MAO: Efficient Model-Agnostic Optimization of Prompt Tuning for Vision-Language Models

Haoyang Li[1,2], Siyu Zhou[2], Liang Wang[1,2], Guodong Long[2*]
[1]*School of Mechanical Engineering and Automation, Shanghai University, Shanghai, China*
[2]*Australian Artificial Intelligence Institute, University of Technology Sydney, Sydney, Australia*
{haoyang.li-3, liang.wang-3}@student.uts.edu.au, siyu.zhou01@gmail.com, guodong.long@uts.edu.au

***Abstract*—Though CLIP-based prompt tuning significantly enhances pre-trained Vision-Language Models, existing research focuses on reconstructing the model architecture, e.g., additional loss calculation and meta-networks. These approaches generally lead to increased complexity and extended training cost. To maintain the efficiency of the tuning process, we propose plug-and-play Model-Agnostic Optimization (MAO) for prompt tuning. Without altering any components of the prompt tuning backbone, we introduce a Data-Driven Enhancement framework to optimize the distribution of the initial data, and incorporate an Alterable Regularization module to boost the task-specific feature processing pipeline, thereby improving overall performance while maintaining low computational cost. Extensive experiments on MAO demonstrate its outstanding performance and efficiency.**

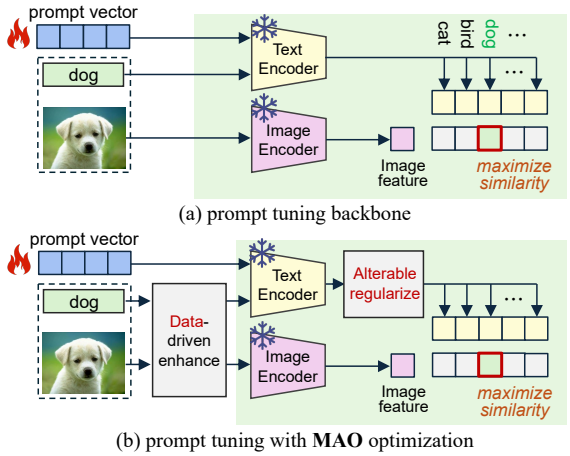
***Index Terms*—Prompt tuning, Vision-language model, Multi-modal learning**

I. Introduction

Vision-Language Models (VLMs) have revealed remarkable capabilities in cross-modal alignment and fusion [1], [2]. Represented by CLIP [3], by pre-training on hyper-scale image-text pairs, VLMs achieve robust open-domain representation and multi-modal understanding (e.g., zero-shot recognition). To further explore the potential of CLIP, Prompt Tuning is proposed as a Parameter-Efficient Fine-Tuning (PEFT) method [4]. Freezing all parameters in foundation CLIP, this approach introduces a lightweight, learnable prompt vector to supersede the original textual or visual input, guiding CLIP's output to fit the distribution of target task.

The objectives of Prompt Tuning can be summarized as: *(1)* enhancing performance on target tasks (**base classes**) through PEFT, and *(2)* maintaining generalization capacity when inferring on unknown images in other out-of-distribution categories (**new classes**). To reach these visions, numerous prompt learners are proposed, containing additional loss functions as constraints [5]–[7], extra meta-net layers for cross-modal alignment [8]–[10], and the incorporation of external knowledge [11], [12]. Unfortunately, though the performance is practically improved, these models also exhibit raised complexity and computational cost. Compared to native CoOp [4], the learnable parameters expand from tens of thousands to millions, and GPU memory usage also increases exponentially. The increase in computational demands limits the flexibility for efficient fine-tuning and elastic deployment.

*Corresponding author.



(a) prompt tuning backbone

(b) prompt tuning with **MAO** optimization

Fig. 1. Architecture comparison between (a) existing prompt tuning backbones and (b) our **M**odel-**A**gnostic **O**ptimization (MAO) framework that introduces Data-Driven Enhancement and Alterable Regularization Module.

To address this issue, herein, we propose a **M**odel-**A**gnostic **O**ptimization (MAO), an efficient plug-and-play prompt tuning method with almost no appended computational overhead. Overall, we observe that existing approaches focus primarily on ameliorating the structure of prompt learners, while ignoring the optimization of the workflow in **data** and **feature processing**. Thus, outside the fine-tuning framework of backbone models, we introduce *Data-Driven Enhancement* to improve the quality of data distribution and an *Alterable Regularization* strategy to optimize feature representation, which are devised separately for base or new tasks without additional parameters.

