UNIFIED VISION AND LANGUAGE PROMPT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Prompt tuning, a parameter- and data-efficient transfer learning paradigm that tunes only a small number of parameters in a pre-trained model's input space, has become a trend in the vision community since the emergence of large vision-language models like CLIP. We present a systematic study on two representative prompt tuning methods, namely text prompt tuning and visual prompt tuning. A major finding is that none of the unimodal prompt tuning methods performs consistently well: text prompt tuning fails on data with high intra-class visual variances while visual prompt tuning cannot handle low inter-class variances. To combine the best from both worlds, we propose a conceptually simple approach called Unified Prompt Tuning (UPT), which learns a tiny neural network to jointly optimize prompts across different modalities. Extensive experiments on over 11 vision datasets show that UPT achieves a better trade-off than the unimodal counterparts on few-shot learning benchmarks, as well as on domain generalization benchmarks. Code and models will be released to facilitate future research.

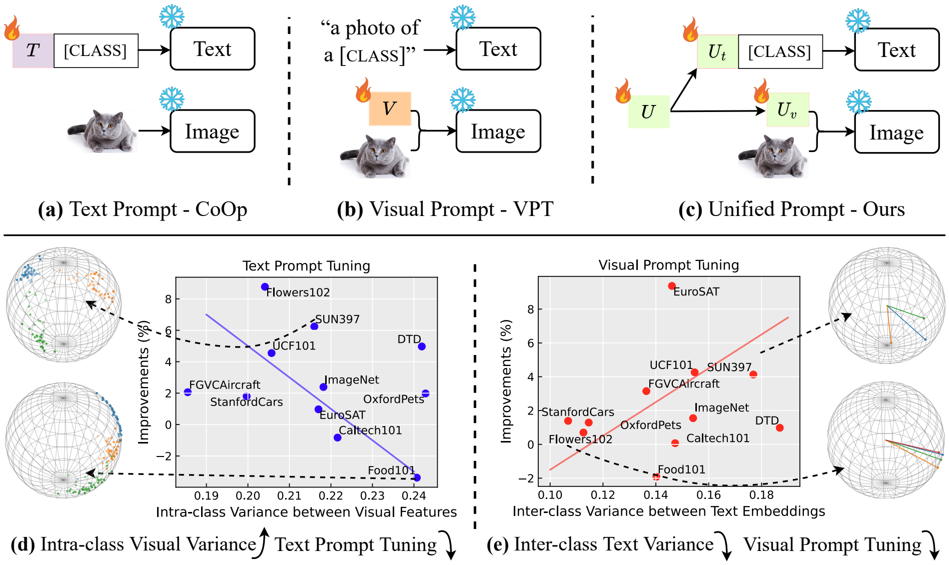
1 INTRODUCTION

Vision-language (VL) models pre-trained on millions of image-text pairs (*e.g.*, CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021)) have shown excellent transferability on a variety of downstream tasks, such as few-shot learning (Zhou et al., 2022a;b; Ju et al., 2021) and open-vocabulary perception (Gu et al., 2022; Zhou et al., 2022c; Zang et al., 2022; Ghiasi et al., 2021). When adapting large VL models to downstream tasks, it is often impractical to fine-tune the entire model directly due to their huge parameter size. To make adaptation more efficiently, many studies (Gao et al., 2021; Li & Liang, 2021; Lester et al., 2021; Zhou et al., 2022a; Lu et al., 2022; Ju et al., 2021; Yao et al., 2021; Jia et al., 2022; Bahng et al., 2022) have explored prompt tuning where the idea is to fine-tune a small number of parameters in a pre-trained model's input space, called *prompt*, while keeping the majority of pre-trained parameters frozen.

A typical VL model consists of two sub-networks—an image encoder and a text encoder—to extract features from visual and textual modalities respectively. Correspondingly, existing prompt tuning approaches can be grouped into two types: text prompt tuning and visual prompt tuning. For text prompt tuning methods, *e.g.*, CoOp (Zhou et al., 2022a), extra text prompt tokens treated as learnable parameters are applied on the text encoder (Fig. 1(a)) to mitigate the issue that hand-crafted text prompt templates (*e.g.*, "`a photo of a [CLASS].`") are often sub-optimal. On the contrary, visual prompt tuning approaches focus on modulating the image encoder (Fig. 1(b)). A representative method is VPT (Jia et al., 2022), which injects learnable parameters into multiple layers of a Vision Transformer. Notably, these prompt-based methods treat the two modalities in isolation.

Despite significant improvements achieved recently, we observe that current prompt tuning approaches (Zhou et al., 2022a; Jia et al., 2022) fail to perform consistently due to inherent variances in visual and text features in downstream tasks. That is to say, using the unimodal prompt may obtain good results on one dataset but not on others.

To analyze the phenomenon, we measure the discrepancy in data distribution focusing on the intra-class variance of *visual* features and inter-class variance of *text* embedding, and study the correlation between data statistics and performance improvement. As shown in Fig. 1(d), when the intra-class variance of image features is large (bottom right), CoOp struggles to learn suitable text prompts for improving the text classifier. As for visual prompt tuning, VPT faces difficulties when the inter-class variance of text features is small, as shown in bottom left of Fig. 1(e). That is, if the text



(a) Text Prompt - CoOp **(b)** Visual Prompt - VPT **(c)** Unified Prompt - Ours

(d) Intra-class Visual Variance↑ Text Prompt Tuning↓ **(e)** Inter-class Text Variance↓ Visual Prompt Tuning↓

Figure 1: **Top**: Architectures of **(a)** *text* prompt tuning (Zhou et al., 2022a), **(b)** *visual* prompt tuning (Jia et al., 2022) and **(c)** our multimodal *unified* prompt tuning (🔥: learnable; ❄: frozen parameters). **Bottom**: the performance improvements (%) of text prompt tuning **(d)** and visual prompt tuning **(e)** compared with the zero-shot CLIP baseline. We show that the variance of visual and text features (x-axis) will affect the improvements (y-axis). We project the text/visual features of the dataset (pointed by the dashed arrow) into a unit sphere to show the variance of different distributions. Please refer to the appendix for the implementation details about how we compute the feature variance.

classifiers are based on text features of low separability, tuning visual prompts would lend little help to improve the final performance. Moreover, intra-class visual variance and inter-class text variance are typically orthogonal. As a consequence, the performance of unimodal prompt tuning methods varies widely across different datasets: CoOp beats VPT by 8.1% on Flowers102 (Nilsback & Zisserman, 2008) while VPT outperforms CoOp by 8.4% on EuroSAT (Helber et al., 2019).

