoOp by 10.13%, 9.91%, and 9.72% on EuroSAT, Flowers102, and StanfordCars, respectively, and the average improvement over 11 datasets is 5.56%. Our method also achieves about 2% improvement on the challenging dataset of ImageNet. Therefore, our method MPL not only enhances the inter-class generalization, but also improves the intra-class discriminative ability.

5.4 Cross-Dataset Transfer

Our method has been shown to have excellent generalization ability in a single dataset. We further evaluate the transferability of our method on more challenging cross-dataset tasks. In this setting, we train multi-modal prompts for 1000 classes on ImageNet. The effectiveness of the learned prompts is then tested on 10 datasets covering general and fine-grained image classification, scene recognition, and texture classification. As shown in Table 3, compared to CoCoOp, MPL has an improvement on 8 out of 10 target domains and achieves the best average accuracy on the 11 datasets. Furthermore, MPL convincingly performs over the recent state-of-the-art MaPLe by 0.84% on average accuracy. This suggests that the mutual prompt of FTP and TVP modules facilitates better generalization for cross-dataset transfer. Finally, we further improve the performance of the conference version CTP+TFT (Long et al., 2023b) on the cross-dataset transfer task.

5.5 Domain Generalization

Domain generalization trains a model on source data and evaluates its generalization ability on a target domain that is different but related to the source domain. To this end, we train the model on the few-shot ImageNet data and test the model on four ImageNet variants, 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). The experiment results are summarized in Table 4. We can clearly see that our MPL performs best on all target datasets. The promising performance shows that our MPL can not only improve the discriminative ability on the source domain data but also enhance the generalization of the target domain. In contrast, although CoCoOp improves the generalization of the target domain, it weakens the discriminative ability of the source domain data. It verifies that our MPL is more domain-generalizable. Last, we achieve new state-of-the-art results on the domain generalization task compared to the conference version CTP+TFT (Long et al., 2023b).

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

	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
UPT	72.63	64.35	48.66	50.66	76.24
CTP+TFT	72.90	64.57	49.11	50.94	76.68
Ours	**72.91**	**64.76**	**49.38**	**51.21**	**76.92**

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

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	81.58	73.90	77.29	76.27	69.73	72.85	97.94	94.75	96.32	**95.55**	97.15	96.34
C	83.36	74.59	78.50	76.43	70.84	73.53	**98.85**	95.18	**96.98**	95.14	95.62	95.38
Ours	**83.67**	**76.48**	**79.71**	**76.81**	**70.85**	**73.71**	98.64	**95.31**	96.95	95.48	**97.65**	**96.55**

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	72.34	72.89	72.61	94.62	70.90	81.06	90.36	91.61	90.98	35.15	30.94	32.91
C	77.30	72.36	74.75	98.14	**78.97**	**87.52**	90.11	91.86	90.98	39.75	31.38	35.07
Ours	76.64	**74.89**	**75.75**	**98.29**	75.89	85.65	**90.48**	**91.89**	**91.18**	**41.48**	**34.37**	**37.59**

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	80.56	74.41	77.36	80.63	56.37	66.35	90.17	77.68	83.46	83.84	76.46	79.98
C	81.48	**77.85**	**79.62**	81.27	58.69	68.16	93.32	73.28	82.09	85.16	74.50	79.47
Ours	**81.66**	77.47	79.51	**81.37**	**61.11**	**69.80**	**93.48**	**82.08**	**87.41**	**85.99**	**79.72**	**82.74**

Table 6 Comparisons of MPL and previous methods in average accuracy, extra parameters (Params), and computation cost (GFLOPs) on the base-to-new generalization task

	Base	New	Hos	Params ↑ (%)	GFLOPs ↑ (%)
CLIP	69.34	74.22	71.70	-	-
CoOp	82.69	63.22	71.66	**0.002**	**0.00**
CoCoOp	80.47	71.69	75.83	0.03	0.00
MaPLe	82.28	75.14	78.55	2.85	1.46
CTP+TFT	83.01	75.72	79.02	2.43	0.15
MPL	**83.67**	**76.48**	**79.71**	2.55	1.35

