DPC: Dual-Prompt Collaboration for Tuning Vision-Language Models

Haoyang Li[1,2] Liang Wang[1,2] Chao Wang[1*] Jing Jiang[2] Yan Peng[1*] Guodong Long[2*]
[1]Shanghai University, [2]University of Technology Sydney
haoyang.li-3@student.uts.edu.au, {cwang, pengyan}@shu.edu.cn, guodong.long@uts.edu.au

Abstract

*The Base-New Trade-off (BNT) problem universally exists during the optimization of CLIP-based prompt tuning, where continuous fine-tuning on base (target) classes leads to a simultaneous decrease of generalization ability on new (unseen) classes. Existing approaches attempt to regulate the prompt tuning process to balance BNT by appending constraints. However, imposed on the same target prompt, these constraints fail to fully avert the mutual exclusivity between the optimization directions for base and new. As a novel solution to this challenge, we propose the plug-and-play **D**ual-**P**rompt **C**ollaboration (`DPC`) framework, the first that decoupling the optimization processes of base and new tasks at the **prompt** level. Specifically, we clone a learnable parallel prompt based on the backbone prompt, and introduce a variable Weighting-Decoupling framework to independently control the optimization directions of dual prompts specific to base or new tasks, thus avoiding the conflict in generalization. Meanwhile, we propose a Dynamic Hard Negative Optimizer, utilizing dual prompts to construct a more challenging optimization task on base classes for enhancement. For interpretability, we prove the feature channel invariance of the prompt vector during the optimization process, providing theoretical support for the Weighting-Decoupling of `DPC`. Extensive experiments on multiple backbones demonstrate that `DPC` can significantly improve base performance without introducing any external knowledge beyond the base classes, while maintaining generalization to new classes. Code is available at: https://github.com/JREion/DPC.*

1. Introduction

Vision-Language Models (VLMs), represented by CLIP [27], have revealed cogent cross-modal open-domain representation and zero-shot capabilities. To further efficiently utilize pre-trained VLMs, Prompt Tuning acquires significant attention as a Parameter-Efficient Fine-Tuning (PEFT) method [48, 49]. Freezing the vision and text encoders, it employs a learnable lightweight prompt vector as a query to guide the output of CLIP towards the target task.

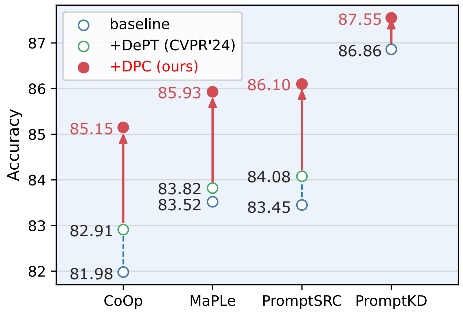
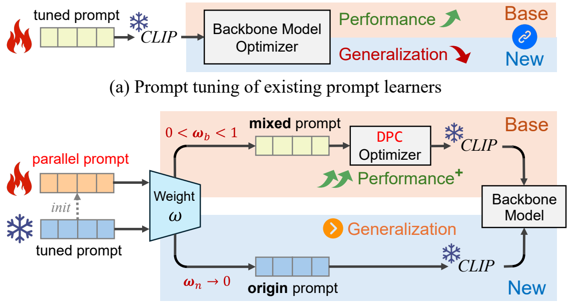


Figure 1. Average classification accuracy of 4 mainstream prompt tuning backbone models on base (target) classes over 11 datasets. `DPC` achieves state-of-the-art performance compared with baselines and another leading plug-and-play model DePT [45].

Unfortunately, the optimization process of prompt tuning often encounters a Base-New Trade-off (BNT) problem [45, 48]. As the prompt increasingly aligns with the target task, the model may overfit to the base (target) classes, resulting in reduced generalization performance on new (unseen) classes. To alleviate the BNT problem, previous approaches attempt to adjust loss functions [15, 42], apply constraints to prompts [43, 48], add extra feature extractors [7, 30], or involve external knowledge [16, 20, 46]. However, all these methods treat prompts as a single entity to be optimized for a balanced performance between base and new classes. Due to the shift of data distribution, the optimization directions of the two are likely to interfere with each other, making it tough to achieve the global optimum.

To mitigate such interference caused by the mutual exclusivity of optimization directions in prompt tuning for base or new tasks, we propose the Dual-Prompt Collaboration (`DPC`) framework. This approach is the first to introduce dual prompts optimized in two distinct directions, overcoming BNT problem by decoupling base and new tasks at the **prompt** level. Since the optimization in prompt tuning basically targets the learnable prompts, we believe

[*] Corresponding authors.



(a) Prompt tuning of existing prompt learners

(b) Decoupled framework of **D**ual-**P**rompt **C**ollaboration *(ours)*

Figure 2. Architecture comparison between (a) existing prompt learners that encounter Base-New Trade-off (BNT) problem and (b) our Dual-Prompt Collaboration framework that decouples the optimization directions of base and new tasks at **prompt** level.

that prompt-level decoupling is more fundamental. Specifically, based on the tuned prompt vector obtained by backbone prompt learner (e.g., CoOp [49]), we initialize and activate a separate parallel prompt. The two prompts are independently utilized for new class generalization and base class enhancement, respectively. To quantitatively regulate the optimization directions of dual prompts, we establish a flexible weight adjustment framework named *Weighting-Decoupling*. This module introduces a task-specific alterable hyperparameter ω, allowing dynamic adjustment of the weight distribution between dual prompts, thus preventing overfitting during base class optimization while fully retaining the generalization ability for unseen classes of the backbone model (typically when $\omega \to 0$). This framework fixes BNT problem by better controlling fine-tuning directions.

For the interpretability of this structure, in Section 4.4, we prove the invariance of feature channels in the prompt vectors during fine-tuning, i.e., continuous prompt optimization of DPC does not compromise the feature distribution of the prompt vectors. This indicates that the prompt tuning process can be correctly measured by DPC weights without causing feature bias.