We argue that the key would be to simultaneously adapt both text and visual prompts to overcome the vast differences across different data distributions. A straightforward solution is to introduce both text and visual prompts to the model and jointly optimize the two modality-specific prompts. However, we find that such a naïve joint training leads to poor performance due to the intrinsic discrepancy between text and image modalities. In particular, the performance is occasionally worse than tuning modality-specific prompts as shown in our experiments.

Solving the aforementioned issues requires modality-agnostic optimization to bridge the isolated prompts. To this end, we present a unified prompt tuning method for both *text* and *visual* modalities, dubbed **U**nified **P**rompt **T**uning (**UPT**). See Fig. 1(c). Specifically, we start with a shared prompt and propose a lightweight self-attention network to generate the prompts for CLIP's *text* and *visual* encoders respectively. We empirically show that such a conceptually simple design can preserve the benefit of individual modalities.

Our contributions are summarized as follows. **1)** We provide a comprehensive study on existing text and visual prompt tuning methods, and identify the shortcoming of unimodal learning. **2)** We present a unified prompt learning method for VL models, which is simple and easy to implement. **3)** We conduct extensive experiments to show that unified prompt tuning outperforms previous unimodal prompt tuning methods under the few-shot learning and domain generalization settings.

2 Methodology

We first introduce vision-language models focusing on CLIP (Radford et al., 2021), in company with text/visual prompt tuning approaches for visual recognition in Sec. 2.1. We then analyze the limitations of previous single-modal prompt tuning approaches in Sec. 2.2. Finally, we present technical details of our proposed unified prompt learning in Sec. 2.3.

2.1 Preliminaries

CLIP. CLIP (Radford et al., 2021) consists of two sub-networks: an image encoder ϕ and a text encoder ψ. These two encoders, respectively, map the text and image inputs into a joint hidden space $\mathbb{R}^d$, where the semantics of vision and language modalities are well-aligned. Here, d refers to the final hidden dimension of the text or image encoder (*e.g.*, $d = 256$ in the ResNet (He et al., 2016) backbone and $d = 512$ in the ViT backbone). Given an input image $\boldsymbol{x}$ and a set of categories $\mathbf{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_k\}$ (*e.g.*, $k = 1000$ for ImageNet (Deng et al., 2009)), the image encoder extracts the corresponding image feature $\boldsymbol{z} = f_\phi(\boldsymbol{x}) \in \mathbb{R}^d$. While the class names in $\mathbf{Y}$ are first filled into a hand-crafted text prompt template `a photo of a [CLASS]` to obtain the text descriptions $\mathbf{A}$, further processed by the text encoder for the text representations: $\mathbf{W} = f_\psi(\mathbf{A}) \in \mathbb{R}^{d\times k}$. The final prediction is computed as follows:

$$p(y = i \mid \boldsymbol{x}) = \frac{\exp\left(\cos\left(\boldsymbol{w}_i, \boldsymbol{z}\right)/\tau\right)}{\sum_{j=1}^{k} \exp\left(\cos\left(\boldsymbol{w}_j, \boldsymbol{z}\right)/\tau\right)}, \tag{1}$$

where $\cos(\cdot,\cdot)$ denotes the cosine similarity and τ is a fixed temperature value (*e.g.*, $\tau = 100$). Conceptually, such a decision process for the input image $\boldsymbol{x}$ in Eq. (1) is formulated in a way that the text encoder ψ takes a role of generating dynamic **classifiers** $\mathbf{W}$ from open-set categories $\mathbf{Y}$, with the image encoder ϕ producing encoded visual **features** $\boldsymbol{z}$. In practice, it is generally infeasible to fine-tune the millions of parameters (*i.e.*, ϕ and ψ) in a VL model for transfer learning in every downstream task.

Text Prompt Tuning. For efficient and effective model adaptation, text prompt tuning approaches consider generating more adaptive **classifiers** without fine-tuning the text encoder ψ. For example, Context Optimization (CoOp) (Zhou et al., 2022a) introduce a set of learnable parameters $\mathbf{T} \in \mathbb{R}^{d\times m}$ to replace the hand-crafted text prompt template (`a photo of a [CLASS]`). The word-embedding of class names in $\mathbf{Y}$ will concatenate with these text prompts in the following form:

$$\hat{\mathbf{T}} = [\boldsymbol{t}_1, \boldsymbol{t}_2, \ldots, \boldsymbol{t}_m, \texttt{CLASS}]. \tag{2}$$

Here, the symbol m denotes the prompt length. The resulting dynamic text representations are extracted by the text encoder: $\mathbf{W} = f_\psi(\hat{\mathbf{T}}) \in \mathbb{R}^{d\times k}$. In each downstream task, the learnable prompts $\mathbf{T}$ will be optimized with each task-specific objective function, *e.g.*, a cross-entropy classification loss $\mathcal{L}_{\text{CE}}(p, y)$ in few-shot learning. Note that both the image and text encoders (ϕ and ψ) are frozen during downstream training. As a result, updating the text prompt $\mathbf{T}$ will correspondingly adjust the decision boundaries with generated classifiers $\mathbf{W}$ for downstream tasks.

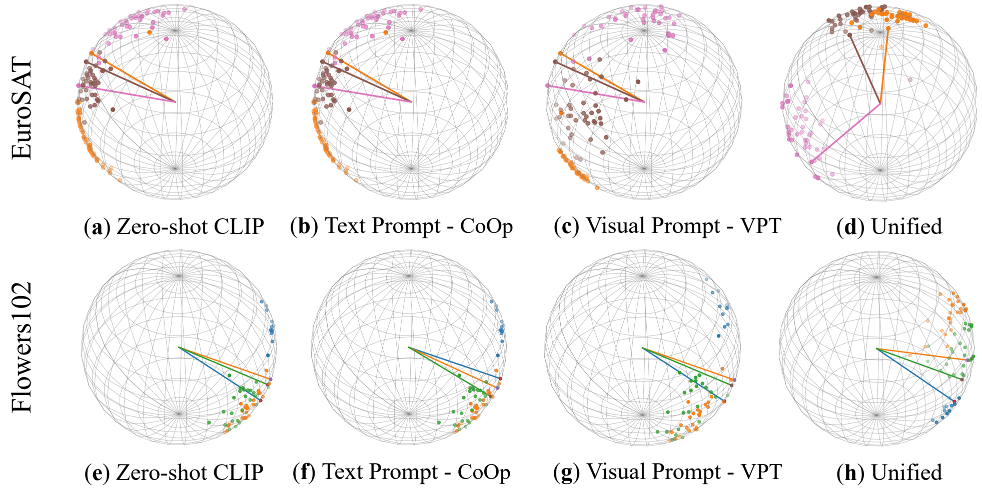
Visual Prompt Tuning. Conversely, visual prompt tuning methods focus on extracting more transferable visual **features** while keeping the visual encoder ϕ unchanged. Following the success of text prompt tuning approaches, recent Visual Prompt Tuning (VPT) (Jia et al., 2022) introduces a similar prompt tuning recipe for the visual encoder ϕ. Suppose the image encoder ϕ contains L Vision Transformer layers, the output of i-th layer, l_i, where $i = 1, 2, \ldots, L$, is given by:

$$[\boldsymbol{c}^{i+1}, \boldsymbol{z}_1^{i+1}, \ldots, \boldsymbol{z}_s^{i+1}] = l_i\left([\boldsymbol{c}^i, \boldsymbol{z}_1^i, \ldots, \boldsymbol{z}_s^i]\right), \tag{3}$$

where $\boldsymbol{c} \in \mathbb{R}^d$ denotes the classification token (`[CLS]`), and $\boldsymbol{Z} = [\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_s] \in \mathbb{R}^{d\times s}$ denotes the input image patch tokens with length s. For the i-th encoder layer, a set of learnable visual prompts $\mathbf{V}^i \in \mathbb{R}^{d\times n}$ are inserted and computed as follows:

$$[\boldsymbol{c}^{i+1}, \ldots, \boldsymbol{Z}^{i+1}] = l_i\left([\boldsymbol{c}^i, \mathbf{V}^i, \boldsymbol{Z}^i]\right), \tag{4}$$

where n stands for the length of visual prompts. Two VPT variants are proposed: VPT-shallow and VPT-deep. For VPT-shallow, the visual prompts are only inserted into the first Transformer layer ($i = 1$). Whereas for VPT-deep, visual prompts are introduced at every layer. The learnable visual prompts are data-independent, which once learned, can modulate the visual features $\boldsymbol{z}$ of input images for better downstream transfer learning.



(**a**) Zero-shot CLIP (**b**) Text Prompt - CoOp (**c**) Visual Prompt - VPT (**d**) Unified

(**e**) Zero-shot CLIP (**f**) Text Prompt - CoOp (**g**) Visual Prompt - VPT (**h**) Unified

Figure 2: Visualization of input features z (projected points) and text classifier $\mathbf{W}$ (projected lines) on EuroSAT and Flowers102.

2.2 Analysis

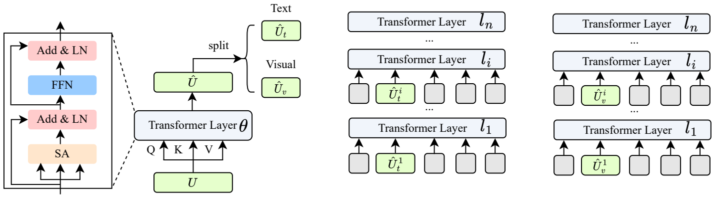
We conduct a series of probing studies to analyze the characteristics of text/visual prompt tuning. First, when adapting the CLIP model with two representative text and visual prompt tuning approaches (CoOp (Zhou et al., 2022a) and VPT (Jia et al., 2022)), we measure the variance of both visual features z and text embeddings $\mathbf{W}$ (*i.e.*, classifiers) for all 11 downstream vision datasets (see Appendix for detailed implementations). For text prompt tuning, as shown in Fig. 1(d), we observe that CoOp performs well on datasets with low intra-class variance between visual features, such as Flowers102, but fails on Food101 dataset with high intra-class feature variance. As for visual prompt tuning, VPT succeeds in improving performance on SUN397 dataset with large inter-class text embeddings, while being less effective on Food101 and Flowers102 with relatively smaller inter-class text embedding variance. The performance improvements of text/visual prompt tuning are highly correlated with the variance of visual features z or text embeddings $\mathbf{W}$ in downstream datasets.

In order to understand this phenomenon, we select two downstream vision datasets (Flowers102 (Nilsback & Zisserman, 2008), EuroSAT (Helber et al., 2019)) for further analysis. During downstream training, we project both visual features z and text embeddings $\mathbf{W}$ (*i.e.*, classifiers) into joint sphere space $\mathbb{R}^3$ for better visualization. As we illustrated in Fig. 2, we can observe that: **1**) For the EuroSAT dataset with high intra-class visual feature variance, text prompts in CoOp fails to adapt the text classifiers $\mathbf{W}$. Clearly, the text classifiers in Fig. 2(**b**) are almost unchanged compared with zero-shot CLIP baseline (Fig. 2(**a**)). **2**) For the Flowers102 dataset with low inter-class text embedding variance, visual prompts in VPT are not effective in modulating the visual features z (Fig. 2(**g**)), thus cannot obtain considerable performance gain.

In conclusion, the single-modal prompt tuning approaches (CoOp and VPT), face the dilemma that consistent improvements over Zero-shot CLIP are hard to achieve due to inherent variances of visual features and text embedding in downstream tasks. Our observation motivates us to present a unified prompt tuning method that tunes the z and $\mathbf{W}$ at the same time.

2.3 Unified Prompt Tuning

Driven by our analysis, we devise a simple yet effective multi-modal **U**nified **P**rompt **T**uning (UPT) approach for adapting VL models. Specifically, instead of introducing two sets of isolated modality-specific prompts (*i.e.*, $\mathbf{T}$ in Eq. (2) and $\mathbf{V}$ in Eq. (4)) for the text and visual encoders, we consider learning a set of unified modality-agnostic prompts for tuning VL models. As shown in Fig. 3, we define a set of learnable prompts $\boldsymbol{U} \in \mathbb{R}^{d \times n}$ with length n. Rather than naïvely appending the unified prompts into the text and visual encoders, we employ a lightweight Transformer layer θ to



(a) Unified Prompt **(b)** Text Encoder **(c)** Image Encoder

Figure 3: The architecture of (**a**) our unified prompt $\boldsymbol{U}$ that is applied to (**b**) CLIP *text* encoder and (**c**) CLIP *image* encoder.

