# **Revisiting Prompt Pretraining of Vision-Language Models**

Zhenyuan Chen<sup>1</sup>, Lingfeng Yang<sup>3</sup>, Shuo Chen<sup>4</sup>, Zhaowei Chen<sup>5</sup>, Jiajun Liang<sup>5</sup>, Xiang Li<sup>2,1\*</sup>

<sup>1</sup>IMPlus, VCIP, College of Computer Science, Nankai University <sup>2</sup>NKIARI, Shenzhen Futian <sup>3</sup>Nanjing University of Science and Technology <sup>4</sup>RIKEN <sup>5</sup>MEGVII Technology

{zhenyuanchen424, chaowechan}@gmail.com, yanglfnjust@njust.edu.cn, shuo.chen.ya@riken.jp, liangjiajun@megvii.com, xiang.li.implus@nankai.edu.cn

## **Abstract**

Prompt learning is an effective method to customize Vision-Language Models (VLMs) for various downstream tasks, involving tuning very few parameters of input prompt tokens. Recently, prompt pretraining in large-scale dataset (e.g., ImageNet-21K) has played a crucial role in prompt learning for universal visual discrimination. However, we revisit and observe that the *limited* learnable prompts could face underfitting risks given the *extensive* images during prompt pretraining, simultaneously leading to poor generalization. To address the above issues, in this paper, we propose a general framework termed Revisiting Prompt Pretraining (**RPP**), which targets at improving the fitting and generalization ability from two aspects: prompt structure and prompt supervision. For prompt structure, we break the restriction in common practice where query, key, and value vectors are derived from the shared learnable prompt token. Instead, we introduce unshared individual query, key, and value learnable prompts, thereby enhancing the model's fitting capacity through increased parameter diversity. For prompt supervision, we additionally utilize soft labels derived from zero-shot probability predictions provided by a pretrained Contrastive Language Image Pretraining (CLIP) teacher model. These soft labels yield more nuanced and general insights into the inter-class relationships, thereby endowing the pretraining process with better generalization ability. RPP produces a more resilient prompt initialization, enhancing its robust transferability across diverse visual recognition tasks. Experiments across various benchmarks consistently confirm the state-of-the-art (SOTA) performance of our pretrained prompts. Codes and models will be made available soon.

## 1 Introduction

Due to the remarkable capabilities of large Vision-Language Models (VLMs) [45, 36, 13], they are widely adopted for visual classification [68, 67, 30], object detection [18, 37], and semantic segmentation [66, 40], etc. In contrast to the traditional backbone fine-tuning or linear probing of vision-only models [19, 61], popular tuning techniques for VLMs include the cross-attention block [1, 35], adapter or projector [17, 41], and prompt learning [68, 30]. Prompt learning methods have gained prominence within the landscape of VLM tuning techniques due to their convenience and lightweight characteristics. These methods focus extensively on specialized tuning by training distinct prompts tailored for each domain or task [68, 67, 29, 30, 5]. Unfortunately, these specialized prompts, while optimized for specific narrow domains, often lack wider generalization capabilities.

To broaden applicability across diverse downstream domains using prompts, one promising solution is to introduce the *pretraining paradigm* of prompts over extensive images. Recently, <u>PrOMpt Pretraining</u> (POMP) [49] makes the first attempt by pretraining a shared token prompt that captures

<sup>\*</sup>Corresponding author. Team site: https://github.com/IMPlus-PCALab

a wide spectrum of visual concepts on the ImageNet-21K [10] dataset. This pretraining on a large-scale dataset is crucial for infusing the token prompt with semantic information, thereby facilitating universal visual discrimination. Nevertheless, we observe that the limited set of learnable prompts (i.e., very few learnable parameters) may exhibit susceptibility to underfitting when confronted with the copious image data encountered during prompt pretraining, in both training epoch and data amount (see Sec. 3.1 for a detailed description). In this study, we introduce a novel framework called Revisiting Prompt Pretraining (RPP) as a solution to tackle the underfitting challenge. RPP is devised to enhance the fitting and generalization capabilities of models by introducing a unique approach focused on two key aspects: prompt structure and prompt supervision.

Regarding prompt structure, we challenge the conventional practice of deriving query, key, and value vectors from a single shared learnable prompt token. Instead, we adopt a novel approach by introducing unshared individual query, key, and value learnable prompts, thereby increasing the parameter space available for optimization. This leads to enhanced fitting capabilities of the model. Particularly, we introduce Self-Attention Prompt Learning (SAPL), a meticulously designed learnable prompt featuring layer-by-layer replaceable query-key-value components. However, there is a substantial distribution discrepancy between the prompt pretraining dataset and CLIP training datasets. Fully aligning with the pretraining dataset might compromise the original generalization capability inherent in CLIP, posing a challenge to maintain its broader generalization applicability.

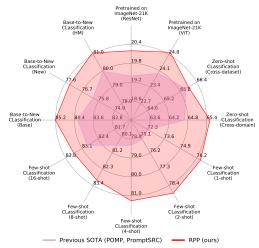


Figure 1: Our method outperforms previous SOTA models on a broad range of visual recognition tasks and datasets.

Further, to address the issue of limited generalization ability while simultaneously maintaining a strong pretraining fitting capability, we introduce a regularization technique termed Prompt Pretraining with Knowledge Distillation (PPKD) for prompt supervision. This approach allows for flexible adjustments to visual and textual prompt tokens while preserving robust generalization capabilities from the larger-scale CLIP teacher. Benefiting from the excellent zero-shot generalization ability of the larger-scale CLIP teacher model on the pretraining dataset, we can enhance generalization while simultaneously preserving fitting capabilities for downstream tasks. Additionally, drawing from previous works such as [6, 7], we provide theoretical analyses for the generalization ability of RPP. We demonstrate that as the weight assigned to the regularization loss increases, the regularization loss gradually decreases. This convergence allows the model to achieve a reduced upper bound on generalization error, consequently enhancing the overall generalization capability of RPP.

The results presented in Fig. 1 highlight the superior performance of RPP over previous SOTA models across various visual recognition tasks and datasets. Specifically, RPP records a 0.9% improvement in ImageNet-21K validation accuracy over POMP [49]. For zero-shot generalization, RPP shows a 0.43% enhancement with an average accuracy across 14 datasets. Notably, in a 16-shot setting, RPP shows enhancements in few-shot and base-to-new generalization accuracies across 11 classification datasets, with improvements of 0.58% and 1.12% respectively, compared to PromptSRC [30].

Our contributions can be summarized as follows:

- We are the first to explicitly identify the underfitting issue encountered during Prompt Pretraining in VLMs.
- We propose the SAPL prompt structure and the PPKD prompt supervision to alleviate the underfitting risk while maintaining its generalization.
- We present a theoretical analysis of generalization ability of RPP, demonstrating a reduction in the upper bound on generalization error that improves overall generalization performance.
- Compared to existing published works, our method consistently achieves SOTA results on multiple datasets under few-shot/base-to-new transfer, with an average improvement of 0.58% and 1.13% points.

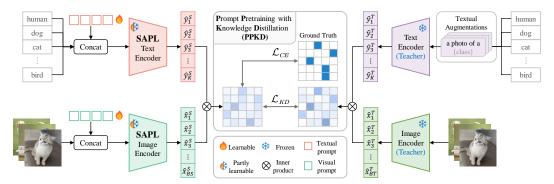


Figure 2: **An overview of our proposed pretraining framework.** Firstly, we propose SAPL Text/Image Encoder, optimizing individual query, key, and value embeddings directly and explicitly (Sec. 3.3). Next, we employ a frozen teacher model to supervise the student model's learnable prompts using regularization loss from knowledge distillation (Sec. 3.4). Further, we provide theoretical results for the generalization error bound of RPP (Sec. 3.5).

#### 2 Related Work

#### 2.1 Prompt Learning

The technique of prompt learning is initially introduced as an improvement to manual prompting in the field of Natural Language Processing (NLP) for transferring knowledge from pretrained models to specific downstream domains [51, 27, 65, 34, 38].

In recent years, the application of prompt learning has expanded into the realms of vision [56, 26, 3, 25] and vision-language [67, 29, 30, 47]. These approaches involve two main branches: the text branch and the image branch. In the text branch, CoOp [68] employs learnable tokens instead of manually designed suffixes in text inputs, such as "a photo of a class", to enhance the transfer capabilities in classification tasks. Subsequently, CoCoOp [67] introduces instance-conditional prompts to mitigate model overfitting. These methods adapt to various tasks, including open vocabulary [16] and visual grounding [46], etc. In the image branch, vision prompt tuning serves as an efficient training technique alongside backbone fine-tuning or linear probing. VPT [56] appends learnable tokens before the image sequence but keeps the entire backbone fixed to transfer vision-only models. A concurrent approach [3] introduces pixel-wise prompts for spatially structured images instead of sequences. Furthermore, some methods leverage joint textual and visual prompt tuning based on CLIP [45] for improvements [63, 29, 30].

Moreover, prompt pretraining with large-scale data has been incorporated into enhanced pretrained models, such as POMP [49]. However, upon revisiting, it is observed that the *limited* learnable prompts may face underfitting risks, especially given the *extensive* images during prompt pretraining. To address the underfitting issue of POMP, this paper introduces a general framework termed RPP, aiming to enhance fitting and generalization abilities by focusing on two key aspects: prompt structure and prompt supervision, as shown in Fig. 2.

#### 2.2 CLIP Distillation

Knowledge Distillation (KD) [24, 23] emerges as a common method for transferring knowledge from teacher models with more parameters and better performance to more deployable student models.

Distillation for VLMs has gained prominence in recent years, with efforts to convert VLMs knowledge into vision-only models during model pretraining [57, 13, 55], open vocabulary detection [18, 64], or semantic segmentation [66, 28]. Additionally, some methods directly distill VLMs to VLMs through ordinary distillation [62, 14, 58, 39], self-distillation [59, 2], and linear probing [33]. DistillVLM [14] transfers knowledge through the intermediate representations of each proposal generated from pretrained detectors. CLIP-KD [62] explores various distillation strategies, including relation, feature, gradient, and contrastive paradigms to assess their impact on CLIP distillation. TinyCLIP [58] learns cross-modal feature alignment from teacher to student in a visual-textual affinity space. CLIPSelf [59] distills CLIP itself using dense feature maps and corresponding predictions. LP-CLIP [33] employs a single linear probing layer for distillation.