For **base**-class tasks, the purpose of MAO is to constrain the tuning process to further fit the distribution of base classes. As Data-Driven Enhancement, we introduce a pre-trained Hard Negative Sampler based on semantic similarity, replacing the random sampling strategy of backbones. This approach enhances the representation of data distribution of base classes by building a denser set with hard negatives. Subsequently, we integrate Alterable Regularization into the flow of feature representation, restricting the model to dynamically learn internal feature relationships of hard negatives for better fitting, while improving generalization through the introduced randomness.

In **new**-class tasks, as extant prompt tuning backbones rely on paired image-text for training, it is tough to effectually

exploit unlabeled images. Recent studies [13] explore applying knowledge distillation to learn from unimodal images. However, this type of approach assumes the existence of a larger pre-tuned teacher prompt model. Moreover, the data requirements and computational overhead are dramatically risen. As a concise and effective alternative, we introduce a rapid pseudo-labeling strategy as Data-Driven Enhancement. Resorting to the outstanding zero-shot capabilities of foundation CLIP, MAO assigns pseudo-labels inferred by CLIP Top-1 to the few-shot unlabeled images for constructing image-text pairs. Additionally, Alterable Regularization is employed to focus on the feature distribution of the pseudo-labels. Without increasing computational cost, this approach efficiently learns new-class features and enhances generalization capacity.

As a model-agnostic optimizer, our MAO can be plug-and-play adapted to most prompt tuning backbones. Extensive experiments verify that compared to the backbones, MAO achieves remarkable integral performance improvements, while maintaining almost unchanged computational cost and inference efficiency. Compared to more progressive models with similar performance, MAO demands less fine-tuning time and only about 30% of the GPU memory.

Our main contributions can be concluded as follows:

1) We propose Model-Agnostic Optimization (MAO), which efficiently optimizes prompt tuning backbones at data and feature level in a plug-and-play manner, consuming almost no additional computational resources.
2) We introduce task-related Data-Driven Enhancement to MAO, improving the data distribution of base and new classes through hard negative sampling and rapid pseudo-label allocation, respectively.
3) We incorporate Alterable Regularization into the procedure of feature processing, constraining the model to dynamically focus more on the features of updated data to enhance performance and generalization.

The code and *Supplementary Material* are available at: https://github.com/JREion/M.A.O.

II. Related Work

CLIP-based VLMs. Vision-Language Models (VLMs) have gained comprehensive attention due to the cross-modal representation capacities [3], [14]. As a representative work of VLMs, CLIP utilizes $\sim 400M$ image-text pairs (with text in the form of "*A photo of a [CLASS]* " as **hard** prompt) to train ViT-based [15] image and text encoders, achieving deep-seated alignment between visual and textual modalities through contrastive learning. The large-scale pre-training endows CLIP with prominent zero-shot multi-modal understanding ability.

In this paper, we adhere to the backbone settings to perform prompt tuning based on frozen CLIP. Additionally, aided by the zero-shot capability of CLIP, we assign pseudo-labels to few-shot unlabeled images in new-class tasks, efficiently enhancing the generalization performance of MAO.

Prompt Tuning. Due to the deep network layers and parameters, full fine-tuning on VLMs is commonly challenging. In contrast, prompt tuning is proposed as a Parameter-Efficient Fine-Tuning (PEFT) strategy, allowing CLIP to rapidly adapt to target tasks [4]. Instead of using hard prompts in the foundation CLIP, it applies a set of learnable lightweight vectors as prompts of the frozen CLIP backbone, which are continuously fine-tuned on particular tasks, acting as queries.

Taking CoOp [4] as origination, comprehensive research is conducted, aiming at reinforcing base-class performance while maintaining generalization to new classes. Proposed approaches cover introducing visual [16] or joint prompt vectors [17], appending more robust constraints (e.g., consistency loss [5], [6]), further cross-modal alignment via auxiliary meta-networks [8]–[10], or fine-tuning guided by external knowledge (e.g., Large Language Models [11]). Obviously, due to the stacking of new learnable modules, while the performance is improved, there is also an expansion in parameters and computational cost of these prompt learners. In contrast, our MAO focuses on model-agnostic optimization strategies by enhancing data and feature processing, thereby achieving performance gains with minimal additional computational cost.

III. Proposed Method

The framework of MAO is illustrated in Fig. 2. As a plug-and-play optimization approach, for an obtained prompt tuning backbone (e.g., CoOp [4]), MAO employs a **two-step** tuning strategy, performing prompt tuning separately by using image-text pairs on base tasks and unlabeled images on new tasks. Both tasks incorporate targeted Data-Driven Enhancement and Alterable Regularization. Details of MAO are as follows.