transform unified prompts $\boldsymbol{U}$ as follows:

$$\begin{aligned} \boldsymbol{U}' &= \mathrm{SA}\left(\boldsymbol{U}\right) + \mathrm{LN}\left(\boldsymbol{U}\right), \\ \hat{\boldsymbol{U}} &= \mathrm{FFN}\left(\mathrm{LN}\left(\boldsymbol{U}'\right)\right) + \mathrm{LN}\left(\boldsymbol{U}'\right), \end{aligned} \tag{5}$$

where the self-attention operator SA, feed-forward network FFN and layer normalization LN are applied to obtain the transformed prompts $\hat{\boldsymbol{U}}$. The self-attention module in the lightweight Transformer layer allows beneficial interaction between two modalities, so as to maximize the complementary effects. Our unified prompts can be introduced into multiple layers of VL models. In particular, for each i-th layer of text and image encoders, we consider learning a set of layer-wise prompts $\boldsymbol{U}^i$, and split transformed $\hat{\boldsymbol{U}}^i$ into two parts $\hat{\boldsymbol{U}}^i = \{\hat{\boldsymbol{U}}_t^i, \hat{\boldsymbol{U}}_v^i\}$, sending into the text and visual encoders respectively. During downstream training, we froze both the text and visual encoder (ψ and ϕ) and only optimize the unified prompts $\boldsymbol{U}$ and the lightweight Transformer layer θ. In this way, both the dynamic classifiers $\mathbf{W}$ and visual features $\boldsymbol{z}$ in Eq. (1) are effectively tuned for reliable prediction in the downstream task. As shown in Fig. 2 (d) and (h), our unified prompts can simultaneously obtain well-aligned text classifiers and separable visual features compared with single-modal counterparts.

3 Experiments

In this section, we conduct experiments under two problem settings, *i.e.*, (i) few-shot image classification (Sec. 3.1) and (ii) domain generalization (Sec. 3.2). We also present ablation studies in Sec. 3.3 on several design choices and extra experimental results about generalizability in Appendix.

Baselines. We compare our approach against the following methods: (1) **Zero-shot CLIP**. This baseline uses hand-crafted text prompt templates and does not involve any prompt-learning strategies. (2) **Single-modal Prompt Tuning** methods, including CoOp (Zhou et al., 2022a) and ProDA (Lu et al., 2022) for the text modality, and VPT (Jia et al., 2022) for the visual modality. In the domain generalization setting, we further compare with CoCoOp (Zhou et al., 2022b), which improves CoOp's generalization performance with an input-conditional design. For VPT, we report the results of both the shallow and deep variants, as described in Sec. 2.1.

3.1 Few-shot Learning

In this section, we measure a model's generalization ability by conducting prompt tuning using different strategies, with just a limited amount of labeled examples per-class in the specific downstream task. Detailed implementation is presented in Appendix.

Datasets. We follow (Zhou et al., 2022b) to use 11 datasets (ImageNet (Deng et al., 2009), Caltech101 (Fei-Fei et al., 2004), OxfordPets (Parkhi et al., 2012), StanfordCars (Krause et al., 2013), Flowers102 (Nilsback & Zisserman, 2008), Food101 (Bossard et al., 2014), FGVC-Aircraft (Maji et al., 2013), SUN397 (Xiao et al., 2010), UCF101 (Soomro et al., 2012), DTD (Cimpoi et al., 2014), EuroSAT (Helber et al., 2019)) as our benchmarks. Following (Zhou et al., 2022a), we use the few-shot evaluation protocol selecting 1/2/4/8/16 shots for training and the whole test set for evaluation. We report averaged results over three runs with different random seeds to reduce the variance. The detailed results are shown in Fig. 4.

Limitation of Single-modal Baselines. Figure 4 shows that the performance improvements of existing *text* prompt tuning method CoOp and *visual* prompt tuning method VPT are not consistent across different datasets. In particular, CoOp obtains better performance than VPT on some datasets, such

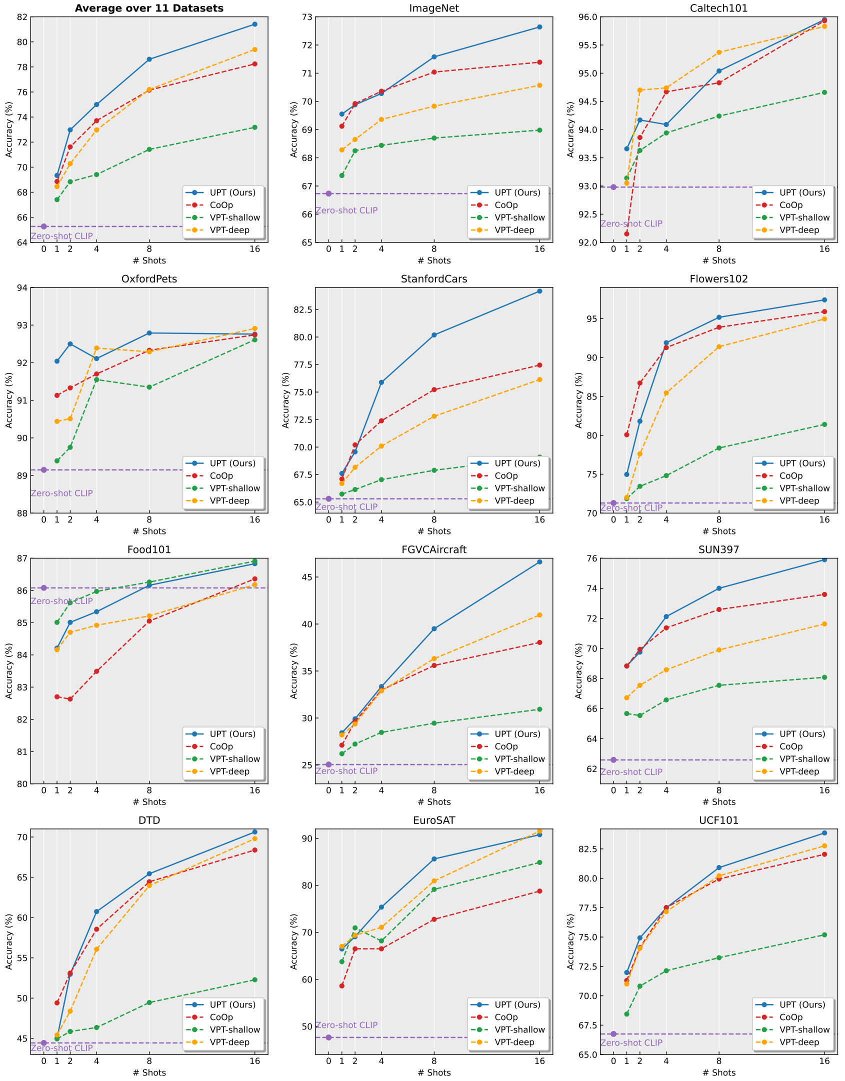


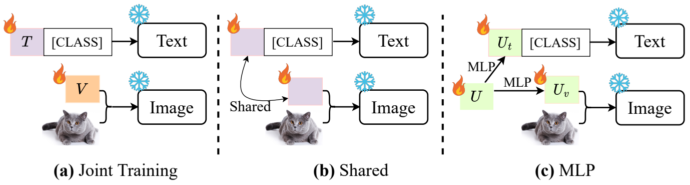
Figure 4: **Main results over 11 datasets under the few-shot learning setting.** We report the average accuracy (%) of 1/2/4/8/16 shots over three runs. Overall, the proposed UPT (blue line) achieves apparent improvements compared with the Zero-shot CLIP and single-modal prompt tuning baselines (CoOp, ProDA and VPT).