In addition to the hard one-hot labels that may present optimization challenges for specific prompt tokens, we introduce soft labels derived from a pretrained large-scale CLIP teacher's zero-shot probability predictions.

#### 3 Method

#### 3.1 Revisiting Prompt Pretraining

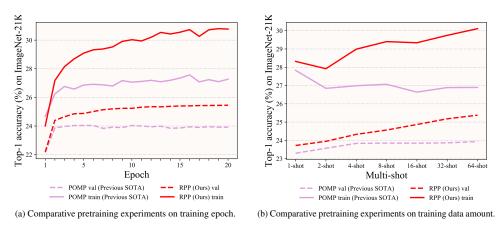


Figure 3: **Empirical evidence of underfitting of current practice.** All methods use the same optimization strategy with 5 epochs and the same data.

Fig. 3 illustrates the training and testing accuracy of our method compared with the POMP method on the ImageNet-21K dataset, in terms of training epoch and data amount, respectively. The pink lines denote the performance of the POMP method, while the red lines represent our approach. In Fig. 3a, the training and testing accuracy of the POMP method become relatively flat and saturated after around 4 epochs, indicating negligible further fitting to the dataset. In Fig. 3b, the POMP method reaches its fitting capacity ceiling at 4-shot, and as training data is further increased, the flattening curve highlights an underfitting issue. In contrast, our method maintains its pretraining fitting capability across extended training epochs and larger data amounts. A detailed description of the method will be provided in Sec. 3.3 and 3.4.

#### 3.2 Preliminaries

CLIP [45] contains an image encoder and a text encoder, denoted as **ImgEnc** and **TextEnc**, respectively. For zero-shot transfer, the input image  $I = \{I_i | I_i \in \mathbb{R}^{3 \times H \times W}, i = 1, 2, \cdots, N\}$  is embedded into a token sequence and passed through the image encoder, denoted as **PatchEmb**. The corresponding text  $T_c$  is the categories prepended with a suffix, appearing as "a photo of a [class]".  $T_c$  is tokenized and embedded into a feature space via the text encoder, denoted as **Tokenizer**:

$$x_i = \operatorname{ImgEnc}(\operatorname{PatchEmb}(I_i)), \tag{1}$$

$$y_c = \mathbf{TextEnc}(\mathbf{Tokenizer}(T_c)), \tag{2}$$

where  $\mathbf{y} = \{\mathbf{y_c} | c = 1, 2, \cdots, C\}$  with  $\mathbf{y_i} \in \mathbb{R}^d$ , and  $\chi = \{\mathbf{x_i} | i = 1, 2, \cdots, N\}$  with  $\mathbf{x_i} \in \mathbb{R}^d$  represent the global features of the image and text outputs, respectively. Then the normalized embeddings are derived as  $\hat{\mathbf{x_i}} = \mathbf{x_i} / \|\mathbf{x_i}\|_2$  and  $\hat{\mathbf{y_c}} = \mathbf{y_c} / \|\mathbf{y_c}\|_2$ . The probabilities concerning each class are calculated as:

$$s_i = \frac{\exp(\hat{\mathbf{x}}_i \cdot \hat{\mathbf{y}}_c / \tau)}{\sum_{j=1}^{C} \exp(\hat{\mathbf{x}}_i \cdot \hat{\mathbf{y}}_j / \tau)},$$
(3)

where C denotes the number of classes,  $(\cdot)$  represents the dot product, and  $\tau > 0$  is the temperature.

Different from manually designed instructions, prompt learning provides a solution to transfer knowledge through learnable tokens in the form of continuous parameters. There are two aspects to prompt learning: the textual branch and the visual branch. In the first one, following common practices [68, 29], "a photo of a [class]" is replaced with " $P^t$  [class]", where  $P^t = \{p^t\}_m$ ,  $(m \in \mathbb{N}_M)$ , is the learnable prompt, and M denotes the number of contents. In terms of visual prompting [26,

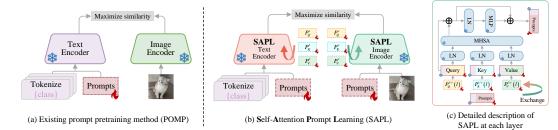


Figure 4: A detailed description of our proposed SAPL prompt structure. (a) Existing methods adopt uni-modal prompting techniques to fine-tune CLIP representations as prompts are learned only in a single branch of CLIP (language or vision). (b) Our SAPL explicitly and directly optimizes the individual query, key, and value embeddings. (c) Detailed description of SAPL at each layer.

63, 29], another set of learnable tokens is extended after the image patches:  $\{I, P^v\}$ . Then they are processed in place of the original image patches and text inputs. Subsequently,  $\{P^t, P^v\}$  are learned to adapt to specific data domains.

In addition to the integration of learnable prompts in the input space, there are methods [29, 26] that propose to incorporate tokens in the deep layers of text / vision encoders, known as multi-layer prompt learning. Denoting the output of the l-th transformer block in the text encoder as  $\{P^t(l), F^t(l)\}$ , where  $P^t(l)$  and  $F^t(l)$  are the intermediate prompt and text features, respectively. Then  $P^t(l)$  is replaced with learnable prompts, which are additionally initialized in the l-th layer of the encoder to introduce learnable tokens in deeper transformer layers. Likewise, within the forward process in the vision encoder,  $\{P^v(l), F^v(l)\}$  undergoes the same operation.

## 3.3 Self-Attention Prompt Learning

As shown in Fig. 4, to mitigate potential underfitting concerns related to the model architecture, we employ a methodology utilizing independent learnable prompts. These prompts are designated for substitution before the self-attention computation at each layer of the model. Particularly, before computing the forward propagation of the current layer with input  $X^{t/v}(l) = \{P^{t/v}(l), F^{t/v}(l)\}$ , we utilize  $\mathbf{Rep}(\cdot,\cdot)$  to replace the learnable prompt from the previous layer with separated learnable prompts  $P_*^{t/v}(l) = \{P_Q^{t/v}(l), P_K^{t/v}(l), P_V^{t/v}(l)\}$ . This approach diverges from traditional methods that typically utilize a single shared prompt, represented by  $P^{t/v}(l)$ .

It is essential to acknowledge that this method applies when the backbone is ViT [12], functioning for both  $X^t(l)$  and  $X^v(l)$ . Conversely, for the ResNet [20] backbone, it exclusively applies to  $X^t(l)$ . This operation aims to broaden the optimization scope within the parameter space while upholding consistent feature scaling. The computation of the current layer is outlined as follows (the following formulas do not distinguish t/v):

$$\begin{cases} X_{Q}^{*}\left(l\right) = \mathbf{LN}\left(\mathbf{Rep}\left(X\left(l-1\right), P_{Q}\left(l\right)\right)\right), \\ X_{K}^{*}\left(l\right) = \mathbf{LN}\left(\mathbf{Rep}\left(X\left(l-1\right), P_{K}\left(l\right)\right)\right), \\ X_{V}^{*}\left(l\right) = \mathbf{LN}\left(\mathbf{Rep}\left(X\left(l-1\right), P_{V}\left(l\right)\right)\right), \end{cases} \tag{4}$$

where  $l \in \{2,3,\ldots,12\}$  represents the l-th layer, and  $\mathbf{LN}(\cdot)$  denotes LayerNorm. We normalize each prompt to maximize the parameter space while simultaneously ensuring that features do not diverge significantly from the original CLIP distribution. Then we proceed with the remaining calculations:

$$\hat{X}(l) = X_Q(l) + \mathbf{MHSA}\left(X_Q^*(l), X_K^*(l), X_V^*(l)\right), 
X(l) = \hat{X}(l) + \mathbf{MLP}\left(\mathbf{LN}\left(\hat{X}(l)\right)\right),$$
(5)

where MHSA  $(\cdot, \cdot, \cdot)$  stands for Multi-head Self Attention, MLP  $(\cdot)$  stands for Multi-Layer Perceptron, and  $X_Q(l)$  means exchanged query prompt without layernorm.

#### 3.4 Prompt Pretraining with Knowledge Distillation

As shown in Fig. 2, we conduct pretraining of multi-layer prompts by leveraging the extensive ImageNet-21K dataset, pioneering the transfer of robust embedded knowledge from a larger-scale

CLIP model to a more computationally efficient scale. As detailed in Sec. 3.3, we utilize the layer-by-layer replaceable  $\{P_*^t(l), P_*^v(l)\}$  as the learnable parameters for the student network, while employing the original pretrained large-scale CLIP model as the teacher network.

We aim to acquire prompt pretrained CLIP endowed with universal generalization capabilities through disciplined pretraining via distillation on the ImageNet-21K dataset. Nevertheless, direct pretraining on 21K data demands over 300GB of GPU memory [49]. To address this challenge, we adopt methodologies inspired by POMP [49], specifically employing local contrast and local correction techniques. In particular, when presented with a batch of input images, our process involves sampling K classes(K << C). This includes selecting the respective ground-truth class and incorporating K-1 randomly selected samples as shared negative classes. Therefore, the probabilities for sample classes are calculated as:

$$\hat{s}_{i} = \frac{\exp(\hat{x}_{i} \cdot \hat{y}_{i}^{ram} / \tau)}{\exp(\hat{x}_{i} \cdot y_{i}^{gt} / \tau) + \sum_{j=1}^{K-1} \exp(\hat{x} \cdot \hat{y}_{j}^{ram} / \tau + m)},$$
(6)

where  $y_i^{gt}$  denotes the ground-truth class of sample i and  $\hat{y}_i^{ram}$  denotes the randomly selected K-1 negative classes. To address the inherent bias in the prompt optimization direction due to the absence of other negative classes, we introduce a local correction term denoted as m into the probability of negative classes. This term aims to incentivize the positive logit to surpass the negative logits by a predefined margin. The formulation for m is as follows:

$$m = -\log((K-1)/(C-1)). (7)$$

Drawing from the aforementioned strategies, we incorporate knowledge distillation loss and cross-entropy loss during the pretraining phase. To enhance text embedding diversity in the teacher model, we employ textual augmentations. This involves the random selection of 60 prompt templates for the text encoder from the comprehensive template list provided in [45]. Then, we introduce logit-level consistency regularization by conditioning the distribution of prompted logits on the teacher logits distribution. This is achieved through the minimization Kullback-Leibler Divergence of the following loss function:

$$\mathcal{L}_{KD} = \frac{1}{N \cdot K} \sum_{i=1}^{N} \sum_{k=1}^{K} s_{ik}^{T} \log \frac{s_{ik}^{T}}{s_{ik}^{S}}, \tag{8}$$

where  $\hat{s}^S_{ik} \in [0,1]$  and  $\hat{s}^T_{ik} \in [0,1]$  represent student's and teacher's probability prediction of i-th training data in the k-th sampled class. For image classification on the ImageNet-21K dataset  $\mathcal{D}$ , learnable prompts  $\{P^t_*(l), P^v_*(l)\}$  interact with frozen **ImgEnc** and **TextEnc** are optimized with the cross-entropy loss,  $\mathcal{L}_{CE}$ , as:

$$\mathcal{L}_{CE} = \mathbb{E}_{(\hat{x}, y^{gt}) \sim \mathcal{D}} \mathcal{L}\left(\hat{s}^{S}\left(\Theta\right), y^{gt}\right). \tag{9}$$

We use  $\lambda > 0$  as the loss balancing hyper-parameters. Our overall pretraining objective thus becomes:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda \mathcal{L}_{KD}. \tag{10}$$

#### 3.5 Theoretical Analysis

In this section, we provide theoretical results for the generalization error bound of RPP. All proofs of theorems are given in the *appendix* Sec. A.

