Self-generated Cross-modal Prompt Tuning

Guiming Cao[1], Zonghan Wu[2], Huan Huo[1], Yuming Ou[1], and Guandong Xu[1,3] (✉)

[1] University of Technology Sydney, Sydney, Australia
Guiming.Cao@student.uts.edu.au, Guandong.Xu@uts.edu.au
[2] East China Normal University, Shanghai, China
[3] The Education University of Hong Kong, Hong Kong, China

Abstract. Training prompt tuning models on task-specific data is a common method for adapting vision-language model knowledge to image recognition downstream tasks. Despite recent advancements in prompt tuning, achieving superior generalization to heterogeneous images, across a wide range of visual characteristics in style, format, and source, remains a significant challenge. To this end, we propose a novel method, namely Self-generated Cross-modal Prompt tuning (SCP), which generates pseudo prompts by applying the frozen knowledge in both the initialization and optimization stages to guide training. Consequently, the model can be trained on available datasets while effectively generalizing to heterogeneous image data in a wide spectrum of textual classes and visual characteristics. Extensive experiments on four benchmarks indicate that our proposed SCP significantly outperforms well-known baselines in generalization performance across a broad spectrum of downstream tasks. Notably, our proposed SCP exhibits significant improvements in both Cross-Dataset and Domain-Shift Generalization, with performance gains of at least 3.63% and 11.71%, respectively. Our code is available at https://github.com/Ghosttimber/Academic.

Keywords: Computer Vision · Multi-Modal · Prompt Tuning.

1 Introduction

Recently developed vision-language models (VLMs) interpret and connect data across a variety of modalities, representing a significant leap forward from the traditional paradigm in multi-modal downstream tasks. Such a model, exemplified by CLIP [28], aligns image and text in a shared space, achieving superior generalization performance without task-specific training. Building upon CLIP, Context Optimization (CoOp) [43] prepends learnable tokens into the prompt of VLMs (learnable prompt), demonstrating improved generalization performance. This paradigm, termed prompt tuning, tailors VLMs to task-specific datasets, broadening the applicability of VLMs to a wide range of downstream tasks.

Expanding upon the foundation laid by CoOp, several subsequent approaches (KgCoOp [36], PromptSRC [16] and TCP [37]) have focused on unlocking the potential of prompt tuning in downstream tasks, retrieving the frozen knowledge

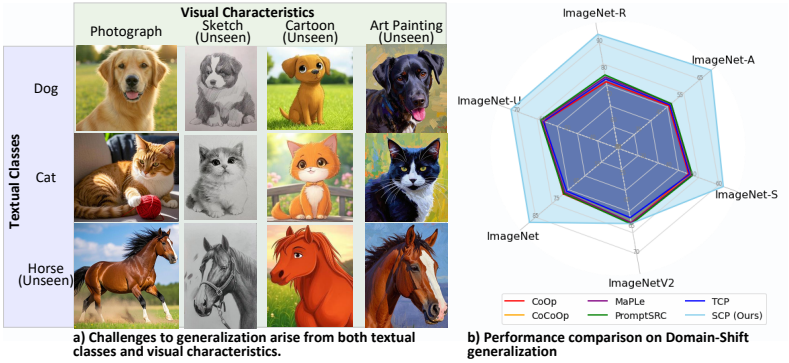


Fig. 1. Challenges addressed by our proposed SCP and its performance compared to the well-known baselines.

lost in training by establishing alignment between the prompted embedding (i.e. the embedding carrying the learnable token) and the frozen CLIP embedding. This mechanism facilitates the knowledge transfer from the training distribution to out-of-distribution, thereby mitigating overfitting to the unseen textual classes from the training data. As a result, further improved generalizability is attainable. Despite the advancement in prompt tuning, a remaining critical challenge stems from achieving superior generalization to heterogeneous images, which encompasses diverse visual characteristics such as style, format, and source, reflecting real-world scenarios as illustrated in Fig. 1(a). In other words, the challenge lies in finding a learnable prompt against heterogeneous images.

Motivated by the challenges, we have found that transforming a textual prompt into its counterpart in the image space results in pseudo prompts independent of visual characteristics. Consequently, incorporating such pseudo prompts as guidance in training enhances the ability to achieve superior generalization across a broader range of downstream tasks as illustrated in Fig. 1(b). Accordingly, based on the aforementioned observation, we propose a novel prompt tuning model agnostic to both textual classes and visual characteristics, and it is referred to as Self-generated Cross-modal Prompt Tuning (SCP).