5.6 Ablation Analysis

Effectiveness of each module. Our framework comprises two key components: a fine-grained text prompt (FTP) and a text-reorganized vision prompt (TVP). In order to evaluate the effectiveness of each module, we conduct comprehensive ablation studies without FTP and TVP on 11 datasets, respectively. As shown in Table 5, both FTP and TVP modules can significantly improve the accuracy compared to the vanilla prompt learning (CoOp), and the model performs better when they work together. Specifically, FTP and TVP improve the average results by 5.63% and 6.84%, respectively, and the combination of them improves the average results by 8.05%. It indicates that FTP and TVP components are beneficial to learning accurate text and visual prompts. In addition, the

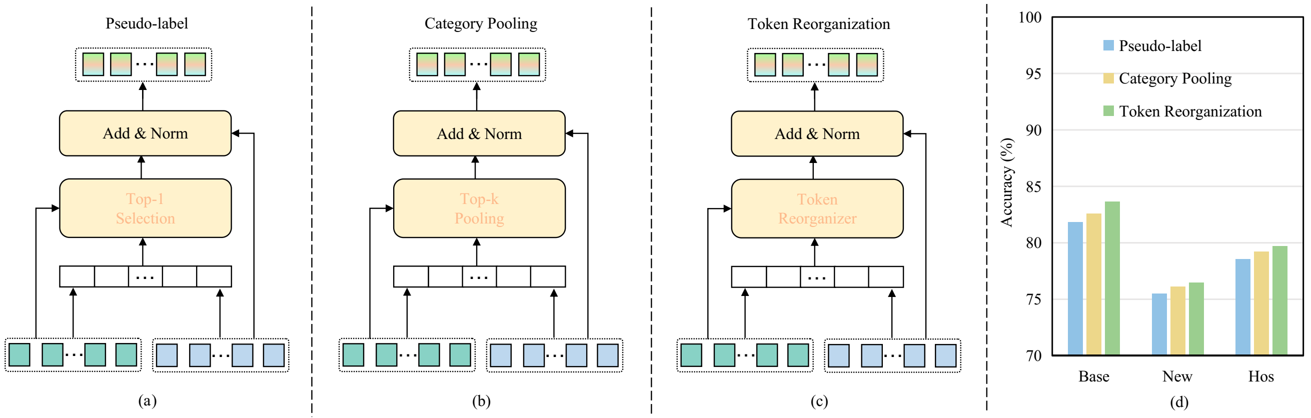

Fig. 6 Different text-reorganized strategies. **a** The pseudo-label method directly selects the category with the top-1 score as the pseudo-label of the current sample according to the importance scores of different categories. **b** The category pooling method selects the k categories with the highest scores according to the importance scores and averages pools the k categories as the supervision information of the current sample. **c** The token reorganization method assigns weights to tokens of different categories according to the importance scores of different categories and merges them into new tokens by weighted summation. **d** The average results of different text-reorganized strategies on the base-to-new generalization task on 11 benchmark datasets

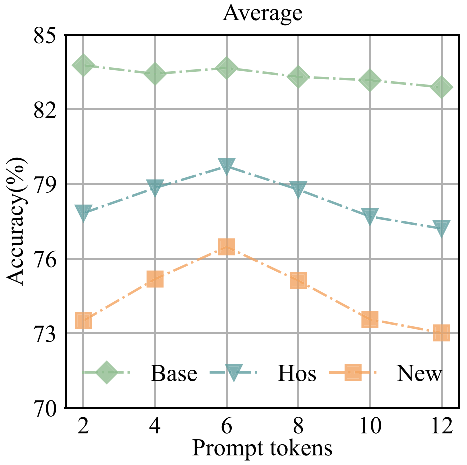

Fig. 7 Ablation studies on the length of prompt on the average accuracy of the 11 datasets

mutual prompt of both modules further improves the performance of downstream tasks.

Computational complexity analysis. Since MPL introduces more computational complexity compared to CoOp and CoCoOp, we report the accuracy, extra parameters and computation cost in the Table 6. Since Time/GPU memory is sensitive to different datasets, we utilize the computational cost (in GFLOPs) to examine the efficiency. Apparently, compared with the existing SOTA method MaPLe, MPL achieves an average accuracy improvement of 1.16% on harmonic mean over 11 classification datasets and requires lower computational cost. Unlike recent SOTA MaPLe, which increases the number of tokens in each layer of the foundation model, we only adopt a single class token as input in the attention. Compared to our baseline CoOp, our method demonstrates a modest increase of only a few percent in the number of additional parameters and computation costs. However, it remarkably leads to a substantial improvement in generalization performance by 12.5% when evaluated on unseen classes.