We define the following optimization objectives according to Eq. (10):

$$\min_{\Theta \in \mathbb{R}^{d}} \underbrace{\frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(\hat{s}_{i}^{S}\left(\Theta\right), y_{i}^{gt}\right)}_{\mathcal{L}_{KD}} + \lambda \underbrace{\mathcal{L}\left(\hat{s}^{S}\left(\Theta\right), \hat{s}^{T}\right)}_{\mathcal{L}_{KD}}, \tag{11}$$

where  $\Theta$  represents the set of learning prompts  $\{P_*^t, P_*^v\}$  of the proposed network, with d denoting the dimensionality of the learning parameters. Now we further analyze the effectiveness of RPP by offering the generalization error bound. Such a bound evaluates the bias between the generalization error  $\varepsilon\left(\Theta\right) := \mathbb{E}_{(\hat{s}^S, y^{g^t}) \sim \mathcal{D}} \mathcal{L}\left(\hat{s}^S\left(\Theta\right), y^{gt}\right)$  and empirical error  $\bar{\varepsilon}_{\chi}\left(\Theta\right) := \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(\hat{s}_i^S\left(\Theta\right), y_i^{gt}\right)$ , where D is the real data distribution and  $\mathbb{E}\left(\cdot\right)$  denotes the expectation function.

Table 2: Cross-dataset and cross-domain evaluation for image classification. The backbone is ViT/B-16. Overall, RPP secures the highest mean accuracy, denoting superior generalization capabilities. The designation "-" signifies that the corresponding entry is unpopulated or not applicable. Each number in the figure represents the validation set accuracy (%) of the current dataset.

	Sou	rce		Target (cross-dataset)							Target (cross-domain)							
	ImageNet-21K	$Im_{ageNet}$	Caltech 101	$O_{\chi for dPet_S}$	$S_{lant}$ ord $C_{ars}$	$F_{lowers102}$	$F_{00dI0I}$	$^{FGV_{CAircnaft}}$	SUN397	DTD	EuroSAT	$^{UCFIOI}$	Average	ImageNetV2	Image/Vet-S	$^{Im_{ageNet}}$	<sup>Im</sup> ageNet-R	Average
CoOp [68]	-	✓	93.7	89.1	64.5	68.7	85.3	18.5	64.2	41.9	46.4	66.6	63.9	64.2	48.0	-	75.2	62.5
CoCoOp [67]	-	✓	94.4	90.1	65.3	71.9	86.1	22.9	67.4	45.7	45.4	68.2	65.7	64.1	48.8	-	76.2	63.0
VPT [11]	-	✓	93.7	90.6	65.0	70.9	86.3	24.9	67.5	46.1	45.9	68.7	66.0	64.2	49.2	-	77.0	63.5
PromptSRC [30]	-	✓	93.6	90.3	65.7	70.3	86.2	24.0	67.1	46.9	45.5	68.8	65.8	64.4	49.6	-	77.8	63.9
hard prompt	-	-	93.3	88.2	65.6	67.4	85.3	23.7	62.6	44.3	42.0	65.1	63.7	60.9	46.1	66.7	74.0	61.9
POMP (our impl.) [49]	✓	-	94.5	90.9	65.5	71.7	86.1	23.9	66.9	44.3	47.3	66.8	65.8	63.4	48.8	69.9	76.9	64.7
RPP (ours)	✓	-	94.5	90.0	66.3	71.4	85.8	23.1	68.3	45.4	47.9	68.4	66.1	63.7	49.5	70.7	78.0	65.5

**Theorem 1.** Assume that  $\Theta^*$  is the solution to Eq. (11). Then we have that for any  $0 < \delta < 1$  with probability  $1 - \delta$ ,

$$\varepsilon\left(\Theta^{*}\right) - \bar{\varepsilon}_{\chi}\left(\Theta^{*}\right) \leq X^{*}\sqrt{2ln\left(1/\delta\right)/N} + B_{\lambda}^{2}R_{N}\left(\mathcal{L}\right)^{3},\tag{12}$$

where 
$$X^* = \max_{r \in \mathbb{N}_N} |\mathcal{L}\left(\hat{s}_r^S(\Theta), y_r^{gt}\right)|$$
, and  $B_{\lambda} \to 0$  as  $\lambda \to +\infty$ .

In Eq. (12) the first term of the upper bound converges with the increasing of the number of training data N. We can also find that the second term converges to 0 with the increasing of  $\lambda$ , which means the regularizer  $\mathcal{L}_{KD}$  effectively improves the generalization ability of RPP.

# 4 Experiments

The experimental section here presents the extensive results of image classification tasks. For details on ablation studies and experiment configurations, please refer to the *appendix*. It is important to highlight that POMP utilizes the ImageNet-21K winter 21 version, requiring the replication of all comparative pretraining experiments on our dataset (ImageNet-21K fall 11 version)<sup>4</sup>.

#### 4.1 Quantitative Experiments

Prompt Pretraining on ImageNet-21K. The results presented in Tab. 1 showcase the outcomes of pretraining experiments conducted on the ImageNet-21K dataset, demonstrating the superior performance of our proposed method compared to ZeroshotCLIP. Notably, when utilizing ResNet50 and ViT-B/16 as backbones, our method exhibits performance leads of 0.8% and 0.9% compared to POMP, respectively. These advancements result from our focused efforts to mitigate underfitting during prompt pretraining, achieved through a refined prompt structure and optimized supervision.

Table 1: **Performance on the ImageNet-21K validation set.** The backbones of our experiments are ResNet-50 (teacher: ResNet-101) and ViT/B-16 (teacher: ViT/L-14). In the upper tier, ZeroshotCLIP and Prompt Ensemble implement zero-shot inference. "." indicates that the item is empty. "our impl." means our implementation of these methods.

Method	ResNet50	ViT-B/16
ZeroshotCLIP [45]	17	20.7
Prompt Ensemble [45]	18.8	23.5
Linear Probing (our impl.) [45]	5.9	20.3
VPT (our impl.) [11]	_	23.6
POMP (our impl.) [49]	19.4	24
RPP (ours)	20.2(+0.8)	24.9(+0.9)

**Zero-shot Image Classification.** Tab. 2 presents the zero-shot experimental outcomes of our method across ten cross-dataset datasets and four cross-domain datasets. Specifically, Our RPP approach outperforms the POMP method by an average margin of 0.3% and 0.8% in cross-dataset

$$^{2}\text{Here }B_{\lambda}=2\mathbb{E}_{\chi,\mathcal{Z}}\left(\sup_{\Theta\in\mathcal{F}(\lambda)}\bar{\varepsilon}_{\mathcal{Z}}\left(\Theta\right)-\bar{\varepsilon}_{\chi}\left(\Theta\right)\right)/\mathbb{E}_{\chi,\mathcal{Z}}\left(\sup_{\Theta\in\mathbb{R}}\bar{\varepsilon}_{\mathcal{Z}}\left(\Theta\right)-\bar{\varepsilon}_{\chi}\left(\Theta\right)\right)\text{ and }\mathcal{F}\left(\lambda\right)\text{ is a shrinking hypothesis space induced by the regularizer }\mathcal{L}_{KD}\left(\Theta\right).$$

<sup>&</sup>lt;sup>3</sup>Here  $R_N\left(\mathcal{L}\right)$  is the Rademacher complexity of the loss function  $\mathcal{L}$  related to the space  $\mathbb{R}$  for N training examples.

<sup>&</sup>lt;sup>4</sup>The complete winter dataset is not available due to server download. All of our implementation experiments will be marked with "(our impl.)"

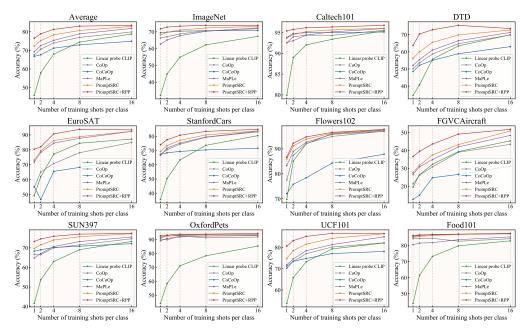


Figure 5: Comparison of RPP performance in a few-shot image recognition scenario. Notably, RPP exhibits superior performance enhancement across various settings, particularly excelling in scenarios characterized by a limited number of shots.

and cross-domain tasks, respectively. Compared with the original CLIP framework, our approach yields improvements of 2.4% and 3.6%.

It is worth noting that when the downstream dataset distribution closely resembles that of ImageNet-21K, our method benefits from the increased fitting capability introduced by our SAPL prompt structure, resulting in improved generalization performance (e.g., in ImageNet-R, our method shows a 1.1% performance boost compared to POMP). Conversely, when the downstream dataset distribution significantly differs from ImageNet-21K, our method further enhances the original CLIP's generalization capability, thanks to our PPKD prompt supervision (e.g., in SUN397, our method exhibits a 1.4% performance improvement compared to POMP).