In essence, SCP systematically leverages frozen CLIP knowledge in both the initialization and optimization phases. In the initialization, the optimal matching embeddings between the textual and visual prompt are selected by the frozen CLIP model and incorporated into learnable prompts representation. The main purpose is to mitigate the negative effects of biases arising from the specificity of textual class and visual characteristics. Consequently, robust learnable prompts can be attained to enhance generalization capability. We refer to this initialization strategy as Cascade Propagation Prompt Initialization (CPPI). As for the optimization, the learnable prompt is aligned with the pseudo prompt embedding (i.e. proxy representation) derived from frozen CLIP knowledge in a cross-modal manner. Note that the proxy representation holds a different role regarding the modality of the learnable prompt in the alignment. The proxy

representation aligned with the visual learnable prompt is generated from the textual prompt of frozen CLIP, serving as pseudo prompts in the image space. We found that it remains independent of visual characteristics. Clearly this enables visual learnable prompts to acquire a robust representation. Similarly, the proxy representation aligned with the textual learnable prompt comes from the visual prompt of frozen CLIP. It offers sufficient samples for the training in alignment while existing baselines rely on limited samples or hand-crafted samples. This optimization strategy is referred to as Self-generated Proxy Alignment (SPA).

In addition, building upon the cross-modal nature of SPA and following [27], we adopt the Euclidean loss in the alignment to reduce the modality gap between textual and visual representations to a desirable level. To mitigate the information loss [15] in the modality conversion of learnable prompts, we propose an Entropy-Regulation (ER) module, which, combined with the adoption of Euclidean loss, further improves the generalization performance.

In summary, this paper makes four major contributions:

- We propose a novel prompt tuning method (Self-generated Cross-modal Prompt tuning), that trains on limited readily available images, adapting prompt tuning to a much broader array of downstream applications.
- In our method, we take advantage of textual and visual knowledge from frozen CLIP into both the initialization and optimization process in a cross-modal manner, conducive to superior generalization capability.
- We introduce the Euclidean loss and the ER to improve the generalization capability, mitigating the challenge discussed by the existing research.
- Extensive experiments on four benchmarks clearly demonstrate that our proposed SCP significantly outperforms the well-known baselines in generalization capability across a wide range of downstream tasks.

2 Related Work

Vision Language Models. Recently, several methods, including BAN [17], Intra-Inter [10], and MCAN [38], adopt the VLMs with the attention-based framework, showing that the utilization of VLMs significantly improves the performance across a wide spectrum of downstream tasks. Subsequently, the methods (ViLBERT [22], LXMERT [32] and UNITER [2]) explore the potential of VLMs models based on BERT-like architectures, attaining further improvement. Methods proposed thereafter, namely CLIP [28] and ALIGN [14], are trained to align a considerable amount of web-scale image-text pair data in a multi-modal architecture, leaping forward the generalization capability to a new level. Meanwhile, this training mechanism has been widely adopted in image recognition [9, 41], object detection [8, 23, 39, 33], and segmentation [6, 20, 29].

Prompt Learning. As a new paradigm to leverage VLMs, prompt tuning manages to significantly address the challenges of CLIP arising from the fact that the hand-crafted text is insufficient for specialized tasks regarding training time and the sensitivity of prompt design. On the other hand, a recent prompt tuning

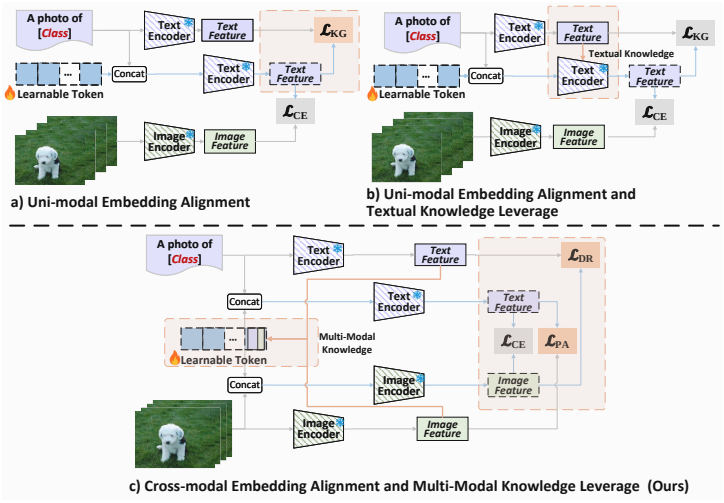


Fig. 2. Schematic architecture of two typical baselines and our proposed method. Compared to (a) and (b), (c) utilizes multi-modal knowledge of frozen CLIP and aligned prompted embedding with frozen CLIP embedding in a cross-modal manner to direct the learnable prompt to obtain a robust representation.