Different text-reorganized strategies. The TVP module constructs the visual prompt by directly focusing on textual domain knowledge. A considerable challenge is the redundant and negative category semantics contained in the textual information. The straightforward idea is to take only the category with the top-1 score as the pseudo-label of the current sample according to the importance scores. However, low experimental accuracy on most datasets leads to poor-quality pseudo-labels. We select the k categories with the highest scores according to the importance scores and average pool the k categories as the supervision information of the current sample. However, the hard approach is very sensitive to the artificially preset K value. Therefore, we design a soft token reorganization strategy that assigns weights to tokens of different categories according to the importance scores and merges them into new tokens by weighted summation. We report the experimental result for three text-reorganized strategies in Fig. 6. With the continuous improvement of the text reorganization strategy, our accuracy on the three experimental indicators of Base, New and Hos is also constantly increased. It indicates our text-reorganized strategy is beneficial to enhance both generalization and discriminative ability.

Text prompt context length. Fig. 7 shows the influence of the context length on the average accuracy of the 11 image recognition datasets. For fair comparison, we adopt random initialization for the context tokens of different lengths. The

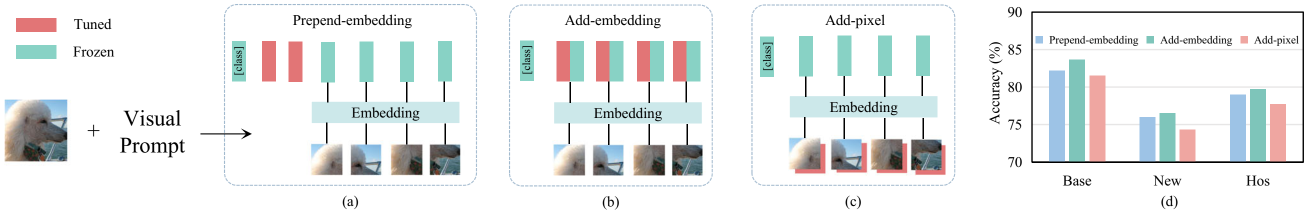

Fig. 8 Ablation studies on the location choice (**a**, **b**, **c**) of visual prompt and the corresponding results (**d**)

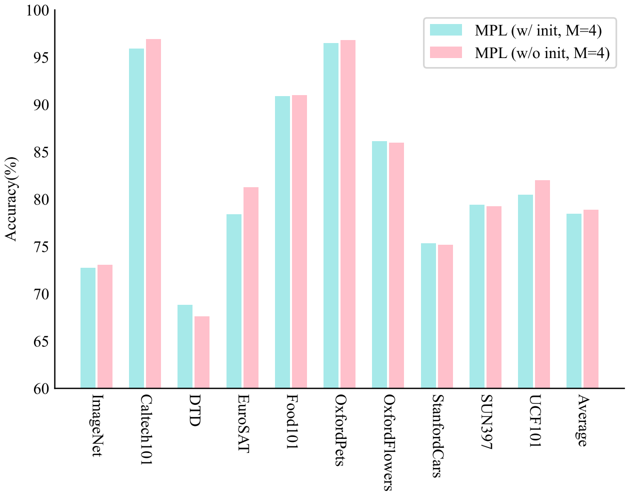

Fig. 9 Effect of different initialization of text prompts

differences in the base classes are fairly small, whereas in the new classes, the models with a longer context length clearly perform better, inconsistent with the findings in CoCoOp (Zhou et al., 2022a). From Fig. 7, we observe that using 6 randomly initialized context tokens is marginally better than using other properly initialized context tokens. Furthermore, we observe that excessive context lengths hurt performance, which may be attributed to overfitting due to more parameters being learned. All in all, choosing an appropriate context length requires a balance between discriminative performance and generalization ability.