Few-shot Image Classification. Tab. 3 displays the average few-shot results of our method across eleven classification datasets. It merits particular attention that when applying RPP to other methodologies, we retain the full structure of RPP while fine-tuning only a subset of model parameters required by the downstream methods. Specifically, when employed within the PromptSRC framework, our approach, following the guidelines provided in the PromptSRC paper, exclusively fine-tunes the first nine layers of the learnable prompts, keeping the final three layers fixed. Under this fine-tuning strat-

Table 3: **Few-shot experiment over 11 image classification datasets,** focusing on the average results of all datasets. The backbone is ViT/B-16. For various shot results of each dataset, see Fig. 5.

Method	validation accuracy (%)
Linear Probing [45]	78.79
CoCoOp [67]	74.90
CoOp [68]	79.89
MaPLe [29]	81.79
PromptSRC [30]	82.87
+ POMP (our impl.) [49]	82.56 ( <b>-0.31</b> )
+ RPP (ours)	83.45 ( <b>+0.58</b> )

egy, our method achieves an enhancement of 0.58% over the existing SOTA approach represented by PromptSRC.

Figure 5 illustrates the experimental results from numerous few-shot learning scenarios across eleven datasets. The red lines at the top of each graph represent our method, showcasing robust performance across all datasets under different few-shot fine-tuning conditions. Owing to an enhanced fitting pretraining while retaining the inherent generalization prowess of CLIP, particularly in data-scarce settings like 1-shot and 2-shot, our method, benefiting from comprehensive pretraining, can improve PromptSRC's average performances by 4.19% and 3.62%, respectively. For detailed outcomes, please refer to the *appendix* Sec. D.

Base-to-New Image Classification. Tab. 4 presents the base-to-new experiments of our method across eleven classification datasets. It is noteworthy that since PromptSRC employs a strategy of incorporating learnable prompts at every layer, POMP is limited to initializing only the input token prompts with pretrained weights, while the remaining learnable parameters are subject to random initialization. This culminates in a compromise when applying it to the PromptSRC; it not only diminishes the generalized capabilities afforded by pretraining but also attenuates the fine-tuning efficacy for

Table 4: Comparison with state-of-the-art methods w/ or w/o RPP on base-to-novel generalization. The backbone is ViT/B-16. The "base-to-new" experiments split datasets into base and novel class sets, training on the base and testing on both to assess the model's generalization to less-represented subdomains.

Method	Base (%)	Novel (%)	HM (%)
CLIP [45]	69.34	74.22	71.70
CoCoOp [67]	80.47	71.69	75.83
MaPLe [29]	82.28	75.14	78.55
CoOp [67]	82.69	63.22	71.66
+ POMP (our impl.)	78.51	70.14	74.09 ( <b>+2.43</b> )
+ RPP (ours)	82.19	76.77	79.39 ( <b>+7.73</b> )
PromptSRC [30]	84.26	76.10	79.97
+ POMP (our impl.) [49]	83.48	73.78	78.33 (-1.64)
+ RPP (ours)	85.21	77.38	81.10 ( <b>+1.13</b> )

downstream tasks, which results in a 1.64% decline in HM. For detailed outcomes, please refer to the *appendix* Sec. D.

**Ablation Study.** We assess the model's fitting ability on the validation set of ImageNet-21K and verify its generalization capacity through the average accuracy obtained from cross-dataset and cross-domain evaluations for image classification. Tab. S2 of *appendix* Sec. C introduces the ablation experiments of SAPL and PPKD in detail. SAPL alone improves the model's fitting ability (21K-val +0.9%), while PPKD alone improves the model's generalization ability (Zero-shot +0.2%). Together, SAPL and PPKD increase 21K-val by 0.9% and Zero-shot by 0.43%.

#### 4.2 Confirmatory Experiments

Table 5: **Quantitative experiment on underfitting.** SAPL\_shared denotes a layer-by-layer replaceable shared QKV. To maintain parity in parameter count between configurations employing shared and unshared QKV, we adjust the number of learnable tokens per layer for the shared QKV to be 3 times that of the unshared QKV.

Cause of underfitting	Method	Difference	21K-train (%)	21K-val (%)	Zero-shot (%)
Parameter Quantity	RPP (PPKD)	Learning params (1-layer)	27.0	24.1	65.64
	RPP (SAPL + PPKD)	Learning params (12-layer)	29.3	24.9	65.92
Parameter Diversity	RPP (SAPL_shared + PPKD)	Shared QKV	26.0	23.3	64.82
	RPP (SAPL + PPKD)	Unshared QKV	29.3	24.9	65.92

Analysis for the underfitting. We consider two potential factors that may lead to underfitting: Parameter Quantity and Parameter Diversity. POMP adds unique learnable parameters only at the input layer, which results in both a smaller Parameter Quantity and lower Parameter Diversity. In contrast, our proposed SAPL is designed with a layer-by-layer replaceable QKV unshared prompt structure, which not only increases Parameter Quantity but also enriches Parameter Diversity, thereby enhancing the model's fitting capability. As shown in Tab. 5, we construct two comparative experiments to validate our hypothesis. The results demonstrate that limited Parameter quantity or diversity diminishes the model's fitting ability, further supporting our argument.

**Trade-off of learnable parameters.** The ablation experiment in Tab. 5's Parameter Diversity illustrates that employing unshared QKV enhances fitting and generalization more effectively than merely adding shared learnable parameters. Our innovative method of enabling prompt diversification within layers through unshared QKV in prompt learning transcends a fundamental trade-off between parameter count and model performance.

Computational resource analysis. Table 6 illustrates that the RPP model requires merely an additional 166K training parameters compared to the baseline method, yet it manages to

Computational resource analysis. Table 6: Computational resource experiment. To maintain fairness in the comparison, we consistently set the image size to 224×224. "Time" means training time.

Method	Flops (GMac)	Params (M)	Time (hour)	21K-val (%)	Zero-shot (%)
POMP (our impl.) [49]	18.31	149.628	1.4	24.0	65.49
RPP (ours)	18.71	149.794	1.6	24.9	65.92

achieve similar inference flops and training durations. Moreover, it demonstrates effective performance across a range of tasks. This suggests that the RPP model strikes an efficient balance between computational efficiency and strong performance.

## 5 Conclusion

We are at the forefront of recognizing and tackling underfitting issues during Prompt Pretraining in VLMs, with a focus on refining prompt structure and supervision. Theoretical analysis corroborates the enhanced generalization capability of our method. Our approach signifies a paradigm shift in prompt learning research, emphasizing the importance of robust initialization strategies. Nevertheless, two challenges persist: 1) Insufficient exploration of the hierarchical category data available in ImageNet-21K. 2) Constraints imposed by the form of supervised pretraining and label limitations impede the acquisition of generalized knowledge. These aspects warrant further investigation in future research endeavors.

# Acknowledgement

This research was supported by the Young Scientists Fund of the National Natural Science Foundation of China (Grant No.62206134), the Fundamental Research Funds for the Central Universities 070-63233084, and the Tianjin Key Laboratory of Visual Computing and Intelligent Perception (VCIP). Computation is supported by the Supercomputing Center of Nankai University (NKSC).

#### References

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, volume 35, pages 23716–23736, 2022. 1
- [2] Alex Andonian, Shixing Chen, and Raffay Hamid. Robust cross-modal representation learning with progressive self-distillation. In *CVPR*, pages 16430–16441, 2022. 3
- [3] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 2022. 3
- [4] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 mining discriminative components with random forests. In *ECCV*, 2014. 16
- [5] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Prompt learning with optimal transport for vision-language models. In *ICLR*, 2023. 1
- [6] Shuo Chen, Chen Gong, Xiang Li, Jian Yang, Gang Niu, and Masashi Sugiyama. Boundary-restricted metric learning. *Machine Learning*, 112(12):4723–4762, 2023. 2
- [7] Shuo Chen, Lei Luo, Jian Yang, Chen Gong, Jun Li, and Heng Huang. Curvilinear distance metric learning. In *NeurIPS*, volume 32, 2019. 2
- [8] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, pages 3606–3613, 2014. 16
- [9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 16
- [10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE, 2009. 2, 16
- [11] Mohammad Mahdi Derakhshani, Enrique Sanchez, Adrian Bulat, Victor Costa, Cees G. M. Snoek, Georgios Tzimiropoulos, and Brais Martínez. Variational prompt tuning improves generalization of vision-language models. *arXiv*, abs/2210.02390, 2022. 7
- [12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 5
- [13] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *CVPR*, pages 19358–19369, 2023. 1, 3

- [14] Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lijuan Wang, Yezhou Yang, and Zicheng Liu. Compressing visual-linguistic model via knowledge distillation. In *ICCV*, pages 1428–1438, 2021. 3
- [15] 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 CVPR Workshop, pages 178–178, 2004. 16
- [16] Chengjian Feng, Yujie Zhong, Zequn Jie, Xiangxiang Chu, Haibing Ren, Xiaolin Wei, Weidi Xie, and Lin Ma. Promptdet: Towards open-vocabulary detection using uncurated images. In *ECCV*, pages 701–717. Springer, 2022. 3
- [17] 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. *IJCV*, pages 1–15, 2023. 1
- [18] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2022. 1, 3
- [19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 1
- [20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, pages 770–778. IEEE, 2016. 5
- [21] Patrick Helber, Benjamin Bischke, Andreas R. Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE J-STARS*, 12:2217–2226, 2019. 16
- [22] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Lixuan Zhu, Samyak Parajuli, Mike Guo, Dawn Xiaodong Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, pages 8320–8329, 2020. 16
- [23] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. A comprehensive overhaul of feature distillation. In *ICCV*, pages 1921–1930, 2019. 3
- [24] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 3
- [25] Qidong Huang, Xiaoyi Dong, Dongdong Chen, Weiming Zhang, Feifei Wang, Gang Hua, and Nenghai Yu. Diversity-aware meta visual prompting. In *CVPR*, pages 10878–10887, 2023. 3
- [26] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, pages 709–727. Springer, 2022. 3, 4, 5
- [27] Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. How can we know what language models know? *TACL*, 8:423–438, 2020. 3
- [28] Siyu Jiao, Yunchao Wei, Yaowei Wang, Yao Zhao, and Humphrey Shi. Learning mask-aware clip representations for zero-shot segmentation. *arXiv preprint arXiv:2310.00240*, 2023. 3
- [29] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *CVPR*, pages 19113–19122, 2023. 1, 3, 4, 5, 8, 9
- [30] 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 *ICCV*, pages 15190–15200, 2023. 1, 2, 3, 7, 8, 9, 16, 17
- [31] Steven G Krantz and Harold R Parks. *A primer of real analytic functions*. Springer Science & Business Media, 2002. 15
- [32] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *CVPR Workshop*, pages 554–561, 2013. 16