A. Preliminaries

Inheriting the settings of extant prompt tuning backbones, MAO introduces a frozen CLIP as foundation model, consisting of ViT-B/16 [15] image and text encoder, which are utilized for mapping image I and text T to embeddings with dimension d, denoted as $f(\cdot)$ and $g(\cdot)$.

The flow of mainstream prompt tuning is displayed in Fig. 1(a). Learnable prompt vector is normally organized as a set of tensors with length L for textual or optional visual input:

$$\boldsymbol{P} = [\theta]_1[\theta]_2 \dots [\theta]_L \tag{1}$$

The textual prompt $\boldsymbol{P}_t$ concatenates $\boldsymbol{P}$ with the $[CLASS]$ tokens containing all candidate classes $\boldsymbol{C} = \{T_i\}_{i=1}^{n}$. In contrast, visual prompt is typically integrated as the prefix of image patch tokens $(\boldsymbol{P}_v, I)$. During fine-tuning phase, prompt tuning applies cross-entropy loss to update the parameters of the learnable prompts:

$$\mathcal{L}_{\mathrm{CE}} = -\sum_i c_i \log p\left(y \mid I\right) \tag{2}$$

$$p(y \mid I) = \frac{\exp\left(\langle g(\boldsymbol{P}_{t_y}), f(\boldsymbol{P}_v, I)\rangle/\tau\right)}{\sum_{i=1}^{n} \exp\left(\langle g(\boldsymbol{P}_{t_i}), f(\boldsymbol{P}_v, I)\rangle/\tau\right)} \tag{3}$$

where c_i is the one-hot label of the i-th candidate in $\boldsymbol{C}$, and $\langle \cdot, \cdot \rangle$ represents cosine similarity. τ is a temperature coefficient defined by CLIP.

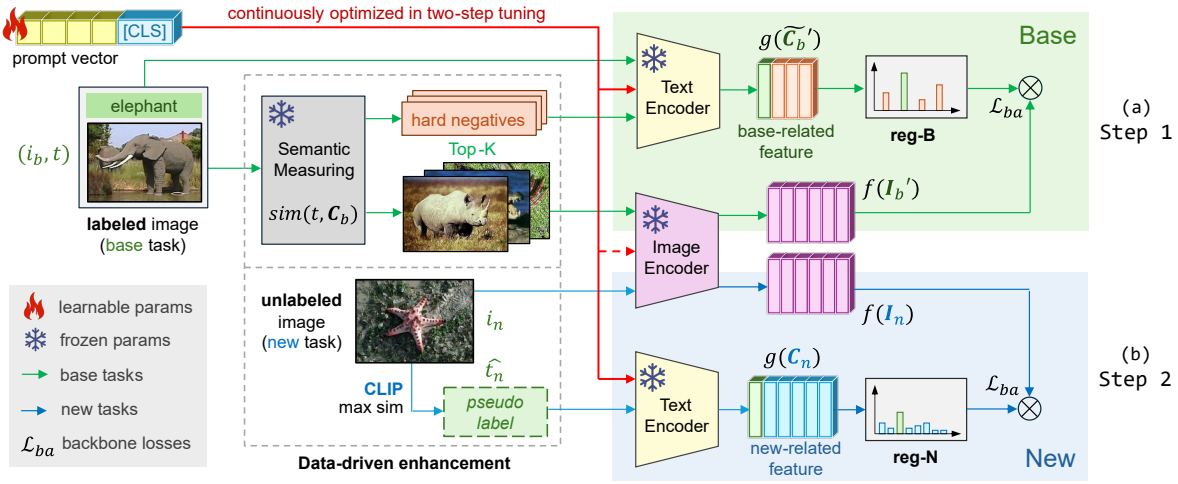


Fig. 2. Framework of proposed MAO. MAO builds a **two-step** fine-tuning structure without altering components of prompt tuning backbones. In (a) **base** tasks, MAO introduces a hard negative sampler as Data-Driven Enhancement (DDE), and an Alterable Regularization (reg-B) that guides the model in learning the feature distribution of hard negatives and keeps generalization. Then in (b) **new** tasks, rapid pseudo-labeling is performed on unlabeled images as DDE using shared-parameter CLIP, followed by reg-N to constrain the fine-tuning on new classes. The inference process follows the settings of the original backbones.

B. Model-Agnostic Optimization on Base-Class Task

To enable prompt tuning to adapt to feature distributions of both base and new tasks without increasing computational cost, MAO's two-step fine-tuning flow evenly splits the original total epoch of the backbone prompt learners. The first half is utilized for fine-tuning on the base-class tasks, while the latter half is dedicated to generalization enhancement by adapting unlabeled images to new-class tasks. To achieve equivalent base-class performance in fewer epochs, MAO introduces Data-Driven Enhancement and Alterable Regularization as below.