method, Distribution-Aware Prompts tuning (DAPT) [3], optimize the learnable prompt by minimising the intra-dispersion and maximizing the inter-dispersion to obtain better generalization performance. Alternatively, Read-only Prompt Optimization (RPO) [19] proposes a set of read-only prompts aiming to avoid the impact on the internal representation of CLIP by using masked attention. Another approach, Decoupled Prompt Tuning (DePT) [40], enhances generalization performance by isolating task-specific knowledge from the channels of feature representation and preserving the shared knowledge.

More recently, two mechanisms, i.e. the multi-modal architecture and the retrieval of frozen CLIP knowledge, are introduced to prompt tuning. Among MaPLe, RPO, DAPT and PromptSRC [16], MaPLe [15] is the early method to apply multi-modal architecture to utilise the knowledge in both text and image encoders. PromptSRC applies the architecture, and concurrently aligns its learnable prompt embedding with frozen CLIP knowledge, which is another pathway to improve the generalization. Alternatively, several methods improve generalization by only retrieving frozen CLIP knowledge, including ProGrad, KgCoOp and TCP. While ProGrad [44] only optimizes the prompt whose gradient is aligned (or non-conflicting) to the frozen CLIP knowledge. KgCoOp [36] explicitly addresses the gap between the embedding of learnable prompts and that of frozen CLIP as shown in Fig. 2(a). As an alternative, TCP further leverages the frozen CLIP knowledge by injecting its text embeddings into the encoder illustrated as Fig. 2(b), enhancing the generalization performance. Note that, these three methods only retrieve the textual knowledge from frozen CLIP. Besides, meth-

ods like PromptKD [21] and HPT [34] apply extra knowledge from fine-tuned models and large language models are not covered in this study.

3 Methodology

In this paper, we propose a novel method, referred to as SCP, to achieve superior generalization by leveraging the frozen CLIP knowledge in both the initialization and optimization stage of the learnable prompt as its schematic architecture shown in Fig. 2(c). Before delineating on SCP, to facilitate understanding, we first review the fundamental knowledge from the existing baseline framework.

3.1 Existing Baseline Framework

The existing baseline, such as CoOp, adopts CLIP for image recognition downstream tasks by prepending learnable tokens to the prompt context. Note that, the pre-trained encoders in CLIP, both image and text, convert the prompt context and the image sample into corresponding embeddings, which are then paired based on the contrastive loss to ensure optimal matching. In specific, the hand-crafted template with c class labels, e.g. "a photo of a $\{Class\}$", $Class : C \in \{1, 2, ...c\}$, is embedded into vectorized textual tokens $T = \{t_i\}_{i=1}^{c}$, and the b learnable tokens $P = \{p_i\}_{i=1}^{b}$ are initialized in text modality space. Then, the text encoder $\mathcal{B}(\cdot)$ interprets the combination of learnable tokens and vectorized textual tokens into the text embedding $W^p = \mathcal{B}([P, T]) = \{\mathrm{w}_i^p\}_{i=1}^{c}$. To infer P, the cosine similarity score $sim(\cdot)$ needs to be maximized. Equivalently, the contrastive loss, between the image embedding x and the prompted text embedding w_y^p, is calculated as

$$\mathcal{L}_{ce} = \frac{\exp(sim(\mathrm{x}, \mathrm{w}_y^p/\tau))}{\sum_{i=1}^{c} \exp(sim(\mathrm{x}, \mathrm{w}_i^p/\tau))}. \tag{1}$$

Here, τ refers to the temperature parameter.