- [33] Clement Laroudie, Andrei Bursuc, Mai Lan Ha, and Gianni Franchi. Improving clip robustness with knowledge distillation and self-training. *arXiv preprint arXiv:2309.10361*, 2023. 3
- [34] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *EMNLP*, pages 3045–3059. ACL, 2021. 3
- [35] 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 *ICML*, volume 202, pages 19730–19742. PMLR, 2023. 1
- [36] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, volume 34, pages 9694–9705, 2021.
- [37] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *CVPR*, pages 10965–10975, 2022. 1
- [38] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *ACL/IJCNLP*, pages 4582–4597. ACL, 2021. 3
- [39] Zheng Li, Xiang Li, Xinyi Fu, Xing Zhang, Weiqiang Wang, and Jian Yang. Promptkd: Unsupervised prompt distillation for vision-language models. In *CVPR*, 2024. 3
- [40] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *CVPR*, pages 7061–7070, 2023. 1
- [41] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv* preprint arXiv:2304.08485, 2023. 1
- [42] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv*, abs/1306.5151, 2013. 16
- [43] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. *Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729, 2008. 16
- [44] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In CVPR, pages 3498–3505. IEEE, 2012. 16
- [45] 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 *ICML*, pages 8748–8763. PMLR, 2021. 1, 3, 4, 6, 7, 8, 9
- [46] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *CVPR*, pages 18082–18091, 2022. 3
- [47] Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Fine-tuned clip models are efficient video learners. In *CVPR*, pages 6545–6554, 2023. 3
- [48] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, 2019. 16
- [49] Shuhuai Ren, Aston Zhang, Yi Zhu, Shuai Zhang, Shuai Zheng, Mu Li, Alex Smola, and Xu Sun. Prompt pre-training with twenty-thousand classes for open-vocabulary visual recognition. In *NeurIPS*, 2023. 1, 2, 3, 6, 7, 8, 9, 17
- [50] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. In *NeurIPS Datasets and Benchmarks*, 2021. 16, 17

- [51] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *EMNLP*, pages 4222–4235. ACL, 2020. 3
- [52] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv*, abs/1212.0402, 2012. 16
- [53] Roman Vershynin. High-dimensional probability: An introduction with applications in data science, volume 47. Cambridge university press, 2018. 14
- [54] Haohan Wang, Songwei Ge, Eric P. Xing, and Zachary Chase Lipton. Learning robust global representations by penalizing local predictive power. In NIPS, 2019. 16
- [55] Zhecan Wang, Noel Codella, Yen-Chun Chen, Luowei Zhou, Jianwei Yang, Xiyang Dai, Bin Xiao, Haoxuan You, Shih-Fu Chang, and Lu Yuan. Clip-td: Clip targeted distillation for vision-language tasks. *arXiv preprint arXiv:2201.05729*, 2022. 3
- [56] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In CVPR, pages 139–149, 2022. 3
- [57] Longhui Wei, Lingxi Xie, Wengang Zhou, Houqiang Li, and Qi Tian. Mvp: Multimodality-guided visual pre-training. In *ECCV*, pages 337–353. Springer, 2022. 3
- [58] Kan Wu, Houwen Peng, Zhenghong Zhou, Bin Xiao, Mengchen Liu, Lu Yuan, Hong Xuan, Michael Valenzuela, Xi Stephen Chen, Xinggang Wang, et al. Tinyclip: Clip distillation via affinity mimicking and weight inheritance. In *ICCV*, pages 21970–21980, 2023. 3
- [59] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. Clipself: Vision transformer distills itself for open-vocabulary dense prediction. *arXiv* preprint arXiv:2310.01403, 2023. 3
- [60] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In CVPR, pages 3485–3492, 2010.
  16
- [61] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In CVPR, pages 9653–9663, 2022.
- [62] Chuanguang Yang, Zhulin An, Libo Huang, Junyu Bi, Xinqiang Yu, Han Yang, and Yongjun Xu. Clip-kd: An empirical study of distilling clip models. *arXiv preprint arXiv:2307.12732*, 2023. 3
- [63] Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Unified vision and language prompt learning. *arXiv preprint arXiv:2210.07225*, 2022. 3, 5
- [64] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In CVPR, pages 16793–16803, 2022. 3
- [65] Zexuan Zhong, Dan Friedman, and Danqi Chen. Factual probing is [mask]: Learning vs. learning to recall. In *NAACL-HLT*, pages 5017–5033. ACL, 2021. 3
- [66] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *ECCV*, pages 696–712. Springer, 2022. 1, 3
- [67] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16816–16825, 2022. 1, 3, 7, 8, 9
- [68] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022. 1, 3, 4, 7, 8, 17

# **Appendix**

In the appendix, we will provide detailed descriptions of formula proofs, parameter settings for all experiments, datasets, and evaluation metrics. We will include comprehensive ablation studies and comparative experiments, and finally, analyze the broader impacts and safeguards.

## **A** Theoretical Proof

This section provides detailed proofs for the Theorem in Sec. 3.5. We introduce the following lemmas for proving our Theorem.

**Lemma 1**(McDiarmid's Inequality [53]). Consider independent random variables  $v_1, v_2, \dots, v_n \in \mathcal{V}$  and a function  $\phi : \mathcal{V}^n \to \mathbb{R}$ . Suppose that for all  $v_1, v_2, \dots, v_n$  and  $v_i' \in \mathcal{V}$   $(i = 1, 2, \dots, n)$ , the function satisfies

$$|\phi(v_1, \dots, V_{i-1}, V_i, V_{i+1}, \dots, V_n) - \phi(v_1, \dots, V_{i-1}, v_i', V_{i+1}, \dots, V_n)| \le c_i,$$
 (S1)

and then it holds that

$$\mathcal{P}\left\{\phi\left(v_{1}, v_{2}, \cdots, v_{n}\right) - \mathbb{E}_{v_{1}, v_{2}, \cdots, v_{n}}\left(\phi\left(v_{1}, v_{2}, \cdots, v_{n}\right)\right) > \mu\right\} \leq e^{-\frac{2\mu^{2}}{\sum_{i=1}^{n} c_{i}^{2}}}.$$
 (S2)

**Lemma 2.** Let  $\Theta^*$  be the solution to the optimization objective

$$\Theta^* \in \min_{\Theta \in \mathbb{R}^d} \frac{1}{N} \sum_{i=1}^N \mathcal{L}\left(\hat{s}_i^S\left(\Theta\right), y_i^{gt}\right) + \lambda \frac{1}{N \cdot K} \sum_{k=1}^K \sum_{i=1}^N \hat{s}_{ik}^T \log \frac{\hat{s}_{ik}^T}{\hat{s}_{ik}^S\left(\Theta\right)}, \tag{S3}$$

then there exists a bounded tensor set  $\mathcal{F}(\lambda)$  such that

$$\Theta^* \in \mathcal{F}(\lambda) = \left\{ \Theta \middle| e^{\frac{C_0}{\lambda s_{ik}^T}} \le s_{ik}^S(\Theta) \le 1, i \in \mathbb{N}_N, k \in \mathbb{N}_K \right\},\tag{S4}$$

where the constant  $C_0 > 0$  is not dependent on  $\lambda$ .

*Proof.* According to the optimality of  $\Theta^*$ , it follows that

$$\frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(\hat{s}_{i}^{S}\left(\Theta^{*}\right), y_{i}^{gt}\right) + \lambda \frac{1}{N \cdot K} \sum_{k=1}^{K} \sum_{i=1}^{N} \hat{s}_{ik}^{T} \log \frac{\hat{s}_{ik}^{T}}{\hat{s}_{ik}^{S}\left(\Theta\right)} \\
\leq \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(\hat{s}_{i}^{T}, y_{i}^{gt}\right) + \lambda \frac{1}{N \cdot K} \sum_{k=1}^{K} \sum_{i=1}^{N} \hat{s}_{ik}^{T} \log \frac{\hat{s}_{ik}^{T}}{\hat{s}_{ik}^{T}} \\
\leq \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(\hat{s}_{i}^{T}, y_{i}^{gt}\right).$$
(S5)

We denote that  $\mathcal{L}_{min} = \inf_{\Theta^*, i=1, 2, \cdots, N} \mathcal{L}\left(\hat{s}_i^S\left(\Theta^*\right), y_i^{gt}\right)$ , and have that

$$\lambda \frac{1}{N \cdot K} \sum_{k=1}^{K} \sum_{i=1}^{N} \hat{s}_{ik}^{T} \log \frac{\hat{s}_{ik}^{T}}{\hat{s}_{ik}^{S}(\Theta)}$$

$$\Leftrightarrow -\lambda \frac{1}{N \cdot K} \sum_{k=1}^{K} \sum_{i=1}^{N} \hat{s}_{ik}^{T} \log \hat{s}_{ik}^{S}(\Theta)$$

$$\leq \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(\hat{s}_{i}^{T}, y_{i}^{gt}\right) - \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(\hat{s}_{i}^{S}(\Theta^{*}), y_{i}^{gt}\right)$$

$$\leq \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(\hat{s}_{i}^{T}, y_{i}^{gt}\right) - \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{min}$$

$$= C_{0}, \tag{S6}$$

where  $C_0 > 0$ . Finally, we have

$$e^{\frac{C_0}{\lambda s_{ik}^T}} \le s_{ik}^S(\Theta) \le 1,\tag{S7}$$

which completes the proof.

The proof of Theorem 1. is given as follows.

**Theorem 1.** Assume that  $\Theta^*$  is the solution to RPP. Then we have that for any  $0 < \delta < 1$  with probability  $1 - \delta$ ,

$$\varepsilon\left(\Theta^{*}\right) - \bar{\varepsilon}_{\chi}\left(\Theta^{*}\right) \le X^{*}\sqrt{2ln\left(1/\delta\right)/N} + B_{\lambda}R_{N}\left(\mathcal{L}\right),$$
 (S8)

where  $B_{\lambda} \to 0$  as  $\lambda \to +\infty$ , and  $X^* = \max_{r \in \mathbb{N}_N} |\mathcal{L}(\hat{s}_r^s(\Theta), y_r^{gt})|$ . Here  $R_N(\mathcal{L})$  is the Rademacher complexity of the loss function  $\mathcal{L}$  related to the space  $\mathbb{R}$  for N training examples.