Visual prompt location. Following VPT (Derakhshani et al., 2022), we ablate the effect of three different insertion locations on the final performance in Fig. 8. i) Prepend-embedding: prepend prompts to the sequence of the image patch embeddings. ii) Add-embedding: add prompts element-wise to image patch embeddings, keeping the transformer's input sequence length constant. iii) Add-pixel: add prompts element-wise to image patches in the pixel level instead of inserting the prompts as latent vectors. We observe that the add-embedding location generally outperforms the other visual prompt locations. In particular, the add-pixel location significantly reduces the experimental accuracy. This suggests that prompts are more accessible to learn task-relevant signals in the latent input space rather than the pixel space.

Text prompt initialization. The zero-shot performance of CLIP (Radford et al., 2021) is sensitive to hand-crafted prompts. We want to know whether artificially initializing prompts have a large influence on the performance of learnable prompt tuning. Therefore, we conduct extensive ablation studies on different datasets to compare a common word vector initialization “a photo of a [CLASS]” with the random initialization. Figure 9 reports the base-to-new generalization results on 10 datasets. We observe that the random initialization outperforms the word vector initialization on average accuracy. However, each dataset has a different result, and the word vector initialization of prompts also provides competitive performance. It indicates that learnable prompt tuning has learned satisfying text prompts with few-shot domain knowledge.

Comparison of different structure designs. To further provide an in-depth analysis of our mutual prompt learning, we further compare two vanilla structures and two task-oriented strategies: (1) MLP-PL: the image features are forwarded to a block of Linear-ReLU-Linear, borrowed from (Zhou et al., 2022a), and then added to the text for augmenting it. (2) MLP-FT: the text prompts are forwarded to the Linear-ReLU-Linear block and then added to the image for augmenting it. (3) CTP: (Long et al., 2023b) proposes class-aware text prompts which generate text prompts based on task-relevant image semantics to avoid semantic ambiguity brought by existing approaches. (4) TFT: (Long et al., 2023b) proposes text-guided feature tuning, which leverages text information to guide the image branch to pay more attention to the task-related representations. We report the results of the different designs in Table 7. In text prompt learning, we observe that the CTP and FTP modules achieve significant performance gains compared to the MLP-PL block. It demonstrates that our design of fine-grained text prompt effectively aligns different text prompt tokens with distinct visual local features rather than the single global features of the MLP-PL block. Noteworthy, compared with CTP, FTP not only improves performance but also is more efficient and parameter-free. In visual prompt learning, it is evident that since the TFT and TVP modules utilize text information to construct the visual prompt, the accuracy increases by 1.30% and 2.56% over MPL-FT, respectively. Furthermore, compared with TFT, TVP further improves performance by

Table 7 Comparison of different structures for text and visual prompt learning

Text prompt learning			Visual prompt learning			Accuracy (%)
MPL-PL	CTP	FTP	MPL-FT	TFT	TVP	
						71.66 (CoOp)
✓						75.83 (+4.17)
	✓					76.64 (+4.98)
		✓				77.29 (+5.63)
			✓			75.94 (+4.28)
				✓		77.24 (+5.58)
					✓	78.50 (+6.84)
✓		✓				77.05 (+5.39)
	✓		✓			79.02 (+7.36)
		✓			✓	**79.71** (+8.05)

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 text and visual prompt modules, MLP-PL and MLP-FT are designed using the Linear-ReLU-Linear block setting of CoCoOp (Zhou et al., 2022a), CTP and TFT are designed using class-aware text prompt and text-guided feature tuning modules of CTP+TFT (Long et al., 2023b)