Meanwhile, to further strengthen base class performance, we devise a *Dynamic Hard Negative Optimizer* to fine-tune the parallel prompt. This module is used to construct and learn to distinguish hard negative samples to further match the base classes. We first reuses the prompt tuning backbone with the collaboration of tuned prompt to spontaneously obtain Top-K negative results on the base classes as hard negative objects. Next, a Feature Filtering module applying L2 normalization is appended to extract hard negative text features aligned with paired images, while maintaining the distribution of base classes. Subsequently, we introduce an improved symmetric cross-entropy loss as an additional optimization term, constructing a more challenging vision-language contrastive learning task. This approach facilitates the DPC to more deeply fit the latent feature distribution of the base classes, while enhancing feature alignment between the vision and language modalities.

Our DPC is orthogonal to most existing prompt tuning backbones, exhibiting outstanding plug-and-play characteristics. Additionally, our model is self-contained, requiring no external knowledge beyond the train splits of base classes. In experimental part, we apply DPC in 4 backbone models [14, 15, 20, 49] with different forms of learnable prompts and conduct base-to-new generalization tasks on 11 recognition datasets. Results in Fig. 1 denote that DPC significantly enhances the base class performance in most of backbone models and datasets, while non-destructively preserves the generalization capability of the backbones.

Our main contributions can be generalized as follows:

1) We propose Dual-Prompt Collaboration (DPC) with flexible Weighting-Decoupling structure. To the best of our knowledge, this is the first prompt tuning enhancement strategy that decouples at the prompt level to overcome the BNT problem.
2) We design a novel Dynamic Hard Negative Optimizer, significantly enhancing the base class performance of DPC by establishing harder visual-text aligning tasks using dual prompts, achieving new State-Of-The-Art.
3) We introduce plug-and-play and self-contained features to the model, endowing it with outstanding adaptability and transferability while minimizing requirements of external knowledge.

2. Related Work

Prompt Tuning in VLMs. As frameworks that deeply integrate and align visual and textual modalities, Vision-Language Models (VLMs) [12, 18, 21, 27] have recently gained extensive attention, demonstrating remarkable potential in multiple cross-modal reasoning tasks. However, with the expansion of VLM network parameters, full-parameter fine-tuning requires substantial computational costs. In contrast, prompt tuning is proposed as a Parameter-Efficient Fine-Tuning (PEFT) technique on frozen pre-trained VLMs [8]. Entirely different from the template-based prompts utilized in CLIP or Large Language Models (LLMs) [19, 37], prompt tuning introduces a set of learnable lightweight vectors for replacing manually constructed hard prompts, and preserves the learnability of only prompt vectors during fine-tuning. As queries, these prompt vectors are continuously optimized, guiding the outputs of VLMs towards domain-specific data. Numerous approaches propose various forms of learnable prompts for constructing text features [34, 49, 50], image features [13, 26, 41], or joint visual-textual encoding [15, 44, 47]. Although the design of prompts varies across models, in our work, DPC constructs parallel prompts following the structure of backbone models, regardless of prompt forms.

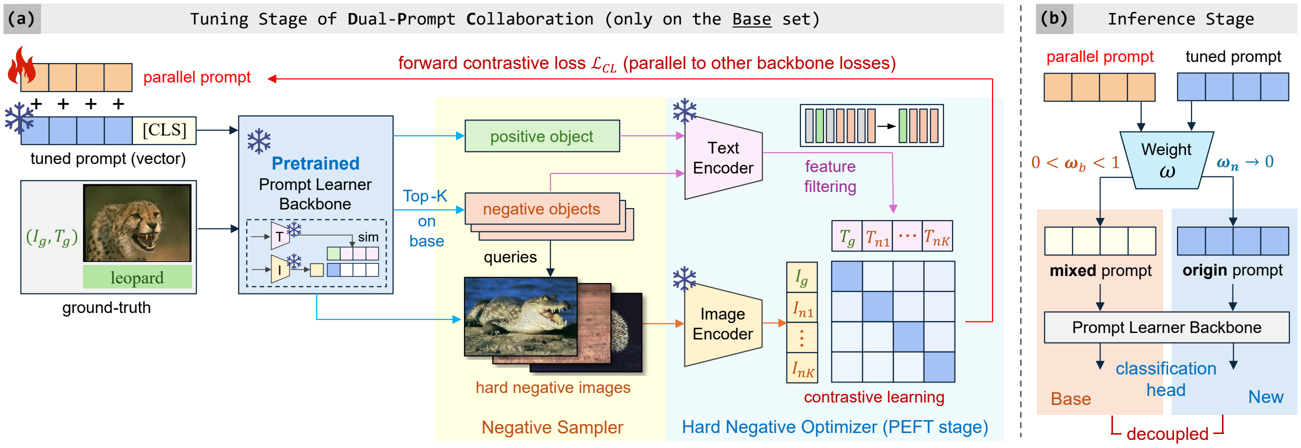


Figure 3. Overview of our proposed DPC. In (a) fine-tuning stage, DPC initializes parallel prompt $\boldsymbol{P}'$ based on tuned prompt $\boldsymbol{P}$ obtained by fine-tuning backbone. Negative Sampler applies tuned prompt $\boldsymbol{P}$ as query to sample hard negatives, then feed them into HNO optimizer to enhance base tasks. In (b) inference stage, DPC decouples base and new tasks by independent weight accumulation on dual prompts.

Base-New Trade-off of Prompt Tuning. The Base-New Trade-off (BNT) problem of CLIP-based prompt learner is first put forward in CoCoOp [48]. Essentially, it is due to the overfitting of prompt vector to the distribution of base classes after optimization, thereby reducing generalization to new classes. Numerous efforts are made to mitigate the impact of the BNT problem. Some approaches focus on appending generalization-related constraints during the prompt optimization process, like conditional context [48], semantic distance balancing [42], or consistency loss [15]. Other studies introduce additional feature extractors like Adapters [7, 30, 31] to address BNT problem by incorporating multi-scale feature mixing. Recently, methods involving the introduction of LLMs [36, 46] or knowledge distillation based on unlabeled images [20, 23] are also explored to enhance generalization through external data.