Data-Driven Enhancement. In base-class tasks, this process aims to construct a denser data distribution, enabling efficient learning of base-class features. Herein, MAO proposes a Hard Negative Sampler to guide prompt tuning in learning reconstructed image-text pairs that are tough to classify precisely, thereby achieving further fitting to the base class.

Specifically, as the Hard Negative Sampler, a pre-trained MiniLM [18] with semantic similarity metric is introduced, which is a compressed Transformer-based model, demonstrating remarkable performance in real-time inference on classification tasks. For each image-text pair (i_b, t) in base tasks passed by the original prompt tuning backbone, cosine similarity is utilized to filter the Top-K categories from the set of base classes $\boldsymbol{C}_b$ that possess the closet semantic distance to the embedding $\boldsymbol{e}_t$ tokenized from t, thereby constructing hard negatives $\boldsymbol{T}_b'$:

$$\boldsymbol{T}_b' = \mathrm{topK}_{c_i \in \boldsymbol{C}_b} \left(\frac{\langle \boldsymbol{e}_t, \boldsymbol{e}_{c_i} \rangle}{\|\boldsymbol{e}_t\| \, \|\boldsymbol{e}_{c_i}\|} \right), \quad \forall c_i \in \boldsymbol{C}_b \tag{4}$$

Afterwards, objects in $\boldsymbol{T}_b'$ are utilized as indices to randomly sample matching images from the pre-constructed training set, organized as a set of hard negative image-text pairs $(\boldsymbol{T}_b', \boldsymbol{I}_b')$. The effectiveness is verified in *Supplementary Material*.

Alterable Regularization. In prompt tuning backbones, cross-entropy loss $\mathcal{L}_{CE}$ is typically measured over entire base-class candidates $\boldsymbol{C}_b$, making it tough to specifically generalize the features of hard negatives. As an improvement, MAO introduces an online dynamic cross-entropy as Alterable Regularization (*reg-B* in Fig. 2), constraining the model by focusing on the feature distribution of hard negatives, while avoiding overfitting by importing randomness of dynamic perturbations.

For the mini-batch consisting of hard negatives $(\boldsymbol{T}_b', \boldsymbol{I}_b')$, MAO extracts all the contained classes and deletes duplicates to organize a candidate set $\widetilde{\boldsymbol{C}}_b' \subset \boldsymbol{T}_b'$ specific to the hard negatives. Since duplicates are excluded, it possesses a dynamic length $H \leq b \cdot \mathrm{top}\,K$, b signifies batch size. Under the constraint of $\widetilde{\boldsymbol{C}}_b'$, MAO obtains corresponding textual features $g(\widetilde{\boldsymbol{C}}_b') \in \mathbb{R}^{H \times d}$ through prompt tuning backbone, followed by L2 normalization to scale cross-modal feature distribution:

$$\hat{g}(\widetilde{\boldsymbol{C}}_b') = \frac{g(\widetilde{\boldsymbol{C}}_b')}{\|g(\widetilde{\boldsymbol{C}}_b')\|_2} \in \mathbb{R}^{H \times d} \tag{5}$$

Next, based on image-text features, an improved cross-entropy loss is proposed with only $\widetilde{\boldsymbol{C}}_b'$ as candidates:

$$\mathcal{L}_{CE}^{base} = -\sum_{i=1}^{H} c_i \log p\left(y \mid I_b\right), \; I_b \in \boldsymbol{I}_b', \; c_i \in \widetilde{\boldsymbol{C}_b'} \tag{6}$$

Beyond that, other possible loss functions in the prompt tuning backbones are maintained unchanged.

Since hard negatives are obtained online, they introduce dynamic **prior constraints** to the model, as well as a degree of perturbation for randomness. Overall, the above procedure can be considered as a type of implicit regularization. Theoretical explanation is visible in *Supplementary Material*. Experiments in Section IV-B verify its beneficial effect on generalization.

C. Model-Agnostic Optimization on New-Class Task

Inheriting the results on base tasks, in the latter half of the two-step fine-tuning, MAO exerts optimization on new classes

TABLE I
BASE-TO-NEW GENERALIZATION PERFORMANCE (%) OF 3 BACKBONE MODELS W/ OR W/O OUR MAO ON 11 DATASETS.