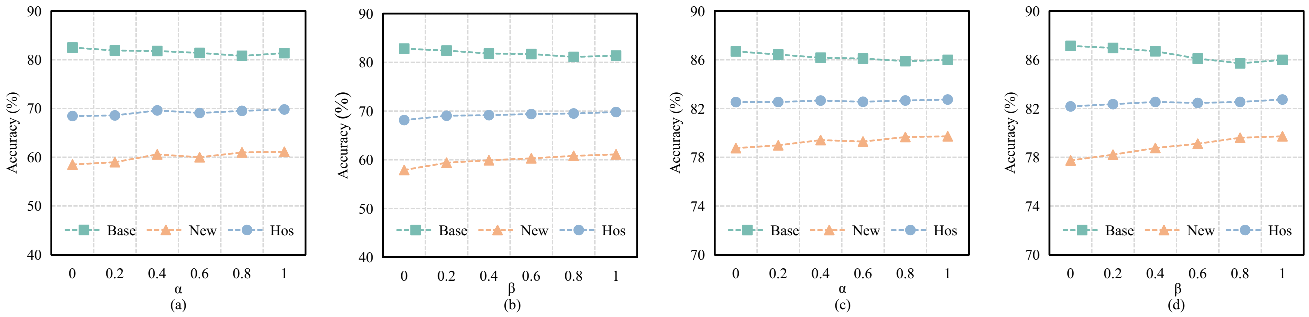

Fig. 10 **Sensitivity analysis of α and** β, with base, new, and hos metrics, on UCF101 (**a**, **b**) and Caltech101 (**c**, **d**) datasets

Table 8 The nearest word for each of the 16 learnable text tokens learned by MPL

#	OxfordPets	StanfodCars	Flowers102	Food101	FGVCAircraft
1	Flyeagles	Shut	Stone	Meat	Purely
2	Coat	Automatic	Resolution	N/A	Sarcastic
3	Insta	Door	Tane	Coles	Randomly
4	Weather	Ka	Homegrown	Wh	Maybe
5	Nir	Main	Surrounds	Slices	Specifically
6	Rag	Both	Daniels	Homemade	N/A
7	Haz	Multiple	Sights	Spring	Exactly
8	Bur	Sheikh	Frameworks	Flat	Towards
9	Tur	Batteries	Burne	Gred	Transferred
10	Coscino	dca	N/A	Spag	React
11	Physis	Unique	Segments	Grit	Vie
12	Marsh	Large	Yellow	Tor	Fuse
13	Brindle	N/A	Quarter	Dant	N/A
14	Nuke	Reality	Ton	Vised	Turb
15	N/A	Du	Ding	N/A	Pre
16	Ated	Power	Leaves	Cleans	Promos

N/A means non-Latin characters

removing the redundant and negative category semantics contained in the textual information. In addition, we find that combining MLP-PL & MLP-FT, CTP & TFT, and FTP & TVP can both improve the results compared with using either of them. It indicates that both text prompt learning and visual prompt learning are essential to achieve better results.

Sensitivity analysis. In Fig. 10, we investigate the sensitivity of two hyper-parameters: α and β of Equation (14). α and β are adopted to weigh the importance of TVP and FTP modules, respectively. When α and β equals 0.0, the model is equivalent to vanilla prompt learning (CoOp). As α and β increase, the prompted model is encouraged to pay more attention to the general knowledge. As a result, the generalization performance (New) will increase while the discrimination accuracy (Base) drops and the overall performance (Hos) gradually increases. In addition, the best performance is achieved when α and β equals 1.0.

Interpreting of text prompts. Learnable prompts are difficult for humans to understand since they transform the context vector distribution from discrete to continuous. Following CoOp, we visualize the text prompts searching the words closest to the learned text tokens in the embedding space. As shown in Table 8, we observed that the nearest words for a few learnable tokens are somewhat related to the corresponding dataset, such as "flyeagles" from OxfordPets, "yellow" from Flowers102, and "slices" from Food101. It demonstrates that each learnable token may focus on one or a subset of characteristics of the corresponding dataset.

6 Conclusion

In this paper, we propose a mutual prompt learning (MPL) approach consisting of a fine-grained text prompt (FTP) and a text-reorganized vision prompt (TVP) to re-activate the task-related representations abilities of VLMs. The FTP decomposes the single global image features into several finer-grained semantics to fuse text and image tokens at the same granularity. On the other hand, the TVP reorganizes the text descriptions of the current image and enables a more precise construction of the visual prompt by synergistically leveraging both image and text knowledge. In addition, we exploit TVP and FTP to mutually prompt and fully unleash the potential representation capabilities of both modalities. As a result, our method achieves excellent performance on 11 classification benchmarks and outperforms other prompt tuning approaches by a large margin. We hope that MPL can be a strong baseline for VLMs adaptation.