Although the aforementioned methods alleviate the BNT problem to some extent, we believe that there is still a common limitation in existing research: the tuning approaches are all applied to the same set of prompt vectors, potentially facing interference between the optimization directions of base and new classes. In contrast, our Dual-Prompt Collaboration strategy decouples the optimization processes on base and new at the prompt level, providing a more fundamental approach to effectively overcome the BNT problem.

3. Proposed Method

The framework of DPC is demonstrated in Fig. 3. As a plug-and-play enhancement method, for an obtained pretrained prompt tuning backbone, we first continuously optimize the newly established parallel prompt (§3.2) on base classes through Dynamic Hard Negative Optimizer (§3.3). Subsequently, during the inference phase, we employ the Weighting-Decoupling module (§3.4) to perform decoupled generalization tasks for base and new classes based on dual prompts. The details of DPC are introduced as follows.

3.1. Preliminaries

Identical to extant research on prompt tuning, DPC utilizes CLIP as the pre-trained VLM backbone model for image-text feature extraction and modality interaction. CLIP employs "A photo of a [CLASS]" as the prompt template for the text modality and imports ViT-based [5] visual encoder $f(\cdot)$ and text encoder $g(\cdot)$ to transform image V and text T into patch embedding and word embedding, respectively. During zero-shot inference, for a set of candidate objects $\boldsymbol{C} = \{T_i\}_{i=1}^{n}$, the matching probability between image features $f(V)$ and text features $g(T)$ is given by:

$$p(y \mid V) = \frac{\exp\left(sim(g(T_y), f(V))/\tau\right)}{\sum_{i=1}^{n} \exp\left(sim(g(T_i), f(V))/\tau\right)} \quad (1)$$

where $sim(\cdot,\cdot)$ denotes cosine similarity and τ is introduced as a temperature coefficient.

For transferring the foundation CLIP model to particular downstream tasks, prompt learner freezes the encoders $f(\cdot)$ and $g(\cdot)$, and appends a set of learnable vectors of length M to the textual or visual inputs, typically organized as:

$$\boldsymbol{P} = [\mathrm{p}]_1[\mathrm{p}]_2 \dots [\mathrm{p}]_M[CLASS] \quad (2)$$

In most existing studies, text prompts $\boldsymbol{P}_t$ replace original prompt template of CLIP as the input for text modality. Optional visual prompts $\boldsymbol{P}_v$ are commonly joined as prefix to visual modality, concatenated with image patch tokens as $(\boldsymbol{P}_v, V)$. During tuning process, cross-entropy loss is normally applied to continuously optimize prompt vectors:

$$\mathcal{L}_{\mathrm{CE}} = -\sum_{i} h_i \log p_\theta \left(y \mid V\right) \quad (3)$$

$$p_\theta(y \mid V) = \frac{\exp\left(sim(g(\boldsymbol{P}_{ty}), f(\boldsymbol{P}_v, V))/\tau\right)}{\sum_{i=1}^{n}\exp\left(sim(g(\boldsymbol{P}_{t_i}), f(\boldsymbol{P}_v, V))/\tau\right)} \quad (4)$$

where h_i is the one-hot label of the candidate object set $\boldsymbol{C}$.

3.2. Dual Prompt Initialization

In the initialization process of DPC, as a two-step tuning, we first execute moderate fine-tuning on the original prompt vector in prompt tuning backbone model, entirely adhering to the baseline settings to obtain the tuned prompt $\boldsymbol{P}$. Next, we freeze the tuned prompt and establish a set of learnable parallel prompt vectors $\boldsymbol{P}'$ based on it, with the form, size, and parameters cloned from the backbone model.

$$\boldsymbol{P}' := \boldsymbol{P} \quad (5)$$

The dual prompts are designed to separately store latent features specific to base and new classes, thus decoupling the tasks for base and new at the prompt level. The frozen tuned prompt $\boldsymbol{P}$ is applied to guarantee generalization during new-class inference, while the parallel prompt $\boldsymbol{P}'$ is utilized for deeper optimization specific to base classes during fine-tuning stage. Detailed pipelines of backbone models that DPC used are listed in *Supplementary Material A.2*.

It is noteworthy that if the epochs or time length of fine-tuning are strictly limited, the tuning epochs on the backbone model can be halved, with the latter half replaced by fine-tuning based on the DPC optimizer. Through ablation experiments in Sec. 4.3, we demonstrate that this setup can still achieve equivalent overall performance improvements for DPC without increasing computational cost.

3.3. Dynamic Hard Negative Optimizer

Independent of the original optimization process of the backbone, this module continuously fine-tunes the parallel prompt $\boldsymbol{P}'$ by constructing a more challenging optimization task on base, effectively enhancing base class performance. It consists of three sub-modules: *Negative Sampler*, *Feature Filtering* and *Hard Negative Optimizing*.

Negative Sampler. Replacing the random sampling strategy used by the backbone model, the Negative Sampler encourages the model to construct mini-batches utilizing hard negative samples that are tough to classify accurately. By reinforcing the difficulty of sample matching, the parallel prompt can be facilitated to fit the base class with tuning.

As the distribution of data shifts from zero-shot to base classes, the Top-K results inferred by the fine-tuned prompt learner generally exhibit more approximate semantics on base tasks. We verify this character in *Supplementary Material B.3*. Therefore, for the ground-truth image-text pairs (I_g, T_g) in the original mini-batch, we directly reuse the prompt tuning backbone, applying the frozen tuned prompt $\boldsymbol{P}$ as a query to dynamically obtain the Top-K inference results, and treat the $K-1$ samples other than the positive object T_g as hard negative objects T^-.

As subsequent process, hard negative objects and the positive object are concatenated to serve as the labels of the updated mini-batch $\boldsymbol{C}' = \left\{T_g, T_j^-\right\}_{j=1}^{K-1}$. Internal filtering is performed to exclude any identical objects within the mini-batch, and images matching the corresponding negative objects $\left\{V_j^-\right\}_{j=1}^{K-1}$ are randomly sampled from the training set to accommodate the following contrastive learning task. This process finally yields dynamic image-text pairs with size $L \leq b \cdot K$, where b denotes the batch size.