*Proof.* Firstly, we denote that

$$\varepsilon\left(\Theta^{*}\right) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(\hat{s}_{i}^{S}\left(\Theta\right), y_{i}^{gt}\right),\tag{S9}$$

and

$$\bar{\varepsilon}_{\chi,r}\left(\Theta^{*}\right) = \frac{1}{N} \left( \sum_{i=1, i \neq r}^{N} \mathcal{L}\left(\hat{s}_{i}^{S}\left(\Theta\right), y_{i}^{gt}\right) + \mathcal{L}\left(\hat{a}_{r}^{S}\left(\Theta\right), b_{r}^{gt}\right) \right), \tag{S10}$$

where  $(\hat{a}_k^S(\Theta), b_k^{gt})$  is an arbitrary data pair from the sample space with similarity label  $b_k^{gt}$ . Then we have that

$$\begin{aligned} |\varepsilon\left(\Theta^{*}\right) - \bar{\varepsilon}_{\chi,r}\left(\Theta^{*}\right)| &= \frac{1}{N} \left| \mathcal{L}\left(\hat{s}_{r}^{S}\left(\Theta\right), y_{r}^{gt}\right) - \mathcal{L}\left(\hat{a}_{r}^{S}\left(\Theta\right), b_{r}^{gt}\right) \right| \\ &\leq \frac{1}{N} \left( \left| \mathcal{L}\left(\hat{s}_{r}^{S}\left(\Theta\right), y_{r}^{gt}\right) \right| + \left| \mathcal{L}\left(\hat{a}_{r}^{S}\left(\Theta\right), b_{r}^{gt}\right) \right| \right) \\ &\leq \frac{2}{N} X^{*}, \end{aligned} \tag{S11}$$

where  $X^* = \max_{r \in \mathbb{N}_N} \left| \mathcal{L}\left(\hat{s}_r^S\left(\Theta\right), y_r^{gt}\right) \right|$ . Then we apply Lemma 2 to the term  $\varepsilon\left(\Theta^*\right) - \bar{\varepsilon}_{\chi}\left(\Theta^*\right)$  and have that with probability  $1 - \delta$  it holds that

$$\varepsilon\left(\Theta^{*}\right) - \bar{\varepsilon}_{\chi}\left(\Theta^{*}\right) \leq \mathbb{E}_{\chi}\left(\varepsilon\left(\Theta^{*}\right) - \bar{\varepsilon}_{\chi}\left(\Theta^{*}\right)\right) + X^{*}\sqrt{2\ln\left(1/\delta\right)/N}.\tag{S12}$$

Now we only need to estimate the first term of the right-hand side of the above inequality. Specifically, there holds

$$\mathbb{E}_{\chi}\left(\varepsilon\left(\Theta^{*}\right) - \bar{\varepsilon}_{\chi}\left(\Theta^{*}\right)\right) = \mathbb{E}_{\chi}\left(\mathbb{E}_{\mathcal{Z}}\varepsilon_{\mathcal{Z}}\left(\Theta^{*}\right) - \bar{\varepsilon}_{\chi}\left(\Theta^{*}\right)\right) \leq \mathbb{E}_{\chi,\mathcal{Z}}\left(\varepsilon_{\mathcal{Z}}\left(\Theta^{*}\right) - \bar{\varepsilon}_{\chi}\left(\Theta^{*}\right)\right), \quad (S13)$$

where  $\mathcal{Z} = \{z_i | z_i \sim \mathcal{D}, i \in \mathbb{N}_N\}$  are independent identically distributed (i.i.d.) samples which are independent of  $\chi = \{x_i | x_i \sim \mathcal{D}, i \in \mathbb{N}_N\}$ . By Lemma 2, we know that there exists the bounded tensor set  $\mathcal{F}(\lambda)$  such that

$$\Theta^* \in \mathcal{F}(\lambda) = \left\{ \Theta | e^{\frac{C_0}{\lambda s_{ik}^T}} \le s_{ik}^S(\Theta) \le 1, i \in \mathbb{N}_N, k \in \mathbb{N}_K \right\}, \tag{S14}$$

where  $C_0 > 0$  is a constant. Let the function

$$B_{\lambda} = 2\mathbb{E}_{\chi,\mathcal{Z}} \left( \sup_{\Theta \in \mathcal{F}(\lambda)} \bar{\varepsilon}_{\mathcal{Z}} \left( \Theta \right) - \bar{\varepsilon}_{\chi} \left( \Theta \right) \right) / \mathbb{E}_{\chi,\mathcal{Z}} \left( \sup_{\Theta \in \mathbb{R}} \bar{\varepsilon}_{\mathcal{Z}} \left( \Theta \right) - \bar{\varepsilon}_{\chi} \left( \Theta \right) \right). \tag{S15}$$

By Levi's Monotone Convergence Theorem [31], we have

$$\lim_{\lambda \to \infty} \mathbb{E}_{\chi, \mathcal{Z}} \left( \sup_{\Theta \in \mathcal{F}(\lambda)} \bar{\varepsilon}_{\mathcal{Z}}(\Theta) - \bar{\varepsilon}_{\chi}(\Theta) \right)$$

$$= \mathbb{E}_{\chi, \mathcal{Z}} \left( \lim_{\lambda \to \infty} \sup_{\Theta \in \mathcal{F}(\lambda)} \bar{\varepsilon}_{\mathcal{Z}}(\Theta) - \sup_{\Theta \in \mathcal{F}(\lambda)} \bar{\varepsilon}_{\chi}(\Theta) \right)$$

$$= \mathbb{E}_{\chi, \mathcal{Z}} \left( \bar{\varepsilon}_{\mathcal{Z}}(0) - \bar{\varepsilon}_{\chi}(0) \right)$$

$$= \mathbb{E}_{\mathcal{Z}} \left( \bar{\varepsilon}_{\mathcal{Z}}(0) \right) - \mathbb{E}_{\chi, (\bar{\varepsilon}_{\chi}(0))} = 0.$$
(S16)

Therefore, we obtain  $\lim_{\lambda\to\infty} B_{\lambda} = 0$ . By standard symmetrization techniques for i.i.d. Rademacher variables  $\sigma = (\sigma_1, \sigma_2, \cdots, \sigma_N)^{\mathsf{T}}$ , it follows that

$$\mathbb{E}_{\chi,\mathcal{Z}}\left(\bar{\varepsilon}_{\mathcal{Z}}\left(\Theta^{*}\right) - \bar{\varepsilon}_{\chi}\left(\Theta^{*}\right)\right) \leq \mathbb{E}_{\chi,\mathcal{Z}}\left(\sup_{\Theta \in \mathcal{F}(\lambda)} \bar{\varepsilon}_{\mathcal{Z}}\left(\Theta\right) - \bar{\varepsilon}_{\chi}\left(\Theta\right)\right) \\
= \frac{B_{\lambda}}{2} \mathbb{E}_{\chi,\mathcal{Z}}\left(\sup_{\Theta \in \mathbb{R}} \bar{\varepsilon}_{\mathcal{Z}}\left(\Theta\right) - \bar{\varepsilon}_{\chi}\left(\Theta\right)\right) \\
= \frac{B_{\lambda}}{2N} \mathbb{E}_{\chi,\mathcal{Z},\sigma}\left(\sup_{\Theta^{*} \in \mathbb{R}} \sum_{i=1}^{N} \sigma_{i}\left(\mathcal{L}\left(s_{i}^{S}\left(\Theta\right)\right) - \mathcal{L}\left(z_{i}^{S}\left(\Theta\right)\right)\right)\right) \\
= \frac{B_{\lambda}}{N} \mathbb{E}_{\chi,\sigma}\left(\sup_{\Theta^{*} \in \mathbb{R}} \sum_{i=1}^{N} \sigma_{i}\mathcal{L}\left(s_{i}^{S}\left(\Theta\right)\right)\right) \\
= B_{\lambda}R_{N}\left(\mathcal{L}\right), \tag{S17}$$

where  $P\{\sigma_i = 0\} = P\{\sigma_i = 1\} = 0.5$  for  $i \in \mathbb{N}_N$ , and  $R_N(\mathcal{L})$  is the Rademacher complexity of L. Finally, combining the above inequality with Eqn. (S12) and Eqn. (S13) completes the proof.  $\square$ 

# **B** Training details

#### **B.1** Text templates and Datasets

Table S1: Datasets in our experiments. Text template utilized by senior CLIP teacher for different datasets.

Dataset	Text template	Classes	Train Size	Test Size	Metric
Prompt Pretraining ImageNet-21K [10]	" a photo of a [class]. "	11,221	12,358,688	561,050	accuracy
Datasets of Image Classification					
Caltech-101 [15]	" a photo of a [class]. "	102	3,060	6,086	mean per-class accuracy
Oxford-Pets [44]	" a photo of a [class], a type of pet. "	37	3,680	3,669	mean per-class accuracy
Stanford Cars [32]	" a photo of a [class]. "	196	8,144	8,041	accuracy
Oxford Flowers-102 [43]	" a photo of a [class], a type of flower. "	102	2,040	6,149	mean per-class accuracy
Food-101 [4]	" a photo of [class], a type of food. "	101	75,750	25,250	accuracy
FGVC Aircraft [42]	" a photo of a [class], a type of aircraft. "	100	6,667	3,333	mean per-class accuracy
SUN-397 [60]	" a photo of a [class]. "	397	15,880	19,850	accuracy
Describable Textures (DTD) [8]	" [class] texture. "	47	3,760	1,880	accuracy
EuroSAT [21]	" a centered satellite photo of [class]. "	10	10,000	5,000	accuracy
UCF-101 [52]	" a photo of a person doing [class]. "	101	7,639	3,783	accuracy
ImageNet [9]	" a photo of a [class]. "	1000	1,281,167	50,000	accuracy
ImageNetV2 [48]	" a photo of a [class]. "	1,000	10,000	10,000	accuracy
ImageNet-S [54]	" a photo of a [class]. "	1,000	50,889	50,889	accuracy
ImageNet-R [22]	" a photo of a [class]. "	200	30,000	30,000	accuracy

**Text templates for senior teacher CLIP** According to prior research [30], different prompt templates are employed for various datasets to enhance the text representation capability of the senior CLIP and to improve the distillation effect of the boosting prompts. Specifically, the corresponding template for each dataset is shown in Tab. S1.