Considering that MPL further increases the training costs, a direct and effective approach is to reduce the computational complexity of the model. Inspired by efficient vision transformer methods (Liang et al., 2022; Xu et al., 2022; Meng et al., 2022; Long et al., 2023a), we can design a token pruning strategy for the visual encoder. Specifically, we can decouple the attentive and inattentive tokens based on the class token attention. In addition to preserving the most discriminative local tokens, we merge similar inattentive tokens and match homogeneous attentive tokens to maintain the model's generalization ability. This will improve training efficiency while maintaining model performance.

Limitations. Similar to CoCoOp (Zhou et al., 2022a), MPL learns image-conditioned text prompts, which may slow down the training speed. This happens because image-conditioned text prompts require an independent forward pass of instance-specific prompts through the text encoder for each image, rather than just a single forward pass of prompts through the text encoder for any size batch. We will to solve this efficiency issue in the future work.

Acknowledgements This work is supported by the National Natural Science Foundation of China (62376106), the National Key Research and Development Program of China (No. 2020YFA0714103), the Fundamental Research Funds for the Central Universities (JLU). Zichang Tan is the project lead.

Data availability The datasets used in this study are publicly available online.

References

Ba, J.L., Kiros, J.R., & Hinton, G.E. (2016). Layer normalization. arXiv preprint arXiv:1607.06450.

Bahng, H., Jahanian, A., Sankaranarayanan, S., et al. (2022). Exploring visual prompts for adapting large-scale models. arXiv preprint arXiv:2203.17274 1(3):4.

Bossard, L., Guillaumin, M., & Van Gool, L. (2014). 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, Springer, (pp. 446–461).

Brown, T., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

Bulat, A., & Tzimiropoulos, G. (2022). Language-aware soft prompting for vision & language foundation models. arXiv preprint arXiv:2210.01115.

Carion, N., Massa, F., Synnaeve, G., et al. (2020). End-to-end object detection with transformers. In *European conference on computer vision*, Springer, (pp. 213–229).

Chen, G., Yao, W., Song, X., et al. (2023). PLOT: Prompt learning with optimal transport for vision-language models. In *The Eleventh International Conference on Learning Representations*, https://openreview.net/forum?id=zqwryBoXYnh.

Chen, Y., Hu, X., Fan, W., et al. (2020). Fast density peak clustering for large scale data based on knn. *Knowledge-Based Systems, 187*(104), 824.

Chen, Y.C., Li, L., Yu, L., et al. (2020b). Uniter: Universal image-text representation learning. In *Computer Vision–ECCV 2020: 16th European Conference*, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX, Springer, (pp. 104–120).

Cimpoi, M., Maji, S., Kokkinos, I., et al. (2014). Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 3606–3613).

Dai, Z., Cai, B., Lin, Y., et al. (2021). Up-detr: Unsupervised pre-training for object detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, (pp. 1601–1610).

Davison, J., Feldman, J., & Rush, A.M. (2019). 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)*, (pp. 1173–1178).

Deng, J., Dong, W., Socher, R., et al. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE*, (pp. 248–255).

Derakhshani, M.M., Sanchez, E., Bulat, A., et al. (2022). Variational prompt tuning improves generalization of vision-language models. arXiv preprint arXiv:2210.02390.

Devlin, J., Chang, M.W., Lee, K., et al. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dong, X., Zheng, Y., Bao, J., et al. (2022). Maskclip: Masked self-distillation advances contrastive language-image pretraining. arXiv preprint arXiv:2208.12262.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Dou, Z.Y., Xu, Y., Gan, Z., et al. (2022). An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 18,166–18,176).

Du, M., Ding, S., & Jia, H. (2016). Study on density peaks clustering based on k-nearest neighbors and principal component analysis. *Knowledge-Based Systems, 99*, 135–145.

Fei-Fei, L., Fergus, R., & Perona, P. (2004). 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, IEEE*, (pp. 178–178).

Gao, P., Geng, S., Zhang, R., et al. (2021). Clip-adapter: Better vision-language models with feature adapters. arXiv preprint arXiv:2110.04544.

Gao, T., Fisch, A., & Chen, D. (2020). Making pre-trained language models better few-shot learners. arXiv preprint arXiv:2012.15723.

Graham, B., El-Nouby, A., Touvron, H., et al. (2021). Levit: a vision transformer in convnet's clothing for faster inference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (pp. 12,259–12,269).