Model	Average of all			ImageNet			Caltech101			OxfordPets			StanfordCars			Flowers102		
	Base	New	H	Base	New	H	Base	New	H	Base	New	H	Base	New	H	Base	New	H
CoOp [4]	81.98	68.84	74.84	76.41	**68.85**	72.43	97.55	**94.65**	96.08	95.06	97.60	96.31	75.69	70.14	72.81	**96.96**	68.37	80.19
+MAO	**82.48**	**74.12**	**78.08**	**76.53**	68.82	**72.47**	**98.06**	94.20	**96.09**	**95.53**	**98.32**	**96.90**	**77.24**	**75.32**	**76.27**	96.77	**77.38**	**86.00**
MaPLe [9]	83.52	73.31	78.08	**76.91**	67.96	72.16	97.98	**94.50**	96.21	95.23	**97.67**	96.44	77.63	71.21	74.28	97.03	72.67	83.10
+MAO	**84.17**	**74.68**	**79.14**	76.79	**68.72**	**72.53**	**98.13**	94.47	**96.27**	**95.85**	97.54	**96.69**	**79.90**	**75.12**	**77.43**	**97.06**	**77.47**	**86.17**
PromptSRC [6]	83.45	74.78	78.87	**77.28**	70.72	73.85	97.93	**94.21**	96.03	95.41	**97.30**	**96.34**	76.34	74.98	75.65	**97.06**	73.19	83.45
+MAO	**84.53**	**75.38**	**79.69**	76.51	**72.53**	**74.47**	**98.13**	94.12	**96.08**	**95.59**	96.92	96.25	**80.91**	**76.01**	**78.38**	95.54	**77.94**	**85.85**

Method	Food101			FGVCAircraft			SUN397			DTD			EuroSAT			UCF101		
	Base	New	H	Base	New	H	Base	New	H	Base	New	H	Base	New	H	Base	New	H
CoOp [4]	90.49	91.47	90.98	37.33	24.24	29.39	**80.99**	74.10	77.39	80.09	49.88	61.47	**87.60**	51.62	64.96	83.66	66.31	73.98
+MAO	**91.03**	**91.63**	**91.33**	**39.56**	**31.79**	**35.25**	80.73	**76.29**	**78.45**	**80.44**	**59.66**	**68.51**	87.12	**67.74**	**76.22**	**84.28**	**74.20**	**78.92**
MaPLe [9]	89.85	90.47	90.16	40.82	**34.01**	**37.11**	**81.54**	75.93	78.63	82.18	55.63	66.35	94.96	**72.19**	**82.02**	84.55	74.15	79.01
+MAO	91.14	**91.23**	**91.18**	**41.88**	32.45	36.57	81.43	**76.78**	**79.04**	**83.14**	**62.02**	**71.04**	**95.65**	70.87	81.42	**84.89**	**74.83**	**79.54**
PromptSRC [6]	**90.83**	**91.58**	**91.20**	39.20	**35.33**	37.16	**82.28**	**78.08**	**80.13**	83.45	54.31	65.80	92.84	**74.73**	**82.80**	85.28	**78.13**	**81.55**
+MAO	90.79	91.23	91.01	**45.32**	33.29	**38.38**	81.51	77.42	79.41	**84.26**	**64.25**	**72.91**	**94.88**	70.10	80.63	**86.40**	75.39	80.52

TABLE II
CROSS-DATASET GENERALIZATION OF 3 BACKBONE MODELS W/ OR W/O OUR MAO ON IMAGENET AS SOURCE AND OTHER 10 DATASETS AS TARGETS.

Model	Source	Target										
	ImageNet	Avg.	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101
CoOp	71.25	64.98	**93.91**	89.97	**65.56**	67.88	85.86	22.11	**66.92**	42.55	**47.77**	67.30
+MAO	**71.33**	**65.54**	93.75	**90.73**	65.03	**69.96**	**85.88**	**23.01**	66.14	**45.92**	46.78	**68.20**
MaPLe	70.11	64.79	**93.67**	89.72	63.90	69.63	**85.79**	21.24	**67.05**	44.92	**44.84**	**67.17**
+MAO	**71.86**	**64.95**	93.35	**90.49**	**64.42**	**71.01**	85.44	**24.96**	65.82	**45.21**	43.37	65.45
PromptSRC	70.65	65.64	**93.43**	**89.92**	65.95	**71.05**	**86.21**	24.03	**67.63**	46.22	42.59	**69.39**
+MAO	**70.97**	**65.68**	93.35	89.02	**66.20**	67.93	86.02	**25.23**	67.08	**47.10**	**46.51**	68.33

through continual learning, exploiting unlabeled images from new classes. To sustain efficiency, MAO continues to apply the identical **few-shot** setup to sample unimodal images I_n in new classes (instead of loading entire dataset as in knowledge distillation-based methods [13]). Similarly, Data-Driven Enhancement and Alterable Regularization are introduced.