as StanfordCars and SUN397. However, for other datasets with high intra-class visual variances, VPT is much more effective than CoOp. For instance, on the EuroSAT dataset, VPT-deep beats CoOp by over 12%. The discrepancy of previous single-modal baselines is also consistent with our motivation in Fig. 1(d) and Fig. 1(e). According to VPT (Jia et al., 2022), VPT-deep is more effective than VPT-shallow, and our experimental results also verify this point. We later show that VPT-shallow obtains much stronger performance than the VPT-deep in the domain generalization setting (Sec. 3.2).

UPT vs. Single-modal Baselines. Our UPT achieves clear advantages over the single-modal prompt-tuning counterparts CoOp, ProDA and VPT, as suggested by the averaged performance (top-left of Fig. 4). In general, the average performance gap between UPT and baselines increases with the shot number available for prompt tuning. Specifically, UPT obtains 0.48/1.36/1.29/2.46/3.19(%) accuracy improvements compared with the *text* prompt tuning method CoOp on 1/2/4/8/16 shots settings. Even compared with the strong *text* prompt tuning baseline ProDA, UPT still boosts the accuracy of 0.11/1.01/0.67/1.5/1.61(%). Similarly, UPT achieves 0.89/2.70/2.03/2.40/2.01(%) ac-

Table 1: **Main results under the domain generalization setting.** We report the average accuracy (%) of 16 shots over three runs. The **best** and **second best** methods are highlighted in red and orange, respectively.

#	Method	Source	Target				*Overall*	*OOD*
		ImageNet	-V2	-S	-A	-R	Average	Average
1	CoOp	71.51	64.20	47.99	49.71	75.21	61.72	59.28
2	CoCoOp	71.02	64.07	**48.75**	50.63	76.18	62.13	59.91
3	VPT-shallow	68.98	62.10	47.68	47.19	76.10	60.38	58.27
4	VPT-deep	70.57	63.67	47.66	43.85	74.42	60.04	57.40
5	Joint Training	71.42	64.36	48.20	49.71	76.23	61.97	59.61
6	Shared	71.46	64.43	48.13	50.03	75.76	61.96	59.55
7	MLP	71.00	64.11	48.65	48.76	76.14	61.78	59.48
8	UPT	**72.63**	64.35	48.66	**50.66**	**76.24**	**62.51**	**59.98**



(a) Joint Training

(b) Shared

(c) MLP

Figure 5: Ablation studies on different design choices. **(a)**: jointly train the existing *text* and *visual* prompt tuning approaches; **(b)**: shared prompts for all modalities; **(c)**: using two MLP layers to generate the prompts.

curacy gains over the *visual* prompt tuning approach VPT-deep. Notably, UPT significantly boosts the performance over CoOp and VPT-deep on challenging large datasets, such as ImageNet with 1,000 classes and SUN397 with 397 categories. UPT also surpasses CoOp and VPT-deep on fine-grained datasets such as StanfordCars and FGVC Aircraft. We also observe that UPT shows less improvement on the two datasets (OxfordPets and Food101), possibly caused by the noisy training data (Zhou et al., 2022a; Bossard et al., 2014). Overall, the experimental results in Fig. 4 demonstrate the effectiveness of our proposed UPT.

3.2 Domain Generalization

Pre-trained VL models like CLIP have shown strong generalization ability. However, the prompt tuned on a specific downstream dataset may hinder the generalization ability on categories outside the training set. In this section, we evaluate the generalization ability of different prompt tuning methods on out-of-distribution (OOD) data.

Datasets. We follow (Zhou et al., 2022a) to use five datasets (ImageNet (Deng et al., 2009), ImageNet V2 (Recht et al., 2019), ImageNet-Sketch (Wang et al., 2019), ImageNet-A (Hendrycks et al., 2021b) and ImageNet-R (Hendrycks et al., 2021a)) for evaluation. Following the protocol, we train a model on ImageNet and evaluate it on four other variants of ImageNet with their domains shifted.

Results. Table 1 summarizes the results. We report the average accuracy on both the source and target datasets (penultimate column), and the OOD average accuracy on target datasets (last column). The results show that VPT-shallow (row #2) achieves higher OOD accuracy than VPT-deep (row #3), and *text* prompt tuning methods outperform *visual* prompt tuning approaches. Furthermore, the proposed UPT (row #8) is generally a better option than single-modal baselines (rows #1-#4) and obtains comparable performance with CoCoOp. Our UPT achieves the best results three times on five datasets, showing that UPT is a reliable prompt tuning method among its competitors in the domain generalization setting.

3.3 Ablation Studies

Comparison with the Joint Training Baseline. As shown in Fig. 5(a), a straightforward approach for multi-modal prompts is tune the *text* prompt (using CoOp) and *visual* prompt (using VPT) jointly.

Table 2: Ablation studies on different multi-modal prompt design choices in Fig. 5 over 11 datasets. We report the accuracy results under the 16 shots setting. The **best** and **second best** methods are highlighted in red and orange, respectively.