Hartigan, J. A., & Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society Series C (Applied Statistics), 28*(1), 100–108.

Hastie, T., Tibshirani, R., Friedman, J. H., et al. (2009). *The elements of statistical learning: data mining, inference, and prediction,* (Vol. 2). Springer.

Haviv, A., Berant, J., & Globerson, A. (2021). Bertese: Learning to speak to bert. arXiv preprint arXiv:2103.05327.

He, K., Zhang, X., Ren, S., et al. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 770–778).

Helber, P., Bischke, B., Dengel, A., et al. (2019). 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.

Hendrycks, D., Basart, S., Mu, N., et al. (2021a). 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).

Hendrycks, D., Zhao, K., Basart, S., et al. (2021b). Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 15,262–15,271).

Houlsby, N., Giurgiu, A., Jastrzebski, S., et al. (2019). Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning, PMLR*, (pp. 2790–2799).

Jia, C., Yang, Y., Xia, Y., et al. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning, PMLR*, (pp. 4904–4916).

Jia, M., Tang, L., Chen, B.C., et al. (2022). Visual prompt tuning. In *Computer Vision–ECCV 2022: 17th European Conference*, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII, Springer, (pp. 709–727).

Jiang, Z., Xu, F. F., Araki, J., et al. (2020). How can we know what language models know? *Transactions of the Association for Computational Linguistics, 8*, 423–438.

Khattak, M.U., Rasheed, H., Maaz, M., et al. (2023). Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 19,113–19,122).

Kim, W., Son, B., & Kim, I. (2021). Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning, PMLR*, (pp. 5583–5594).

Krause, J., Stark, M., Deng, J., et al. (2013). 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, (pp. 554–561).

Lester, B., Al-Rfou, R., & Constant, N. (2021). The power of scale for parameter-efficient prompt tuning. arXiv preprint arXiv:2104.08691.

Li, L.H., Yatskar, M., Yin, D., et al. (2019). Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557.

Li, X.L., & Liang, P. (2021). Prefix-tuning: Optimizing continuous prompts for generation. arXiv preprint arXiv:2101.00190.

Li, Y., Fan, H., Hu, R., et al. (2022a). Scaling language-image pre-training via masking. arXiv preprint arXiv:2212.00794.

Li, Z., Wang, W., Xie, E., et al. (2022b) Panoptic segformer: Delving deeper into panoptic segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 1280–1289).

Liang, Y., Ge, C., Tong, Z., et al. (2022). Not all patches are what you need: Expediting vision transformers via token reorganizations. arXiv preprint arXiv:2202.07800.

Liu, P., Yuan, W., Fu, J., et al. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys, 55*(9), 1–35.

Liu, X., Zheng, Y., Du, Z., et al. (2021a). Gpt understands, too. arXiv preprint arXiv:2103.10385.

Liu, Z., Lin, Y., Cao, Y., et al. (2021b). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (pp. 10,012–10,022).

Liu, Z., Yu, X., Fang, Y., et al. (2023b). Graphprompt: Unifying pre-training and downstream tasks for graph neural networks. arXiv preprint arXiv:2302.08043.

Loedeman, J., Stol, M.C., Han, T., et al. (2022). Prompt generation networks for efficient adaptation of frozen vision transformers. arXiv preprint arXiv:2210.06466.

Long, S., Zhao, Z., Pi, J., et al. (2023a). Beyond attentive tokens: Incorporating token importance and diversity for efficient vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 10,334–10,343).

Long, S., Zhao, Z., Yuan, J., et al. (2023b). Task-oriented multi-modal mutual leaning for vision-language models. arXiv preprint arXiv:2303.17169.

Lu, Y., Liu, J., Zhang, Y., et al. (2022). Prompt distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 5206–5215).

Maji, S., Rahtu, E., Kannala, J., et al. (2013). Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151.

Meng, L., Li, H., Chen, B.C., et al. (2022). Adavit: Adaptive vision transformers for efficient image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 12,309–12,318).

Mu, N., Kirillov, A., Wagner, D., et al. (2022). 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, Springer, (pp. 529–544).

Nilsback, M. E., & Zisserman, A. (2008). Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision* (pp. 722–729). IEEE: Graphics & Image Processing.