Data-Driven Enhancement. To exploit out-of-distribution unlabeled images from new classes without modifying any backbone components, MAO proposes a rapid pseudo-labeling strategy. With no attached parameters or additional losses, this approach reuses the foundation CLIP to sample pseudo-labels for few-shot unlabeled images as supervision signals, thus integrating them into fine-tuning.

Sharing image encoder $f(\cdot)$ and text encoder $g(\cdot)$ with prompt learner backbones, MAO performs zero-shot inference on unlabeled images i_n supervised by all new-class candidates $\boldsymbol{C}_n$, picking Top-1 with the highest confidence score as its pseudo-label $\hat{t_n}$. The calculation is executed by a similar approach as Eqn. 3:

$$\hat{t_n} = \underset{c \in \boldsymbol{C}_n}{\arg\max} \frac{\exp\left(\langle f(i_n), g(c) \rangle / \tau\right)}{\sum_{c' \in \boldsymbol{C}_n} \exp\left(\langle f(i_n), g\left(c'\right) \rangle / \tau\right)} \tag{7}$$

Resorting to the zero-shot ability of CLIP, this process constructs pseudo image-text pairs $(\boldsymbol{I}_n, \hat{\boldsymbol{T}}_n)$ with acceptable quality, boosting generalization by increasing data diversity.

Alterable Regularization. Feature optimization for new classes (*reg-N* in Fig. 2) is approximate to base tasks. The discrepancy is that during fine-tuning on new, MAO supersedes base-class objects with all new-class candidate $\boldsymbol{C}_n$, achieving implicit regularization by altering prior constraints of data distribution. Since the tokenization and normalization are already handled by Data-Driven Enhancement, the computational load can be further reduced. Analogously, the cross-entropy loss for new classes is formulated as:

$$\mathcal{L}_{CE}^{new} = -\sum_{i=1}^{N} c_i \log p\left(y = \hat{t_n} \mid I_n\right), \; I_n \in \boldsymbol{I}_n, c_i \in \boldsymbol{C}_n \tag{8}$$

Overall, through pseudo-label allocation and transformation of feature distribution, MAO organizes a tuning task that focuses on learning latent new-class representation without augmenting computational overhead. Such a design encourages prompt tuning backbone to benefit from new-class data, thereby effectually enhancing generalization capacity.

IV. EXPERIMENTS

A. Experimental Setup

Datasets. Following CoOp's benchmark [4], we apply 11 recognition-related datasets with various data distributions to

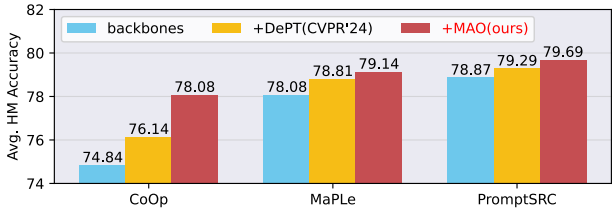


Fig. 3. Average HM performance of base-to-new generalization tasks of 3 backbones with plug-and-play methods, DePT [10] and our MAO.

TABLE III
COMPUTATION COST OF MAINSTREAM PROMPT TUNING BACKBONES AND OUR MAO ON FLOWERS102 DATASET.

Model	Learnable Params	Memory (MB)	Tuning Time Per Epoch	Inference FPS	HM Acc.
CoOp	**8K**	**1103.7**	**22.0s** (0.69x)	767.7	80.19
+MAO	**8K**	1176.9	31.9s (1x)	**772.0**	**86.00**
MaPLe	3.55M	3288.5	37.1s (1.16x)	765.6	83.10
CoPrompt	4.74M	3697.6	46.5s (1.46x)	768.1	85.71

make sufficient evaluation. These datasets are listed in Tab. I and Tab. II, containing generic and fine-grained objects.

Baselines. For comparison, 3 widely-recognized prompt learners, containing CoOp [4], MaPLe [9] and PromptSRC [6], are employed as baselines and backbone models for our MAO. Another leading plug-and-play module DePT [10] is also imported to validate MAO's adaptability. Additionally, we introduce CoPrompt [11] to contrast computational cost.

Implementation Details. As a model-agnostic approach, MAO thoroughly applies the initial parameter settings of the backbone models. For a fair comparison, all 3 original backbones are uniformly fine-tuned with $epoch = 20$ and batch size $b = 32$. In contrast, following the proposed two-step tuning strategy (Section III) , MAO assigns 10 epochs for base and new class optimization, respectively, and regulates the learning rate to $lr = 0.0035$. More details are in *Supplementary Material*.