It is crucial that to avoid data leakage, the Top-K candidates of the negative sampler only contain base classes, and the image sampling range for constructing sample pairs is also restricted to the prebuilt train split. Compared to other hard negative samplers [32, 40], DPC achieves fully autonomous sample filtering without introducing any additional network parameters or external knowledge.

Feature Filtering. To maintain the complete performance of backbones, for obtaining the text features $g(\boldsymbol{C}') \in \mathbb{R}^{L\times d}$ generated by the text encoder from the set of hard negative objects $\boldsymbol{C}'$, DPC first performs L2 normalization on the text modality $\boldsymbol{C}$ during fine-tuning, which is constructed by the parallel prompt $\boldsymbol{P}'$ from all candidates n in base classes. The purpose is to keep the global feature distribution of prompt learner for base classes unchanged, preventing parameter shift when collaborating with the tuned prompt $\boldsymbol{P}$.

$$\hat{g}(\boldsymbol{C}) = \frac{g(\boldsymbol{C})}{\|g(\boldsymbol{C})\|_2} \in \mathbb{R}^{n\times d}, \;\; \boldsymbol{C} = \{T_i\}_{i=1}^{n} \quad (6)$$

Next, a selection matrix $Q \in \mathbb{R}^{L\times n}$ is introduced for extracting text features associated with hard negatives.

$$g(\boldsymbol{C}') = Q \cdot \hat{g}(\boldsymbol{C}) \quad (7)$$

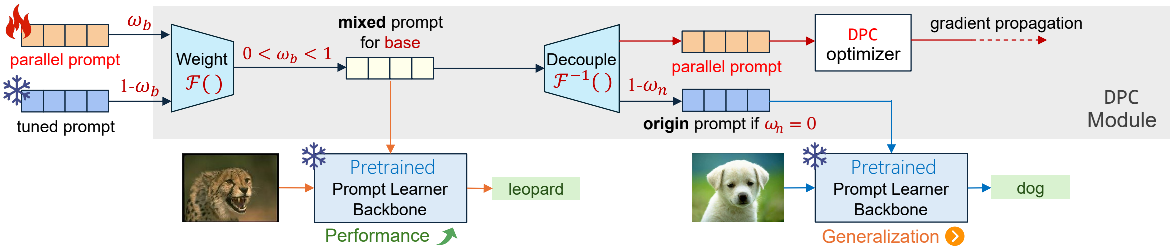
Q can be expressed as follows. $\boldsymbol{e}_{i_j}$ is the standard basis vector in $\boldsymbol{C}'$ where the label index position i_j corresponding to the j-th hard negative object is 1 and the rest are 0.

$$Q = \left(\mathbf{e}_{i_1},\, \mathbf{e}_{i_2},\, \ldots,\, \mathbf{e}_{i_L}\right), \;\; i \in \boldsymbol{C}' \quad (8)$$

With Feature Filtering, DPC reorganizes the image and text features input to the Hard Negative Optimizing process for subsequent optimization.

Hard Negative Optimizing. To achieve more robust cross-modal alignment on base classes, we upgrade the cross-entropy loss of the traditional prompt learner to stronger image-text contrastive loss for hard negatives. For a mini-batch $(V', \boldsymbol{C}')$ composed of hard negative image-text pairs with length L, we employ the InfoNCE loss function [38] to create a symmetric image-text contrastive learning task:

$$\mathcal{L}_{\mathrm{CL}} = -\frac{1}{L}\sum_{i=1}^{L}\left(\log p_\theta\left(y \mid \boldsymbol{C}'_i\right) + \log p_\theta\left(y \mid V'_i\right)\right) \quad (9)$$



(a) Inference on base classes ($0<\omega_b<1$)

(b) Inference on new classes ($\omega_n \rightarrow 0$)

Figure 4. Weighting-Decoupling structure in DPC. This structure allows DPC to continuously optimize the parallel prompt $\boldsymbol{P}'$ during the tuing phase and to endow separate accumulated weights to dual prompts ($\boldsymbol{P}$ and $\boldsymbol{P}'$) during (a) inference stage on base classes and (b) inference stage on new classes.

Among them, $p_\theta\left(y \mid \boldsymbol{C}_i'\right)$ represents the matching score of the text feature for the target image V_i', and vice versa.

$$p_\theta(y \mid \mathbf{C}_i') = \frac{\exp\left(\operatorname{sim}\left(f\left(V_i'\right), g\left(\boldsymbol{C}_i'\right)\right)/\tau\right)}{\sum_{j=1}^{L} \exp\left(\operatorname{sim}\left(f\left(V_i'\right), g\left(\boldsymbol{C}_j'\right)\right)/\tau\right)} \quad (10)$$

$$p_\theta(y \mid V_i') = \frac{\exp\left(\operatorname{sim}\left(g\left(\boldsymbol{C}_i'\right), f\left(V_i'\right)\right)/\tau\right)}{\sum_{j=1}^{L} \exp\left(\operatorname{sim}\left(g\left(\boldsymbol{C}_i'\right), f\left(V_j'\right)\right)/\tau\right)} \quad (11)$$

If other losses are appended to the backbone model, $\mathcal{L}_{\mathrm{CL}}$ is computed in parallel as a plug-and-play optimizer. We believe that the above setups can benefit both visual prompts and textual prompts during optimization.

Overall, the Dynamic Hard Negative Optimizer enables parallel prompt $\boldsymbol{P}'$ to better fit the base classes. Relying on the decoupled design of dual prompts, DPC optimizer does not compromise the generalization ability for new classes.

3.4. Weighting-Decoupling Module

Weighting-Decoupling Module (WDM) integrates the tuning and inference processes, allowing the tuned prompt $\boldsymbol{P}$ and parallel prompt $\boldsymbol{P}'$ to decouple and collaborate flexibly.