Parkhi, O.M., Vedaldi, A., Zisserman, A., et al. (2012). Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE*, (pp. 3498–3505).

Petroni, F., Rocktäschel, T., Lewis, P., et al. (2019). Language models as knowledge bases? arXiv preprint arXiv:1909.01066.

Radford, A., Kim, J.W., Hallacy, C., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning, PMLR*, (pp. 8748–8763).

Raffel, C., Shazeer, N., Roberts, A., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research, 21*(1), 5485–5551.

Rao, Y., Zhao, W., Chen, G., et al. (2022). Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 18,082–18,091).

Recht, B., Roelofs, R., Schmidt, L., et al. (2019). Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, PMLR, (pp. 5389–5400).

Rodriguez, A., & Laio, A. (2014). Clustering by fast search and find of density peaks. *science, 344*(6191), 1492–1496.

Scao, T.L., Fan, A., Akiki, C., et al. (2022). Bloom: A 176b-parameter open-access multilingual language model. arXiv preprint arXiv:2211.05100.

Shin, T., Razeghi, Y., Logan IV. R.L., et al. (2020). Autoprompt: Eliciting knowledge from language models with automatically generated prompts. arXiv preprint arXiv:2010.15980.

Soomro, K., Zamir, A. R., & Shah, M. (2012). A dataset of 101 human action classes from videos in the wild. *Center for Research in Computer Vision, 2*(11).

Su, W., Zhu, X., Cao, Y., et al. (2019). Vl-bert: Pre-training of generic visual-linguistic representations. arXiv preprint arXiv:1908.08530.

Touvron, H., Cord, M., Douze, M., et al. (2021). Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning, PMLR*, (pp. 10,347–10,357).

Touvron, H., Lavril, T., Izacard, G., et al. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Wallace, E., Feng, S., Kandpal, N., et al. (2019). Universal adversarial triggers for attacking and analyzing nlp. arXiv preprint arXiv:1908.07125.

Wang, H., Ge, S., Lipton, Z., et al. (2019). Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, *32*.

Wang, M., Xing, J., & Liu, Y. (2021a). Actionclip: A new paradigm for video action recognition. arXiv preprint arXiv:2109.08472.

Wang, W., Xie, E., Li, X., et al. (2021b). Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (pp. 568–578).

Wu, H., Xiao, B., Codella, N., et al. (2021). Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (pp. 22–31).

Wu, J., Li, X., Wei, C., et al. (2022). Unleashing the power of visual prompting at the pixel level. arXiv preprint arXiv:2212.10556.

Xiao, J., Hays, J., Ehinger, K.A., et al. (2010). Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE*, (pp. 3485–3492).

Xie, E., Wang, W., Yu, Z., et al. (2021). Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems, 34*, 12,077-12,090.

Xu, Y., Zhang, Z., Zhang, M., et al. (2022). Evo-vit: Slow-fast token evolution for dynamic vision transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, (pp. 2964–2972).

Yang, Y., Huang, W., Wei, Y., et al. (2022). Attentive mask clip. arXiv preprint arXiv:2212.08653.

Yu, T., Lu, Z., Jin, X., et al. (2023). Task residual for tuning vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 10,899–10,909).

Yuan, L., Chen, Y., Wang, T., et al. (2021a). Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (pp. 558–567).

Yuan, W., Neubig, G., & Liu, P. (2021). Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems, 34*, 27,263-27,277.

Zang, Y., Li, W., Zhou, K., et al. (2022). Unified vision and language prompt learning. arXiv preprint arXiv:2210.07225.

Zhang, R., Fang, R., Zhang, W., et al. (2021). Tip-adapter: Training-free clip-adapter for better vision-language modeling. arXiv preprint arXiv:2111.03930.

Zhang, S., Jiang, T., Wang, T., et al. (2020). Devlbert: Learning deconfounded visio-linguistic representations. In *Proceedings of the 28th ACM International Conference on Multimedia*, (pp. 4373–4382).

Zhou, K., Yang, J., Loy, C.C., et al, (2022a). Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 16,816–16,825).

Zhou, K., Yang, J., Loy, C. C., et al. (2022). Learning to prompt for vision-language models. *International Journal of Computer Vision, 130*(9), 2337–2348.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.