B. Experimental Results

Base-to-New Generalization. Abided by the baselines' design, categories in each dataset are equally divided into base and new classes. MAO performs fine-tuning utilizing image-text pairs from base classes and unlabeled images from new classes, followed by accuracy evaluations on both test sets. The Harmonic Mean (HM) of base and new tasks is also calculated. As exhibited in Tab. I, MAO surpasses all 3 backbones in overall performance, with the most significant enhancement in new-class generalization compared to CoOp. Results demonstrate that without modification of model architecture, prompt tuning can be optimized by simply ameliorating the distribution of data and features.

Cross-Dataset Generalization. Using ImageNet tuned on all classes as source, in Tab. II, we conduct **zero-shot** inference on remaining 10 datasets to evaluate the transferability across diverse distributions. While source accuracy improves, MAO also attains higher accuracy on multiple target datasets. Remarkably, this is achieved without any target-task fine-tuning. We attribute this to MAO's Alterable Regularization design, which mitigates overfitting to the ImageNet source, thus guaranteeing favorable generalization to out-of-distribution data.

Comparison with Plug-and-play Baseline. We contrast MAO with another progressive plug-and-play model, DePT [10]. As illustrated in Fig. 3, MAO consistently surpasses DePT in HM accuracy, revealing better optimization level.

TABLE IV
ABLATION OF THE COMPONENTS IN MAO WITH PROMPTSRC BASELINE ON BASE-TO-NEW TASKS OVER 11 DATASETS. DDE: DATA-DRIVEN ENHANCEMENT. AR: ALTERABLE REGULARIZATION.

	Base	New		Average Acc.			Δ
	DDE+AR	DDE	AR	Base	New	H	
				83.45	74.78	78.87	
(a)	✓			84.53	74.95	79.45	+0.58
(b)		✓	✓	83.45	75.05	79.03	+0.16
(c)	✓	✓		84.53	75.02	79.49	+0.62
(d)	✓	✓	✓	84.53	75.38	79.69	+0.82

TABLE V
ABLATION OF THE PSEUDO-LABEL SAMPLER IN MAO WITH PROMPTSRC.

Model	Pseudo-Label Sampler	HM Acc.	Δ
PromptSRC	-	78.87	
+MAO	Foundation CLIP	79.69	+0.82
+MAO	Fine-tuned prompt	79.31	+0.44

C. Computational Cost

To confirm the efficiency of MAO, we employ multiple metrics to quantify the differences of computational cost between MAO, the associated backbone, and other prompt tuning approaches. As revealed in Table III, taking Flowers102 dataset as a paradigm, we contrast CoOp-based MAO with CoOp backbone, as well as MaPLe [9] and CoPrompt [11], which possess approximate HM accuracy.

Clearly, due to the model-agnostic characteristic of MAO, the quantity of learnable parameters sustains identical to CoOp, much less than MaPLe and CoPrompt. Meanwhile, GPU memory and inference speed of MAO are basically the same as CoOp. This implies that the hardware resource demand of MAO does not expand, supporting flexible deployment of prompt learners. Moreover, compared to CoPrompt, CoOp-based MAO acquires an equivalent level of performance while expending only 68.6% of fine-tuning time and 31.8% of GPU memory. More analyses are detailed in ablation study (Section IV-D).

D. Ablation Study

Validity of Proposed Components. Effect of components in MAO is examined in Table IV. Since Data-Driven Enhancement (DDE) and Alterable Regularization (AR) are bound together in base tasks, only their combination is considered. Compared with PromptSRC backbone, (a) importing base-class optimization improves base accuracy, and the introduction of Alterable Regularization also enhances the zero-shot generalization on new tasks to a certain extent (detailed analysis and verification

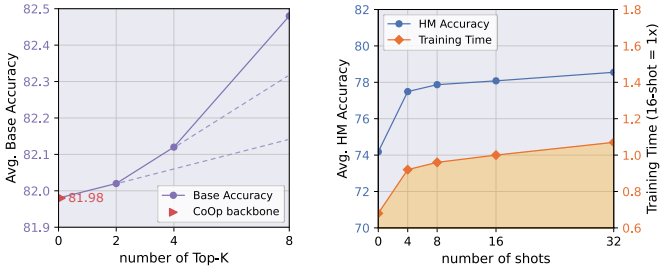


Fig. 4. The impact of (**Left**) the number of Top-K in Data-Driven Enhancement for base-class tasks and (**Right**) shots of unlabeled images for new-class tasks on accuracy and computational cost of CoOp-based MAO.

are available in *Supplementary Material*). Additionally, the absence of base-class optimization in (b) and AR module in (c) prevents the model from reaching optimal performance. Among them, the gap between (b) and (d) proves that prompt vector fine-tuned on the base class can serve as an effective supervision for generalization in new-class fine-tuning. In contrast, (d) with full setting performs the best, demonstrating the necessity of each component in MAO.