WDM uniformly acts on the input of dual prompts during both tuning and inference stage. As demonstrated in Fig. 4 (a), during model initialization, the Weighting sub-module $\mathcal{F}()$ is introduced, which combines the tuned prompt and parallel prompt into a mixed prompt $\widetilde{\boldsymbol{P}}_b$ by controlling the base-class-specific weighting coefficient ω_b constructed for base class inference.

$$\widetilde{\boldsymbol{P}}_b = \mathcal{F}(\boldsymbol{P}') = \omega_b \boldsymbol{P}' + (1-\omega_b)\,\boldsymbol{P} \quad (12)$$

Subsequently, the mixed prompt is passed into the Decoupling sub-module, which is the inverse transformation of the Weighting module. As illustrated in Fig. 4 (b), during this process, the mixed prompt is decomposed back into parallel prompt $\boldsymbol{P}'$ and original tuned prompt $\boldsymbol{P}$ by $\mathcal{F}^{-1}()$. The former is imported to the DPC optimizer in tuning process to realize integrated gradient propagation. In contrast, during new class inference, both are reassigned by a new-class-specific weighting coefficient ω_n to obtain $\widetilde{\boldsymbol{P}}_n$:

$$\widetilde{\boldsymbol{P}}_n = \omega_n \mathcal{F}^{-1}(\widetilde{\boldsymbol{P}}_b) + (1-\omega_n)\,\boldsymbol{P} \quad (13)$$

Above-mentioned design guarantees the model integrity while allowing independent weighting coefficients applied for base and new class inference, flexibly balancing the base class performance optimized by the parallel prompt $\boldsymbol{P}'$ and the latent features for new class generalization of the tuned prompt $\boldsymbol{P}$. We discuss the range of ω_b and ω_n by ablation study in Section 4.3.

4. Experiments

4.1. Experimental Setup

Datasets. Following the benchmark setting of CoOp [49], for tasks of base-to-new generalization and cross-dataset transfer, we use 11 recognition-related datasets with diverse data distributions, including ImageNet [3], Caltech101 [6], OxfordPets [25], StanfordCars [17], Flowers102 [24], Food101 [1], FGVCAircraft [22], SUN397 [39], DTD [2], EuroSAT [9] and UCF101 [33]. For cross-domain tasks, we select ImageNet-V2 [29], ImageNet-Sketch [35], ImageNet-A [11] and ImageNet-R [10], which exhibit domain shifts compared to ImageNet.

Baselines. We select 4 influential prompt learners as baselines and backbone models for our plug-and-play module. These contain CoOp [49] using *textual* prompts, MaPLe [14] employing *integrated visual and textual* prompts, and PromptSRC [15] and PromptKD [20], which utilize *separate visual and textual* prompts. Additionally, we compare the leading plug-and-play module for prompt learners, DePT [45], to validate the superiority of the prompt-level decoupling strategy of our DPC.

Implementation Details. We strictly follow the primary settings of the prompt tuning baselines, fine-tuning the backbone to obtain the tuned prompt, and subsequently fine-tuning the DPC optimizer utilizing the same hyperparameters. For a fair comparison, we set the batch size of

Method	**Avg. over 11 datasets**			ImageNet			Caltech101			OxfordPets		
	Base	New	H	Base	New	H	Base	New	H	Base	New	H
CoOp	81.98	68.84	74.84	76.41	68.85	72.43	97.55	94.65	96.08	95.06	97.60	96.31
+DPC	**85.15**	**68.84**	**76.13**	**77.72**	**68.85**	**73.02**	**98.58**	**94.65**	**96.58**	**95.80**	**97.60**	**96.69**
MaPLe	83.52	73.31	78.08	76.91	67.96	72.16	97.98	94.50	96.21	95.23	97.67	96.44
+DPC	**85.93**	**73.31**	**79.12**	**77.94**	**67.96**	**72.61**	**98.64**	**94.50**	**96.53**	**95.82**	**97.67**	**96.73**
PromptSRC	83.45	74.78	78.87	77.28	70.72	73.85	97.93	94.21	96.03	95.41	97.30	96.34
+DPC	**86.10**	**74.78**	**80.04**	**78.48**	**70.72**	**74.40**	**98.90**	**94.21**	**96.50**	**96.13**	**97.30**	**96.71**
PromptKD	86.86	80.55	83.59	**80.82**	74.66	**77.62**	**98.90**	96.29	**97.58**	**96.44**	97.99	**97.21**
+DPC	**87.55**	**80.55**	**83.91**	80.25	**74.66**	77.35	98.77	**96.29**	97.51	96.07	**97.99**	97.02

Method	StanfordCars			Flowers102			Food101			FGVCAircraft		
	Base	New	H	Base	New	H	Base	New	H	Base	New	H
CoOp	75.69	70.14	72.81	96.96	68.37	80.19	90.49	91.47	90.98	37.33	24.24	29.39
+DPC	**81.13**	**70.14**	**75.24**	**98.86**	**68.37**	**80.84**	**91.15**	**91.47**	**91.31**	**45.56**	**24.24**	**31.64**
MaPLe	77.63	71.21	74.28	97.03	72.67	83.10	89.85	90.47	90.16	40.82	34.01	37.11
+DPC	**79.56**	**71.21**	**75.15**	**98.20**	**72.67**	**83.53**	**91.35**	**90.47**	**90.90**	**49.78**	**34.01**	**40.41**
PromptSRC	76.34	74.98	75.65	97.06	73.19	83.45	90.83	91.58	91.20	39.20	35.33	37.16
+DPC	**82.28**	**74.98**	**78.46**	**97.44**	**73.19**	**83.59**	**91.40**	**91.58**	**91.49**	**46.74**	**35.33**	**40.24**
PromptKD	82.41	82.80	82.60	**99.24**	82.91	**90.34**	**92.59**	93.73	**93.15**	48.80	41.75	45.00
+DPC	**84.17**	**82.80**	**83.48**	98.96	**82.91**	90.23	92.41	**93.73**	93.07	**52.94**	**41.75**	**46.68**