#	Method	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	*Average*
1	CoOp	71.36	95.93	92.74	77.45	95.90	86.36	38.04	73.59	68.38	78.77	82.04	78.24
2	VPT-shallow	68.98	94.66	92.61	69.09	81.40	86.91	30.93	68.08	52.28	84.87	75.19	73.18
3	VPT-deep	70.57	95.83	92.91	76.13	94.96	86.18	40.96	71.63	69.79	**91.53**	82.76	79.39
4	Joint Training	71.42	95.84	93.34	79.02	95.25	86.55	40.56	74.17	67.83	78.94	82.81	78.70
5	Shared	71.46	95.50	92.99	78.66	95.55	86.67	39.18	73.64	67.69	73.36	82.06	77.88
6	MLP	71.00	95.59	**93.74**	75.88	93.38	**87.20**	37.17	72.74	67.31	90.66	81.43	78.73
7	UPT (Ours)	**72.63**	**95.94**	92.95	**84.33**	**97.11**	85.00	**46.80**	**75.92**	**70.65**	90.51	**84.03**	**81.44**

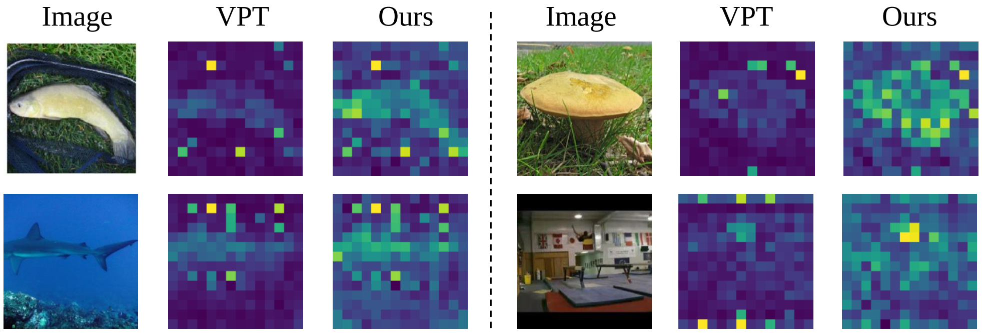


Figure 6: Visualization of attention response map between visual prompts and image patch tokens. The images are test images from ImageNet. We visualize the self-attention module from the last block of ViT of the CLIP image encoder.

We investigate the effectiveness of such joint training scheme, and report its results in Table 2 row #4 and Table 1 row #5. On the few-shot learning setting, we see that such a joint training solution performs better than CoOp and VPT-shallow, which shows that multi-modal optimization is helpful to a certain extent. But the joint training approach obtains slightly worse accuracy than the VPT-deep (78.70% vs. 79.39%) since VPT-deep involves a large number of parameters. On the domain generalization setting, we find the joint training method performs much better than VPT-deep. Also, the joint training method shows inferior performance to our UPT, demonstrating that our self-attention base mechanism is more effective.

Shared Prompts for *Text* and *Visual* Modalities. We also investigate the results of directly sharing prompts for different modalities. As shown in Fig. 5(b), the shared prompts will be optimized for both *text* and *visual* modalities. This scheme differs from the proposed UPT, where the shared prompts are transformed with self attention. Experimental results are presented in Table 2 row #5 and Table 1 row #6, and we observe that such a prompt sharing strategy achieves worst performance among all the ablation design choices.

MLP Baseline. For our proposed UPT, we use a Transformer layer with the self-attention operator to partially share the hyper-parameters for different modalities. Here, we study a simpler design that generates the unified prompts with two MLP layers. Results are presented on Table 2 row #6 and Table 1 row #7. The MLP baseline is still competitive, yielding best performance on two datasets. Nonetheless, the average results is still poorer than the proposed self-attention based approach.

3.4 QUALITATIVE RESULTS

While it is hard to visualize what have been learned during text prompt tuning, it is possible to visualize the visual prompts learned by VPT and UPT following the self-supervised learning method, DINO (Caron et al., 2021). In particular, for each layer of the Vision Transformer (ViT), we can compute the self-attention response map of visual prompts and image patch tokens. Figure 6 compares such response maps by VPT and the proposed UPT. We find that UPT shows stronger self-attention responses compared with VPT. This could be the possible reason why UPT achieves better performance on the few-shot learning and the OOD generalization settings.

4 RELATED WORK

Vision-Language Models. Recent vision-language pre-trained models (Radford et al., 2021; Jia et al., 2021) use the contrastive loss to align an image encoder (*e.g.*, ViT (Dosovitskiy et al., 2021)) and a text encoder (*e.g.*, BERT (Kenton & Toutanova, 2019)) in a common feature space. These vision-language models are trained on web-scale image-text pairs and are transferable across various downstream tasks such as point cloud classification (Zhang et al., 2022a), video classification (Qian et al., 2022), object detection (Gu et al., 2022; Du et al., 2022; Zhou et al., 2022c; Zang et al., 2022) and semantic segmentation (Ghiasi et al., 2021). In this work, we aim to explore how to adapt the CLIP model to the downstream few-shot recognition task.

Text Prompt Tuning. The concept of prompt tuning was first proposed in the NLP area (Liu et al., 2021; Gao et al., 2021; Li & Liang, 2021; Lester et al., 2021). In particular, a text prompt refers to a task-specific template for language models. For example, in sentiment analysis, the template might be "`I [MASK] the movie.`" where the mask placeholder will be filled with either "love" or "hate." Common practices in text prompt tuning include (i) searching for a specific word in the dictionary, known as *hard* prompt learning (Gao et al., 2021), or (ii) turning masked tokens into learnable vectors, known as *soft* prompt learning (Li & Liang, 2021; Lester et al., 2021). Text prompt tuning has also been applied in computer vision after the emergence of large vision-language models (*e.g.*, CLIP (Radford et al., 2021)), which are too big to fine-tune. A representative work is CoOp (Zhou et al., 2022a), which turned the input context tokens in CLIP's text branch into learnable vectors for adapting CLIP to downstream image recognition. Other follow-ups of CoOp include CoCoOp (Zhou et al., 2022b), DualCoOp (Sun et al., 2022), ProGrad (Xing et al., 2022), and ProDA (Lu et al., 2022).

Visual Prompt Tuning. The idea of visual prompt tuning is to adapt large pre-trained Vision Transformers (Dosovitskiy et al., 2021) by adding learnable parameters in the visual input space, which is analogous to text prompt tuning in NLP. VPT (Jia et al., 2022) and Visual Prompting (Bahng et al., 2022) both add trainable tokens to the input of Transformer models. A recent work, NOAH (Zhang et al., 2022b), uses neural architecture search algorithms to identify the optimal configuration of prompt modules. In comparison to the unimodal prompt learning methods discussed above, our paper provides a timely study on how to achieve a better trade-off using multimodal prompt learning.

5 CONCLUSION

With the rapid scaling of vision models along the size dimension, efficient downstream adaptation methods have become essential for facilitating large-scale deployment of vision models in the wild. Our paper provides a timely and comprehensive study on how to adapt large vision-language models like CLIP from the prompt learning perspective. In particular, our study unveils that the previous unimodal prompt tuning methods do not work consistently well across different computer vision datasets. In contrast, the proposed UPT method, despite having a simple design, achieves a better trade-off compared with the unimodal counterparts. The results suggest that one should exploit correspondences between different modalities for prompt learning.