Pseudo-Label Sampler. We consider applying foundation CLIP or the prompt learner backbone tuned on base classes for pseudo-label sampling in new-class fine-tuning. It can be observed in Tab. V that the former performs better. We believe this is because that the tuned prompt learner backbone tends to fit the base classes, thereby weakening randomness and generalization on new-class sampling. In contrast, resorting to better global generalization, the foundation CLIP assigns pseudo-labels to unlabeled images with preferable quality.

Effect of Top-K in Hard Negative Sampler. As revealed in the left plot of Fig. 4, the base-class performance of MAO improves with the increase of K. Therefore, it is a priority to set a larger K for fine-tuning, while guaranteeing that the length of mini-batch H remains smaller than the total amount of base classes (otherwise, Alterable Regularization for base tasks would be invalidated). Herein, we set $K = 8$.

Impact of Shots. The right plot of Fig. 4 indicates that though the trend of growth gradually moderates, an increased shot of unlabeled images S brings a reinforcement in the HM performance of MAO. We believe that this can be attributed to the introduction of more diversified data in the process of new-class fine-tuning. Meanwhile, this leads to a corresponding increase in computational cost, with its trend approximating the gain in performance. Considering the marginal effect of performance enhancement, we recommend $8 \leq S \leq 32$ to equilibrate performance and computational cost.

V. Conclusion

We propose **M**odel-**A**gnostic **O**ptimization (MAO) for prompt tuning, improving performance by optimizing data distribution and feature representation without further computational cost. In fine-tuning on both base and new tasks, we introduce hard negative sampling and rapid pseudo-labeling as task-related Data-Driven Enhancement, constructing a dynamic dense data distribution for the model, and exploiting unlabeled images that original backbones cannot utilize. Subsequently, Alterable Regularization is applied to append implicit constraints during feature processing stage. Experiments reveal that MAO prominently enhances performance without demanding more computational resources. Overall, MAO provides an important reference and a novel solution for maintaining the lightweight and flexibility of prompt learners.

Acknowledgements

This work is supported by the China Scholarship Council (CSC), the UTS Top-Up Scholarship, and the Shanghai Institute of Intelligent Science and Technology, Tongji University. Computational facilities are provided by the UTS eResearch High Performance Compute Facilities and the Shanghai Technical Service Computing Center of Science and Engineering, Shanghai University.

References

[1] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *ICML*. PMLR, 2023, pp. 19730–19742.

[2] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, et al., "Learning transferable visual models from natural language supervision," in *ICML*. PMLR, 2021, pp. 8748–8763.

[4] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[5] H. Yao, R. Zhang, and C. Xu, "Visual-language prompt tuning with knowledge-guided context optimization," in *CVPR*, 2023, pp. 6757–6767.

[6] M. U. Khattak, S. T. Wasim, M. Naseer, S. Khan, M. Yang, and F. S. Khan, "Self-regulating prompts: Foundational model adaptation without forgetting," in *ICCV*, 2023, pp. 15190–15200.

[7] H. Li, L. Wang, C. Wang, J. Jiang, Y. Peng, and G. Long, "Dpc: Dual-prompt collaboration for tuning vision-language models," *arXiv preprint arXiv:2503.13443*, 2025.

[8] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *CVPR*, 2022, pp. 16816–16825.

[9] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan, "Maple: Multi-modal prompt learning," in *CVPR*, 2023, pp. 19113–19122.

[10] Ji Zhang, Shihan Wu, Lianli Gao, Heng Tao Shen, and Jingkuan Song, "Dept: Decoupled prompt tuning," in *CVPR*, 2024, pp. 12924–12933.

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

[12] K. Cai, K. Song, Y. Pan, and H. Lai, "Malip: Improving few-shot image classification with multimodal fusion enhancement," in *ICME*. IEEE, 2024, pp. 1–6.

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

[14] C. Jia, Y. Yang, Y. Xia, Y. Chen, Z. Parekh, H. Pham, Le, et al., "Scaling up visual and vision-language representation learning with noisy text supervision," in *ICML*. PMLR, 2021, pp. 4904–4916.

[15] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[16] M. Jia, L. Tang, B. Chen, C. Cardie, S. Belongie, B. Hariharan, and S. Lim, "Visual prompt tuning," in *ECCV*. Springer, 2022, pp. 709–727.

[17] Y. Zang, W. Li, K. Zhou, C. Huang, and C. Loy, "Unified vision and language prompt learning," *arXiv preprint arXiv:2210.07225*, 2022.

[18] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers," *Advances in Neural Information Processing Systems*, vol. 33, pp. 5776–5788, 2020.