Method	SUN397			DTD			EuroSAT			UCF101		
	Base	New	H	Base	New	H	Base	New	H	Base	New	H
CoOp	80.99	74.10	77.39	80.09	49.88	61.47	87.60	51.62	64.96	83.66	66.31	73.98
+DPC	**82.81**	**74.10**	**78.21**	**84.61**	**49.88**	**62.76**	**93.40**	**51.62**	**66.49**	**87.02**	**66.31**	**75.27**
MaPLe	81.54	75.93	78.63	82.18	55.63	66.35	94.96	72.19	82.02	84.55	74.15	79.01
+DPC	**82.02**	**75.93**	**78.86**	**85.48**	**55.63**	**67.40**	**98.33**	**72.19**	**83.26**	**88.14**	**74.15**	**80.54**
PromptSRC	82.28	78.08	80.13	83.45	54.31	65.80	92.84	74.73	82.80	85.28	78.13	81.55
+DPC	**83.63**	**78.08**	**80.76**	**86.88**	**54.31**	**66.84**	**96.25**	**74.73**	**84.13**	**88.99**	**78.13**	**83.21**
PromptKD	**83.53**	81.07	**82.28**	85.42	71.01	77.55	97.20	82.35	89.16	90.12	81.50	85.59
+DPC	83.28	**81.07**	82.17	**87.73**	**71.01**	**78.49**	**98.29**	**82.35**	**89.61**	**90.18**	**81.50**	**85.62**

Table 1. Base-to-new generalization performance of 4 backbone models w/ or w/o our `DPC` on 11 datasets. Benefiting from the decoupling structure at prompt level, `DPC` achieves general base class performance improvements while fully retaining new class generalization.

Method	**Source**	**Target**		Method	**Source**	**Target**	
	ImageNet	Avg. of cross-dataset	Avg. of cross-domain		ImageNet	Avg. of cross-dataset	Avg. of cross-domain
CoOp	71.25	64.98	60.31	PromptSRC	70.65	65.64	60.58
+DPC	**71.80**	**64.98**	**60.31**	**+DPC**	**71.42**	**65.64**	**60.58**
MaPLe	70.11	64.79	60.11	PromptKD	72.42	70.77	71.47
+DPC	**71.36**	**64.79**	**60.11**	**+DPC**	**74.43**	**70.77**	**71.47**

Table 2. Average performance of cross-dataset and cross-domain generalization tasks of 4 backbone models w/ or w/o our `DPC`.

the backbone to 32, while for `DPC`, we select a batch size of 4 and set the Top-K number of the Negative Sampler to $K = 8$, ensuring that the size of the mini-batch remains consistent during fine-tuning. According to ablation study, the collaboration weights are set to $\omega_b = 0.2$ (in MaPLe, $\omega_b = 1.0$) and $\omega_n = 1e$-6. Exceptionally, in PromptKD, due to its disparate settings (loading entire classes and fine-tuning based on all images in the dataset), we first maintain the original settings to obtain the tuned prompt. Next, the model is adjusted to sample few-shot image-text pair on base classes like other backbones, rendering the `DPC` optimizer learnable. Detailed implementation specifics of all

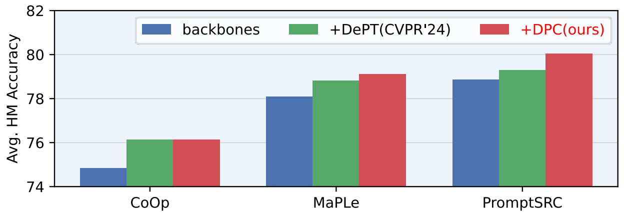
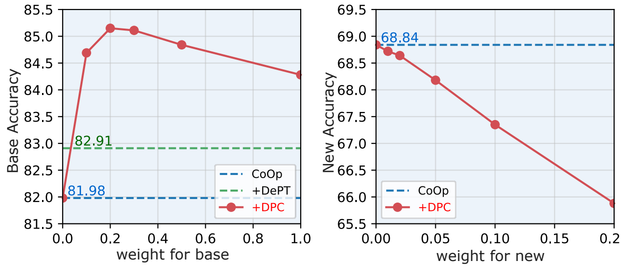


Figure 5. Average HM performance of base-to-new generalization tasks of 3 backbones with plug-and-play methods, DePT [45] and our DPC.



(a) impact of ω_b on base

(b) impact of ω_n on new

Figure 6. Impact of the collaboration weight for (a) base tasks and (b) new tasks in DPC. Detailed results are visible in Sup.Mat.B.2.

models are enumerated in *Supplementary Material* A.2.

4.2. Experimental Results

Base-to-New Generalization. Adhering to the baselines, categories in each dataset are evenly divided into base classes and new classes. Fine-tuning of both the backbone and DPC is executed only on the base classes, followed by inference on both base and new classes. H denotes the Harmonic Mean (HM) of the accuracy on base and new tasks. In conclusion, DPC achieves superior HM performance across all 4 backbones, while surpassing the current State-Of-The-Art PromptKD [20] on multiple datasets. As exhibited in Tab. 1, the performance improvement mainly stems from the optimization on the base classes. Additionally, for tasks on new classes, the decoupled structure of DPC is validated to thoroughly maintain the generalization performance of the original backbone.

Cross-Dataset and Cross-Domain Transfer. We train ImageNet on all classes as source and evaluate it on other datasets mentioned in Sec. 4.1 under zero-shot setting. Approximate to base-to-new tasks, DPC optimizes the performance on ImageNet, gaining enhancements across all backbones in Tab. 2. Meanwhile, for cross-dataset and cross-domain tasks involving unseen data distributions, the generalization level is consistently maintained through the coverage of raw tuned prompt $\boldsymbol{P}$.

Compare with Another Plug-and-play Method. To validate the effectiveness of DPC as a transferable module, we compare its HM performance with DePT [45], a plug-and-play prompt learner that decouples base and new tasks at the **feature** level, as displayed in Fig. 5. It is evident that the enhancement effect of DPC is superior or equal to DePT across the 3 baselines. We attribute this to the fact that decoupling at the **prompt** level is more thorough, furnishing a broader optimization space for the plug-and-play model. More detailed data is listed in *Supplementary Material B.7*.