**Prompt Pretraining.** Following the conventional ImageNet-21K pretraining approach [50], we undertake the following three processes on the dataset: (1) Invalid class filtering: To mitigate the influence of an extremely long-tail distribution on experimental outcomes, classes with fewer than 500 images are excluded. Consequently, from the fall 11 release, the dataset comprises 12,358,688 images spanning 11,221 classes. (It's noteworthy that POMP utilizes the winter 21 version, necessitating all comparison experiments with POMP to be replicated on our dataset). (2) Creation of a validation set: For standardization and future benchmarking purposes, we allocate 50 images per class for a uniform validation split. (3) Image resizing: To facilitate accessibility and expedite training, all images within the ImageNet-21K dataset are resized to a resolution of 224 during the preprocessing phase. All methods train on the full ImageNet-21K dataset (fall 11 version). The specifics of the pretraining dataset, ImageNet-21K, utilized for Prompt Pretraining, are presented in Tab. S1.

It's imperative to note that the results shown are replicated on our proprietary dataset due to discrepancies with the dataset employed by POMP. Furthermore, our advocated SAPL prompt structure necessitates tailored modifications to align with the ViT architecture. Consequently, when employing ResNet as the backbone, this innovative structure is exclusively integrated within the text encoder.

Table S2: Ablation Study Results. By default, ablation experiments employ ViT-B/16 as the backbone, with ViT-L/14 serving as the Teacher model. All ablation studies are conducted using the same GPUs. "Nctx" means the number of prompts that can be learned per layer. "Pretrained val." means ImageNet-21K validation Top-1 accuracy. "Zero-shot" means average accuracy on cross-dataset and cross-domain evaluation for image classification. "-" indicates that the item is empty. We performed ablation experiments on ImageNet-21K and 14 zero-shot datasets.

Method	SAPL	PPKD	Nctx	PPKD temperature	PPKD loss-weight	Pretrained val. (%)	Zero-shot (%)
	-	-	16	-	-	24.0	65.49
RPP	$\checkmark$	-	4	1.0	0.0	24.9 ↑	64.03 ↓
KFF	-	$\checkmark$	4	1.0	1.0	24.1 -	65.64 ↑
	$\checkmark$	$\checkmark$	4	1.0	1.0	24.0 24.9 ↑	65.70 ↑
			8	1.0	1.0	25.4	65.59
			16	1.0	1.0	25.5	65.16
			32	1.0	1.0	25.6	65.17
			4	0.1	1.0	25.0	64.76
			4	0.2	1.0	25.1	64.66
			4	2.0	1.0	24.9	65.66
A 1-1-4: C41	,	,	4	4.0	1.0	24.7	65.27
Ablation Study	✓	✓	4	8.0	1.0	24.6	64.96
			4	1.0	0.1	25.1	64.54
			4	1.0	0.5	25.2	65.28
			4	1.0	1.0	25.2	65.70
			4	1.0	1.5	25.0	65.41
			4	1.0	2.0	24.9	65.92
			4	1.0	4.0	24.7	65.41

**Image Classification.** For zero/few-shot and base-to-new image classification, we evaluate the performance of PPKD on 11 downstream datasets, including Caltech-101, Oxford-Pets, Stanford Cars, Oxford-Flowers102, Food-101, FGVC Aircraft, EuroSAT, SUN-397, Describable Textures (DTD), UCF-101, ImageNet. We also conduct zero-shot evaluation on 3 out-of-domain datasets including ImageNetV2 ImageNet-S, and ImageNet-R, to evaluate the domain generalization capability of our method. The specifics of the downstream datasets utilized for classification are presented in Tab. S1.

#### **B.2** Task Setups and Implementation Details

# **B.2.1 Pretraining Details**

Our experiments are conducted utilizing  $8 \times \text{Nvidia}$  A40 GPUs. For the pretraining phase, we employ the proposed ImageNet-21K fall 11 version [50]. Following POMP [49], each class is represented by 16 training samples (16-shot), and the prompt length is set to 4. The batch size used is 32, and the maximum epoch is limited to 5. At each training step, we sample 1,000 classes, denoted as K=1000. We employ the SGD optimizer with an initial learning rate (lr) of 0.016, decay according to the cosine annealing rule.

#### **B.2.2** Setting for Image Classification

Our experiments are conducted utilizing 8×Nvidia A40 GPUs. We perform an extensive comparative experimental analysis of our proposed method in comparison with the prior SOTA POMP approach. This evaluation involves utilizing the PromptSRC [30] and CoOp[68] frameworks within the base-to-new experimental setting.

In the few-shot experiments, we utilize a standardized experimental setup, maintaining a fixed batch size of 4 and conducting a maximum of 50 epochs. Specifically within the PromptSRC framework, the lr is set at 0.02. However, in the base-to-new experiments, we reduce the maximum number of epochs from 50 to 20. In the PromptSRC framework, the lr remained at 0.02, while it is adjusted to 0.036 within the CoOp framework.

Our novel prompt structure necessitates a tailored approach to model weight initialization, which may involve fine-tuning to achieve compatibility with established training protocols or selectively merging new structural elements with extant pretrained weights.

# C Ablation study

We initially conduct ablation studies focusing on two aspects: prompt structure (SAPL) and prompt supervision (PPKD). According to the data presented in the second and third rows of Tab. S2, it is

Table S3: The complete experiment from base-to-novel. Comparison with state-of-the-art methods w/ or w/o RPP on base-to-novel generalization. Our RPP consistently improves baseline model performance on 11 datasets.

Method	Base (%)	Novel (%)		Method	Base (%)	Novel (%)	HM (%)	Method	Base (%)		
CLIP	69.34 80.47	74.22 71.69	71.70 75.83	CLIP	72.43 75.98	68.14 70.43	70.22 73.10	CLIP	96.84	94.00	95.40
CoCoOp MaPLe	82.28	75.14	78.55	CoCoOp MaPLe	76.66	70.43	73.47	CoCoOp MaPLe	97.96 97.74	93.81 94.36	95.84 96.02
	82.69	63.22	71.66			67.88	71.92		98.00	89.81	93.73
CoOp + POMP (our impl.)	78.51	70.14	74.09 (+2.43)	CoOp + POMP (our impl.)	76.47 74.43	67.88	70.55 (-1.37)	CoOp + POMP (our impl.)	98.00 97.93	93.27	95.73 95.54 (+1.81
+ RPP (ours)	82.19	76.77	79.39 (+7.73)	+ RPP (ours)	76.77	72.50	74.57 (+2.65)	+ RPP (ours)	98.67	95.33	96.97 (+3.24
PromptSRC	84.26	76.10	79.97	PromptSRC	77.60	70.73	74.01	PromptSRC	98.10	94.03	96.02
+ POMP (our impl.)	83.48	73.78	78.33 (-1.64)	+ POMP (our impl.)	77.53	70.47	73.83 (-0.18)	+ POMP (our impl.)	98.40	95.37	96.86 (+0.84
+ RPP (ours)	85.21	77.38	81.10 (+1.13)	+ RPP (ours)	77.93	72.50	75.12 (+1.11)	+ RPP (ours)	98.67	95.00	96.80 (+0.78
(a) Av	erage ove	r 11 datase	ets.		(b) Ima	geNet			(c) Calte	ch101	
Method	Base (%)	Novel (%)	HM (%)	Method	Base (%)	Novel (%)	HM (%)	Method	Base (%)	Novel (%)	HM (%)
CLIP	91.17	97.26	94.12	CLIP	63.37	74.89	68.65	CLIP	72.08	77.80	74.83
CoCoOp	95.20	97.69	96.43	CoCoOp	70.49	73.59	72.01	CoCoOp	94.87	71.75	81.71
MaPLe	95.43	97.76	96.58	MaPLe	72.94	74.00	73.47	MaPLe	95.92	72.46	82.56
CoOp	93.67	95.29	94.47	CoOp	78.12	60.40	68.13	CoOp	97.60	59.67	74.06
+ POMP (our impl.) + RPP (ours)	95.57 96.17	97.40 97.77	96.47 (+2.00) 96.96 (+2.49)	+ POMP (our impl.) + RPP (ours)	71.27 75.13	71.90 75.77	71.58 (+3.45) 75.45 (+7.32)	+ POMP (our impl.) + RPP (ours)	93.43 96.67	72.60 77.60	81.71 (+7.65) 86.09 (+12.03)
PromptSRC + POMP (our impl.)	95.33 96.43	97.30 97.43	96.30 96.93 (+0.63)	PromptSRC + POMP (our impl.)	78.27 77.07	74.97 74.30	76.58 75.66 (-0.92)	PromptSRC + POMP (our impl.)	98.07 97.37	76.50 73.67	85.95 83.87 (-2.08)
+ RPP (ours)	96.37	97.93	97.14 (+0.84)	+ RPP (ours)	80.87	75.83	78.27 (+1.69)	+ RPP (ours)	98.00	77.73	86.70 (+0.75)
Method	Base (%)	Novel (%)	HM (%)	Method	Base (%)	Novel (%)	HM (%)	Method	Base (%)	Novel (%)	HM (%)
CLIP	90.10	91.22	90,66	CLIP	27.19	36.29	31.09	CLIP	69.36	75.35	72.23
CoCoOp	90.70	91.29	90.99	CoCoOp	33.41	23.71	27.74	CoCoOp	79.74	76.86	78.27
MaPLe	90.71	92.05	91.38	MaPLe	37.44	35.61	36.50	MaPLe	80.82	78.70	79.75
CoOp	88.33	82.26	85.19	CoOp	40.44	22.30	28.75	CoOp	80.60	65.89	72.51
+ POMP (our impl.) + RPP (ours)	89.73 90.07	90.87 91.53	90.30 (+5.11) 90.79 (+5.60)	+ POMP (our impl.) + RPP (ours)	21.43 37.07	23.77 36.10	22.54 (-6.21) 36.58 (+7.83)	+ POMP (our impl.) + RPP (ours)	79.70 81.63	74.80 79.20	77.17 (+4.66) 80.40 (+7.89)
PromptSRC + POMP (our impl.)	90.67 90.50	91.53 91.67	91.10 91.08 (-0.02)	PromptSRC + POMP (our impl.)	42.73 33.90	37.87 28.63	40.15 31.04 (-9.11)	PromptSRC + POMP (our impl.)	82.67 82.37	78.47 78.53	80.52 80.40 (-0.12)
+ RPP (ours)	90.70	92.07	91.38 (+0.28)	+ RPP (ours)	43.73	38.20	40.78 (+0.63)	+ RPP (ours)	82.87	79.20	80.99 (+0.47)
	(g) Foo	d101		(	h) FGVC	Aircraft			(i) SUN	1397	
Method	Base (%)	Novel (%)	HM (%)	Method	Base (%)	Novel (%)	HM (%)	Method	Base (%)	Novel (%)	HM (%)
CLIP	53.24	59.90	56.37	CLIP	56.48	64.05	60.03	CLIP	70.53	77.50	73.85
CoCoOp MaPLe	77.01 80.36	56.00 59.18	64.85 68.16	CoCoOp MaPLe	87.49 94.07	60.04 73.23	71.21 82.35	CoCoOp MaPLe	82.33 83.00	73.45 78.66	77.64 80.77
	79.44	41.18	54.24	CoOp	92.19	54.74	68.69	CoOp	84.69	56.05	67.46
CoOp + POMP (our impl.)	76.97	44.90	56.71 (+2.47)	+ POMP (our impl.)	80.83	63.10	70.87 (+2.18)	+ POMP (our impl.)	82.27	71.87	76.72 (+9.26)
+ RPP (ours)	81.23	59.37	68.60 (+14.36)	+ RPP (ours)	86.20	81.37	83.71 (+15.02)	+ RPP (ours)	84.50	77.97	81.10 (+13.64)
PromptSRC	83.37	62.97	71.75	PromptSRC	92.90	73.90	82.32	PromptSRC	87.10	78.80	82.74
+ POMP (our impl.)	83.57	50.63	63.06 (-8.69)	+ POMP (our impl.)	94.27	72.97	82.26 (-0.06)	+ POMP (our impl.)	86.90	77.90	82.15 (-0.59)
+ RPP (ours)	84.47	64.73	72.65 (+0.90)	+ RPP (ours)	96.27	79.37	87.01 (+4.69)	+ RPP (ours)	87.43	79.57	83.31 (+0.57)
	(j) DT	D			(k) Euro	SAT			(l) UCF	101	