On the other hand, the results achieved by UPT are by no means perfect: in the ablation studies we observe that some alternative designs, such as using MLP instead of Transformer, might sometimes give better performance. In summary, we believe multimodal prompt learning is a promising framework, and we expect more improvements to be achieved with more advanced (and efficient) designs.

REFERENCES

Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Visual prompting: Modifying pixel space to adapt pre-trained models. *arXiv preprint arXiv:2203.17274*, 2022.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, pp. 446–461. Springer, 2014.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pp. 9650–9660, 2021.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, pp. 3606–3613, 2014.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255. IEEE, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *CVPR*, pp. 14084–14093, 2022.

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*, pp. 178–178. IEEE, 2004.

Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. In *ACL*, 2021.

Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Open-vocabulary image segmentation. *arXiv preprint arXiv:2112.12143*, 2021.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

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 J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 12(7):2217–2226, 2019.

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 *ICCV*, pp. 8340–8349, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, pp. 15262–15271, 2021b.

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 *ICML*, pp. 4904–4916. PMLR, 2021.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. *arXiv preprint arXiv:2203.12119*, 2022.

Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. *arXiv reprint arXiv:2112.04478*, 2021.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pp. 4171–4186, 2019.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV workshops*, pp. 554–561, 2013.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *EMNLP*, 2021.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *ACL*, 2021.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*, 2021.

Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *CVPR*, 2022.

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

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, pp. 722–729. IEEE, 2008.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, pp. 3498–3505. IEEE, 2012.

Rui Qian, Yeqing Li, Zheng Xu, Ming-Hsuan Yang, Serge Belongie, and Yin Cui. Multimodal open-vocabulary video classification via pre-trained vision and language models. *arXiv preprint arXiv:2207.07646*, 2022.

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*, pp. 8748–8763. PMLR, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, pp. 5389–5400. PMLR, 2019.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Ximeng Sun, Ping Hu, and Kate Saenko. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. *arXiv preprint arXiv:2206.09541*, 2022.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, pp. 10506–10518, 2019.

Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *CVPR*, pp. 4582–4591, 2017.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pp. 3485–3492. IEEE, 2010.

Yinghui Xing, Qirui Wu, De Cheng, Shizhou Zhang, Guoqiang Liang, and Yanning Zhang. Class-aware visual prompt tuning for vision-language pre-trained model. *arXiv preprint arXiv:2208.08340*, 2022.

Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Cpt: Colorful prompt tuning for pre-trained vision-language models. *arXiv preprint arXiv:2109.11797*, 2021.

Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. Open-vocabulary detr with conditional matching. In *ECCV*, 2022.

Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *CVPR*, pp. 8552–8562, 2022a.

Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. Neural prompt search. *arXiv preprint arXiv:2206.04673*, 2022b.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022b.

Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022c.

Appendix

In the supplementary materials, we discuss the implementation details and more experimental results. Section A explains how we compute the intra-/inter- class variance for Fig.(1) of the main paper. Section B reports the implementation details of our paper. Section C presents more experimental results under the base-to-new generalization and cross-dataset transfer settings.

A Intra-/Inter- Class Variance

In this section, we provide the implementation details about how we compute the *intra-class* visual variance and *inter-class* text variance for different datasets (Fig.1 (d)(e) in the main paper.

Intra-class Visual Variance. Given one dataset with k classes in total, for each image $\boldsymbol{x}$ that belongs to class c, we first use the CLIP image encoder ϕ to extract the corresponding image feature $f_\phi(\boldsymbol{x})$. Then we get the intra-class variance of class c as:

$$\text{var}_c = \frac{1}{||X_c||} \sum_{x \in X_c} \left(f_\phi(\boldsymbol{x}) - \bar{f}_\phi(\boldsymbol{x}) \right)^2 , \tag{6}$$

where X_c denotes to the set of images that have the ground-truth class label c, and $\bar{f}_\phi(\boldsymbol{x})$ refers to the mean values of class c. Then we can compute the intra-class variance Var_v for all the k classes as

$$\text{Var}_\text{v} = \frac{1}{k} \sum_{c=1}^{k} \text{var}_c . \tag{7}$$

Inter-class Text Variance. For each dataset, we first compute the CLIP text features $\boldsymbol{w}$ of classs c, and the mean value $\bar{\boldsymbol{w}}$ of all the k classes. Then we get the inter-class text variance Var_t as:

$$\text{Var}_\text{t} = \frac{1}{k} \sum_{c=1}^{k} (w_c - \bar{w})^2 . \tag{8}$$

B Implementation Details

Our implementation is based on the source code of CoOp (Zhou et al., 2022a). We use ViT-B/16 as the CLIP backbone (Radford et al., 2021). Following (Zhou et al., 2022b), we set the context length of CoOp as $m = 4$ (same for VPT). For Zero-shot CLIP and VPT, we use the default prompt template, "`a photo of a [CLS].`" We use SGD as the optimizer, with an initial learning rate of 0.002, which is decayed by the cosine annealing rule. The batch size is set to 32 for all datasets.

C More Experimental Results

Recent work CoCoOp (Zhou et al., 2022b) points out that the text prompts learned by CoOp (Zhou et al., 2022a) are not generalizable to novel classes and out-of-distribution data. CoCoOp defines two new settings - base-to-new generalization and cross-dataset transfer - to measure the generalizability ability of prompt learning approaches. In this section, we provide the experimental results of our UPT in these two settings.

Datasets. We use the same 11 datasets we used in the few-shot learning setting (section 3.1 in the main paper). Following CoCoOp (Zhou et al., 2022b), we use the 16-shot protocol and report the averaged results over three runs, and set the training schedule as ten epochs. We report the accuracy on base and new classes, and the harmonic mean for base-to-novel trade-off.

C.1 Base-to-new Generalization

In the base-to-new generalization setting, we split the classes into two disjoint groups - base classes and new classes. All the prompt learning approaches are required to train on the base classes, while

Table 3: **Comparison results in the base-to-new generalization setting**. H: Harmonic mean (Xian et al., 2017). The **best** and **second best** methods are highlighted in red and orange, respectively. The method 'VPT-s' refers to VPT-shallow.

(a) **Average over 11 datasets**.