	TS	DHNO	WE	DE	Base	New	H
					81.98	68.84	74.84
(1)	✓				82.69	68.39	74.86
(2)	✓	✓			84.28	64.12	72.83
(3)	✓	✓	✓		85.15	65.88	74.29
(4)	✓		✓	✓	82.23	68.84	74.94
(5)	✓	✓	✓	✓	**85.15**	**68.84**	**76.13**

Table 3. Ablation study of components in DPC with CoOp baseline on base-to-new tasks. TS: Two-Step tuning. DHNO: Dynamic Hard Negative Optimizer. WE & DE: Weighting-Decoupling.

4.3. Ablation Study

In this section, we discuss the impact of each DPC submodules, collaboration weights (ω_b, ω_n), Top-K sampling amount and fine-tuning epoch on model performance. The evaluation is based on base-to-new tasks of CoOp. More detailed experiments are listed in *Supplementary Material*.

Validity of Proposed Components. Tab. 3 illustrates the performance variation by introducing DPC sub-modules into backbones. Comparison between *(1)* tuning the original backbone continuously with another 20 epochs, *(2)* introducing Dynamic Hard Negative Optimizer (DHNO), and *(3)* performing dual-prompt weight accumulation (WE) on base classes, validates the effectiveness of DHNO and WE in enhancing base performance, respectively. However, the BNT problem can be observed, manifesting as a continuous decline of new-class performance. This suggests that methods without prompt-level decoupling tend to overfit to base classes. In contrast, *(4)* model with complete Weighting-Decoupling successfully maintains generalization performance, highlighting its necessity. Nonetheless, the absence of DHNO results in limited promotion of base tasks. By comparison, *(5)* with full configuration performs the best, confirming that both optimization directions for base and new classes are indispensable and proving the superiority of decoupling at the prompt level. Further ablation studies on DHNO sub-modules are in *Supplementary Material B.5*.

Influence of Collaboration Weights. The impact of collaboration weights (ω_b, ω_n) on base and new tasks is reflected in Fig. 6. Overall, DPC reaches optimal performance with ω_b=0.2 and ω_n=1e-6. By prompt decoupling, weights for base and new tasks are independently valued, avoiding the BNT problem in inference stage. Concrete data and

batch size	Top-K	Avg. Accuracy	Time
4	8	85.15 (+3.17)	1X
8	4	84.50 (+2.52)	0.69X
16	2	83.84 (+1.85)	0.59X

Table 4. Ablation study of the amount of Top-K in Negative Sampler. Size of mini-batch is fixed at 32 for fair comparison.

Model	epoch	total	Base	New	H
backbone	20	40	81.98	68.84	74.84
+DPC	+20		**85.15**	**68.84**	**76.13**
backbone	10	20	81.68	70.75	75.82
+DPC	+10		**83.99**	**70.75**	**76.80**
backbone	5	10	79.64	74.19	76.82
+DPC	+5		**82.89**	**74.19**	**78.29**

Table 5. Ablation study of the effect with less fine-tuning epoch.

analyses are detailed in *Supplementary Material B.2*.

Impact of Top-K Sampling Amount. Under the premise of a fixed mini-batch size L, we test diverse Top-K sampling amounts of the Negative Sampler (§3.3) and summarize the comparative results in Tab. 4. We observe that the base performance promotes with the growth of K, demonstrating that the Negative Sampler effectually collects and learns more similar hard negatives. From another perspective, affected by the data interaction bottleneck of the sampler, time of PEFT can be further reduced by decreasing K.

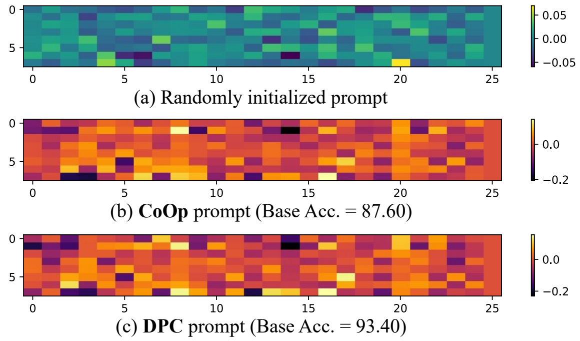
Less Fine-tuning Epochs. As an approach to reduce the computational cost, we consider smaller total amounts of epochs. The strategy is discussed in Section 3.2. As shown in Tab. 5, even when the epochs are cut back to half or a quarter, the base performance of DPC remains superior to the original backbone. Meanwhile, with the intensive generalization performance for new classes, HM performance growth relative to backbones can still be acquired.

Computational Cost. As discussed in *Supplementary Material B.6*, the additional computational cost of DPC is tiny. Compared to baselines, DPC does not introduce a significant increase in parameters, memory cost, or inference time.

4.4. Interpretability and Analysis

To provide a empirical analysis to the mechanism of the DPC weight accumulation structure, in this section, we demonstrate and analyze the feature channel invariance of the prompt vectors during fine-tuning that we discover.

As a visualization, we map the randomly initialized prompt vector, the tuned prompt $\boldsymbol{P}$ optimized by the backbone model, and the parallel prompt $\boldsymbol{P}'$ obtained after DPC optimization onto the feature maps in Fig. 7. We find that the feature distribution of parallel prompt fine-tuned by DPC is highly similar to the original tuned prompt.



(a) Randomly initialized prompt

(b) **CoOp** prompt (Base Acc. = 87.60)

(c) **DPC** prompt (Base Acc. = 93.40)

Figure 7. Visualization of feature maps of (a) randomly initialized prompt before tuning, (b) tuned prompt on CoOp backbone, and (c) optimized parallel prompt of DPC. Prompts are fine-tuned on base classes of EuroSAT [9] and are down-sampled for readability.