evident that both SAPL and PPKD methodologies are capable of enhancing the pretraining model's fitting ability, as reflected by an increase in Pretrained validation accuracy. However, in alignment with the analysis discussed in the main body, excessive fitting may lead to a diminishment in the model's generalization capabilities inherited from CLIP, as indicated by a 1.46% decrease in zero-shot performance reported in the second row. To mitigate this issue, we additionally utilize soft labels derived from zero-shot probability predictions provided by a large-scale CLIP teacher model. As shown in the fourth row, when SAPL and PPKD are employed in conjunction, there is an observed improvement of 1.2% in our pretraining validation set accuracy, along with a 0.21% augmentation in the average precision of zero-shot tasks.

Subsequently, according to the "Ablation Study" of Tab. S2, we undertake ablation experiments related to hyperparameters. Based on these results, we discern that indiscriminately augmenting model parameters results in a profound loss of generalization. Consequently, we ascertain the ideal parameter configuration for the model by equilibrating the degree of pretraining with the capacity for downstream generalization, as illustrated by the gray background. It is worth noting that in ablation experiments, our more selective criterion is the generalization ability of the model under certain fitting ability.

# D Additional experiments

**Base-to-New** Tab. S3 presents the empirical results of the base-to-new generalization experiments comparing our approach with existing methods. Our method demonstrates an average improvement of 7.73% and 1.13% over the baseline methods CoOp and PromptSRC, respectively, across average on eleven datasets. Compared to the prior pretrained method POMP, improvements are observed

Dataset	Method	1 shot (%)	PromptSRC) and 2 shots (%)	4 shots (%)	8 shots (%)	16 shots (%
	Linear probe CLIP	32.13	44.88	54.85	62.23	67.31
	СоОр	66.33	67.07	68.73	70.63	71.87
ImageNet	CoCoOp	69.43	69.78	70.39	70.63	70.83
-	MaPLe	62.67	65.10	67.70	70.30	72.33
	PromptSRC	68.13	69.77	71.07	72.33	73.17
	RPP (ours)	71.83	72.80	73.40	73.87	73.87
	Linear probe CLIP	79.88	89.01	92.05	93.41	95.43
	CoOp	92.60	93.07	94.40	94.37	95.57
Caltech101	CoCoOp	93.83	94.82	94.98		95.16
	MaPLe	92.57	93.97	94.43		96.00
	PromptSRC RPP (ours)	93.67 95.43	94.53 95.77	95.27 96.20		96.07 96.77
	Linear probe CLIP	34.59	40.76	55.71		69.96
DED	CoOp	50.23	53.60	58.70		69.87
DTD	CoCoOp MaDLa	48.54 52.13	52.17	55.04		63.04
	MaPLe PromptSRC	52.13 56.23	55.50 59.97	61.00 65.53		71.33 72.73
	RPP (ours)	63.73	70.43	72.87		73.60
	Linear probe CLIP	49.23	61.98	77.09		87.21
E CAT	CoOp	54.93	65.17	70.80		84.93
EuroSAT	CoCoOp MoDLo	55.33	46.74	65.56		73.32
	MaPLe PromptSRC	71.80 73.13	78.30 79.37	84.50 86.30	62.23 70.63 70.63 70.30 72.33 73.87	92.33 92.43
	RPP (ours)	80.4	82.03	90.67		92.43
	Linear probe CLIP	35.66	50.28	63.38		80.44
StanfordCars	CoOp CoCoOp	67.43 67.22	70.50 68.37	74.47 69.39		83.07 71.57
StaniordCars	MaPLe	66.60	71.60	75.30		83.57
	PromptSRC	69.40	73.40	77.13		83.83
	RPP (ours)	74.10	77.60	80.73		84.87
		69.74	85.07	92.02		97.37
	Linear probe CLIP CoOp	77.53	87.33	92.02	,	97.37 97.07
Flowers102	СоСоОр	72.08	75.79	78.40		87.84
1 10WC13102	MaPLe	83.30	88.93	92.67	73.67 79.30 70.44 79.47 80.97 83.47 96.10 94.97 84.30 95.80 96.27 96.60 39.35 39.00 26.61 42.00 43.27	97.00
	PromptSRC	85.93	91.17	93.87		97.60
	RPP (ours)	86.67	92.27	94.77		97.80
	Linear probe CLIP	19.61	26.41	32.33	30.35	45.36
	CoOp	21.37	26.20	30.83		43.40
FGVCAircraft	CoCoOp	12.68	15.06	24.79		31.21
1 G v Crimerant	MaPLe	26.73	30.90	34.87		48.40
	PromptSRC	27.67	31.70	37.47		50.83
	RPP (ours)	36.57	39.47	43.60	49.10	51.77
	Linear probe CLIP	41.58	53.70	63.00	69.08	73.28
	CoOp	66.77	66.53	69.97		74.67
SUN397	CoCoOp	68.33	69.03	70.21		72.15
	MaPLe	64.77	67.10	70.67		75.53
	PromptSRC	69.67	71.60	74.00		77.23
	RPP (ours)	73.23	74.57	75.93	77.03	77.40
	Linear probe CLIP	44.06	58.37	71.17	78.36	85.34
	СоОр	90.37	89.80	92.57		91.87
OxfordPets	CoCoOp	91.27	92.64	92.81		93.34
	MaPLe	89.10	90.87	91.90		92.83
	PromptSRC	92.00	92.50	93.43		93.67
	RPP (ours)	92.73	93.23	93.90	93.93	94.47
	Linear probe CLIP	53.66	65.78	73.28	79.34	82.11
	CoOp	71.23	73.43	77.10		82.23
UCF101	CoCoOp	70.30	73.51	74.82		78.14
	MaPLe	71.83	74.60	78.47		85.03
	PromptSRC	74.80	78.50	81.57		86.47
	RPP (ours)	80.67	83.23	85.23		86.50
	Linear probe CLIP	43.96	61.51	73.19		82.90
n us:	CoOp	84.33	84.40	84.47		84.20
Food101	CoCoOp	85.65	86.22	86.88		87.25
	MaPLe	80.50	81.47	81.77		85.33
	PromptSRC RPP (ours)	84.87 86.23	85.70 86.57	86.17 86.70		87.5 87.53
	Linear probe CLIP	45.83	57.98	68.01		78.79
	CoOp	67.56	70.65	74.02		79.89
Average	CoCoOp	66.79	67.65	71.21		74.90
	MaPLe PromptSRC	69.27	72.58	75.37		81.79
	PRODUNKI	72.32	75.29	78.35	80.69	82.87

at 5.3% and 2.77%, respectively. Notably, in datasets where CLIP's generalization performance is suboptimal, such as FGVCAircraft and DTD datasets, the use of POMP as a baseline results in a notable decrease in Harmonic Mean (HM) by 9.11% and 8.69%, respectively. This decline is attributable to POMP's underfitting during pretraining on ImageNet-21K and the subsequent loss of CLIP's inherent generalization capabilities, leading to a narrower convergence direction when its pretrained weights are employed for initialization. In contrast, our method, benefiting from both the fitting ability of the pretraining and the original CLIP's generalizability, manages to achieve performance gains of 0.63% and 0.9% over PromptSRC.

**Few-shot** Tab. S4 presents the fine-tuning experimental results of our method compared with existing approaches across a spectrum of shots. Owing to the appropriate fit of pretraining and retention of the inherent generalization capabilities of CLIP, our method yields an exceptionally notable enhancement in performance when training with a minimal number of data samples (1-shot, 2-shot), relative to existing methods. This is particularly evident on challenging datasets such as FGVCAircraft, where our method delivers performance improvements of 8.9% and 7.77% over PromptSRC at 1-shot and 2-shot, respectively. These findings further substantiate the value of the pretraining strategy we propose, termed RPP.

# **Broader Impacts**

Further research and careful consideration are necessary when utilizing this technology, as the presented proposed method relies on statistics derived from training datasets that may possess biases and could potentially result in negative societal impacts.

# Safeguards

Our paper employs the ImageNet-21K dataset for pretraining in an open-source multimodal model. Potential security concerns may arise from biases in the pretraining of open-source data and multimodal models. Please be mindful of biases in the original data and model, as well as the security of the model. We do not release any data or models; we only provide a pretraining approach.