	Base	New	H
CLIP	69.34	74.22	71.70
CoOp	**82.69**	63.22	71.66
CoCoOp	80.47	71.69	75.83
VPT-s	73.32	73.21	73.16
VPT-deep	75.81	72.40	73.97
UPT	76.88	**75.57**	**76.15**

(b) ImageNet.

	Base	New	H
CLIP	72.43	68.14	70.22
CoOp	**76.47**	67.88	71.92
CoCoOp	75.98	70.43	73.10
VPT-s	74.47	69.13	71.70
VPT-deep	75.80	68.76	72.11
UPT	75.83	**70.80**	**73.23**

(c) Caltech101.

	Base	New	H
CLIP	96.84	94.00	95.40
CoOp	**98.00**	89.81	93.73
CoCoOp	97.96	93.81	95.84
VPT-s	97.47	93.80	95.60
VPT-deep	97.50	94.10	95.77
UPT	97.70	**95.63**	**96.14**

(d) OxfordPets.

	Base	New	H
CLIP	91.17	97.26	94.12
CoOp	93.67	95.29	94.47
CoCoOp	95.20	**97.69**	**96.43**
VPT-s	93.90	96.87	95.36
VPT-deep	94.33	95.50	94.91
UPT	**96.07**	97.60	96.32

(e) StanfordCars.

	Base	New	H
CLIP	63.37	74.89	68.65
CoOp	**78.12**	60.40	68.13
CoCoOp	70.49	73.59	**72.01**
VPT-s	66.00	74.23	69.88
VPT-deep	69.23	74.03	71.55
UPT	68.50	**75.37**	71.77

(f) Flowers102.

	Base	New	H
CLIP	72.08	**77.80**	74.83
CoOp	**97.60**	59.67	74.06
CoCoOp	94.87	71.75	**81.71**
VPT-s	75.83	75.73	75.78
VPT-deep	83.63	70.50	76.50
UPT	85.00	77.33	80.99

(g) Food101.

	Base	New	H
CLIP	90.10	91.22	90.66
CoOp	88.33	82.26	85.19
CoCoOp	90.70	91.29	90.99
VPT-s	90.17	90.97	90.56
VPT-deep	90.20	91.17	90.68
UPT	**90.72**	**92.00**	**91.35**

(h) FGVCAircraft.

	Base	New	H
CLIP	27.19	**36.29**	31.09
CoOp	**40.44**	22.30	28.75
CoCoOp	33.41	23.71	27.74
VPT-s	30.83	35.17	32.86
VPT-deep	33.40	35.17	**34.26**
UPT	32.76	36.10	34.53

(i) SUN397.

	Base	New	H
CLIP	69.36	75.35	72.23
CoOp	**80.60**	65.89	72.51
CoCoOp	79.74	76.86	78.27
VPT-s	75.40	77.27	76.32
VPT-deep	78.23	76.63	77.43
UPT	78.90	**78.56**	**78.73**

(j) DTD.

	Base	New	H
CLIP	53.24	59.90	56.37
CoOp	**79.44**	41.18	54.24
CoCoOp	77.01	56.00	64.85
VPT-s	55.27	57.16	56.20
VPT-deep	64.87	55.40	59.76
UPT	69.53	**62.13**	**65.63**

(k) EuroSAT.

	Base	New	H
CLIP	56.48	64.05	60.03
CoOp	**92.19**	54.74	68.69
CoCoOp	87.49	60.04	71.21
VPT-s	71.67	58.87	64.64
VPT-deep	66.70	60.67	63.54
UPT	73.57	**70.43**	**71.96**

(l) UCF101.

	Base	New	H
CLIP	70.53	**77.50**	73.85
CoOp	**84.69**	56.05	67.46
CoCoOp	82.33	73.45	**77.64**
VPT-s	75.60	76.10	75.85
VPT-deep	80.07	74.50	77.18
UPT	78.10	76.33	77.21

evaluation is conducted on the base and new classes separately. The experimental results are shown in Table 3.

Single-modal Baselines. We observe that previous single-modal baselines perform dramatically different on the base and new splits. In particular, the text prompt tuning method CoOp achieves the highest performance on base classes and poor performance on new classes. On the contrary, visual prompt tuning approaches VPT-shallow and VPT-deep obtain high accuracy on base classes, but low accuracy on new classes. Such results show the intrinsic discrepancy between single-modal text and visual prompt tuning methods. CoOp optimizes specifically for base classes but at the expense of generalization ability on new classes. The advanced text prompt tuning approach CoCoOp with the input-conditional design achieves the best base and new trade-off among the single-modal baselines.

Strong Generalizability of UPT. As shown in Table 3, UPT is more generalizable than baseline methods when taking into account both the base and new classes. As for base classes, UPT is better

Table 4: **Comparison results in the cross-dataset transfer setting**. Prompts applied to the 10 target datasets are learned from source ImageNet dataset. The **best** and **second best** methods are highlighted in red and orange, respectively.

	Source	Target										
	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	*Average*
CoOp	**71.51**	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55	63.88
CoCoOp	71.02	**94.43**	90.14	65.32	71.88	86.06	22.94	67.36	45.73	45.37	68.21	65.74
VPT-shallow	68.98	93.07	89.63	63.63	70.50	85.03	**24.01**	66.30	45.13	**45.56**	66.80	65.33
VPT-deep	70.57	90.33	88.50	57.87	63.83	76.90	21.93	63.10	42.13	40.63	64.53	61.85
UPT	70.86	93.31	**90.57**	**65.33**	**72.33**	**86.17**	24.57	**67.66**	45.67	44.94	**68.23**	**65.85**

than VPT but worse than CoOp. This is reasonable because UPT are jointly optimized on the text and visual modalities, and the visual modality branch is not specifically for base classes. For new classes, UPT has significantly improved performance. For instance, UPT obtains $+12.35/+2.36/+3.17$ gains for CoOp/VPT-shallow/VPT-deep. UPT even achieves $+1.35$ gains on new classes compared with the CLIP baseline without prompt learning. In summary, the experimental results under the base-to-new generalization setting show strong generalizability of UPT.

C.2 Cross-dataset Transfer

In the cross-dataset transfer setting, prompts learned from ImageNet are applied to ten other target datasets to evaluate the generalizability. The detailed results are presented in Table 4. We find VPT-shallow achieves higher accuracy than VPT-deep, and the *text* prompt tuning method CoCoOp outperforms *visual* prompt tuning approaches. On the source dataset, UPT obtains better performance than VPT but worse than CoOp. On target datasets, UPT obtains the best performance on six out of ten. The results on the cross-dataset transfer setting also verify that our proposed UPT is more generalizable than single-modal baselines.