We reveal this phenomenon through the following analysis: During DPC optimization, the parallel prompt $\boldsymbol{P}'$ is initialized based on the tuned prompt $\boldsymbol{P}$, and both are fine-tuned on the tasks targeting identical base classes, following the design of the decoupled structure (§3.4). Benefiting from the Feature Filtering module (§3.3) that maintains the original feature distribution of the base, the latent feature channels basically remain unchanged during the DPC optimization on parallel prompt. This characteristic allows DPC to linearly control the shift of the mixed prompt $\widetilde{\boldsymbol{P}}_b$ towards the base classes by dynamically adjusting weights ω_b, thereby maximizing HM performance.

5. Conclusion

We propose DPC, the first approach that decoupling at **prompt** level to address the Base-New Trade-off problem in prompt tuning. During fine-tuning, the tuned prompt obtained from backbone is frozen to maintain generalization to new tasks, while also being applied as a query for Negative Sampler to spontaneously construct hard negatives for optimization. The activated parallel prompt significantly enhances base performance through the Dynamic Hard Negative Optimizer. During inference, by introducing decoupled weights for base and new, features from dual prompts are flexibly coordinated to maximize overall performance.

In future work, we will explore further improvements to DPC, such as adaptive parameterization of the weight coefficient ω and adaptation to a broader range of downstream tasks (e.g., object detection and semantic segmentation).

Acknowledgements

This work is supported by grants from the Natural Science Foundation of Shanghai, China (No. 23ZR1422800), Shanghai Institute of Intelligent Science and Technology in Tongji University, China Scholarship Council (CSC) and UTS Top-Up Scholarship.

References

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, pages 446–461. Springer, 2014. 5

[2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 5, 1

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 4

[5] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[6] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. 5, 4

[7] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2): 581–595, 2024. 1, 3, 6

[8] Zeyu Han, Chao Gao, Jinyang Liu, Sai Qian Zhang, et al. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint arXiv:2403.14608*, 2024. 2

[9] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 5, 8, 1

[10] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349, 2021. 5

[11] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15262–15271, 2021. 5

[12] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2

[13] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 2, 6

[14] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. 2, 5, 3

[15] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15190–15200, 2023. 1, 2, 3, 5

[16] Muhammad Uzair Khattak, Muhammad Ferjad Naeem, Muzammal Naseer, Luc Van Gool, and Federico Tombari. Learning to prompt with text only supervision for vision-language models. *arXiv preprint arXiv:2401.02418*, 2024. 1

[17] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 5

[18] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 2

[19] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021. 2

[20] Zheng Li, Xiang Li, Xinyi Fu, Xin Zhang, Weiqiang Wang, Shuo Chen, and Jian Yang. Promptkd: Unsupervised prompt distillation for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26617–26626, 2024. 1, 2, 3, 5, 7, 4

[21] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 2

[22] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 5

[23] Marco Mistretta, Alberto Baldrati, Marco Bertini, and Andrew D Bagdanov. Improving zero-shot generalization of learned prompts via unsupervised knowledge distillation. *arXiv preprint arXiv:2407.03056*, 2024. 3, 2

[24] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008. 5

[25] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 5, 1

[26] Wenjie Pei, Tongqi Xia, Fanglin Chen, Jinsong Li, Jiandong Tian, and Guangming Lu. Sa^{2}vp: Spatially aligned-and-

adapted visual prompt. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4450–4458, 2024. 2

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

[28] Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Fine-tuned clip models are efficient video learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6545–6554, 2023. 1

[29] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019. 5

[30] Shuvendu Roy and Ali Etemad. Consistency-guided prompt learning for vision-language models. *arXiv preprint arXiv:2306.01195*, 2023. 1, 3, 2

[31] Dominykas Seputis, Serghei Mihailov, Soham Chatterjee, and Zehao Xiao. Multi-modal adapter for vision-language models. *arXiv preprint arXiv:2409.02958*, 2024. 3

[32] Soonyong Song and Heechul Bae. Hard-negative sampling with cascaded fine-tuning network to boost flare removal performance in the nighttime images. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2843–2852. IEEE, 2023. 4

[33] K Soomro. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 5

[34] Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. Argue: Attribute-guided prompt tuning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28578–28587, 2024. 2

[35] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019. 5

[36] Yubin Wang, Xinyang Jiang, De Cheng, Dongsheng Li, and Cairong Zhao. Learning hierarchical prompt with structured linguistic knowledge for vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5749–5757, 2024. 3

[37] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 2

[38] Chuhan Wu, Fangzhao Wu, and Yongfeng Huang. Rethinking infonce: How many negative samples do you need? *arXiv preprint arXiv:2105.13003*, 2021. 4

[39] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 5

[40] Huang Xie, Okko Räsänen, and Tuomas Virtanen. On negative sampling for contrastive audio-text retrieval. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 4

[41] Chen Xu, Yuhan Zhu, Haocheng Shen, Boheng Chen, Yixuan Liao, Xiaoxin Chen, and Limin Wang. Progressive visual prompt learning with contrastive feature re-formation. *International Journal of Computer Vision*, pages 1–16, 2024. 2

[42] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6757–6767, 2023. 1, 3, 2

[43] Hantao Yao, Rui Zhang, and Changsheng Xu. Tcp: Textual-based class-aware prompt tuning for visual-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23438–23448, 2024. 1, 2, 3, 6

[44] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Unified vision and language prompt learning. *arXiv preprint arXiv:2210.07225*, 2022. 2

[45] Ji Zhang, Shihan Wu, Lianli Gao, Heng Tao Shen, and Jingkuan Song. Dept: Decoupled prompt tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12924–12933, 2024. 1, 5, 7

[46] Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Yu Qiao, Peng Gao, and Hongsheng Li. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15211–15222, 2023. 1, 3

[47] Yi Zhang, Ke Yu, Siqi Wu, and Zhihai He. Conceptual codebook learning for vision-language models. *arXiv preprint arXiv:2407.02350*, 2024. 2

[48] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16816–16825, 2022. 1, 3, 2

[49] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 1, 2, 5

[50] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15659–15669, 2023. 2