To further unlock the potential of CLIP, MaPLe extends learnable prompts to the image side and takes advantage of multiple transformer blocks in the learning process. Specifically, the prompted image embedding is generated as $\mathrm{X}^p = \mathcal{V}([\widetilde{P}, Z])$. Here, $\widetilde{P} = \mathcal{F}(P)$ is the operation that converts learnable token to image modality space by a projection function $\mathcal{F}(\cdot)$, Z refers to the vectorized visual tokens. Following Maple [15], the operation for processing prompted text and image embeddings through J transformer layers is

$$W_{j+1}^p = \mathcal{B}_{j+1}([P_j, T_j]) \tag{2}$$

and

$$\mathrm{X}_{j+1}^p = \mathcal{V}_{j+1}([\widetilde{P_j}, Z_j]) = \mathcal{V}_{j+1}([\mathcal{F}(P_j), Z_j]), \tag{3}$$

respectively, where $j \in (0, 1, \cdots, J-1)$. Accordingly, the contrastive loss in MaPLe is calculated as

$$\mathcal{L}_{ce} = \frac{\exp(sim(\mathrm{x}^p, \mathrm{w}_y^p/\tau))}{\sum_{i=1}^{c} \exp(sim(\mathrm{x}^p, \mathrm{w}_i^p/\tau))}. \tag{4}$$

Alternatively, KgCoOp [36] proposed new loss function, i.e.

$$\mathcal{L}_{kg} = \left\| W^{frozen} - W^p \right\|_2^2, \tag{5}$$

to account for the distribution gap between the text embedding of frozen CLIP and that of prompted model. Therefore, KgCoOp has the final loss function as

$$\mathcal{L}_{Total} = \mathcal{L}_{ce} + \omega \mathcal{L}_{kg}, \tag{6}$$

where ω serves as a crucial weighting hyper-parameter.

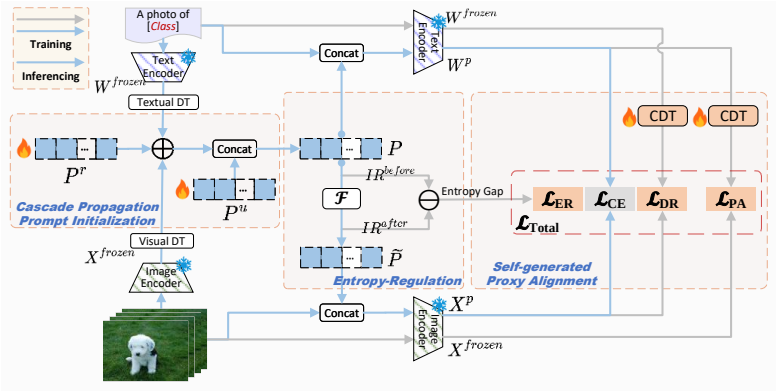


Fig. 3. The framework of Self-generated Cross-model Prompt tuning. Here, Textual DT and Visual DT perform intra-modal dimension transformation, while P^r and P^u refer to restricted tokens and unrestricted tokens, respectively. $\mathcal{F}$ refers to the projection function, and CDT refers to the cross-modal dimension transformation module. In addition, $\mathcal{L}_{CE}$ is the standard cross-entropy loss, $\mathcal{L}_{ER}$ is the proposed Entropy-Regulation constraint to minimize information during the projection function. $\mathcal{L}_{DR}$ and $\mathcal{L}_{PA}$ are the constraints for the proposed Self-generate Proxy Alignment to minimize the discrepancy between the corresponding prompted embedding and frozen CLIP embedding.

3.2 Self-generated Cross-modal Prompt Tuning

In the existing framework, the learnable prompts obtain task-specific knowledge to learn a representation space that potentially overfits the training data.

Clearly, it degrades the generalization capability when the data is not present in training. Recent baselines, such as KgCoOp, PromptSRC, and TCP, draw upon the frozen CLIP knowledge in intra-modal to retrieve the generalization capability of CLIP lost in such training. However, such a mechanism is insufficient to direct the learnable prompt to learn a representation that manages to generalize the heterogeneous images, resulting in a minor improvement in the related evaluation, i.e. Cross-Dataset generalization and Domain-Shift generalization. Accordingly, our proposed method, Self-generated Cross-modal Prompt tuning (SCP), taps the frozen CLIP knowledge in a cross-modal manner to generate the pseudo prompts. Consequently, robust learnable prompts are achieved by aligning pseudo prompts in training, leading to the superior generalization capability on the image recognition downstream task. It mainly consists of three major modules, as shown in Fig. 3, each of which is introduced in the sequel.

Cascade Propagation Prompt Initialization (CPPI). In initialization, CPPI introduces two learnable tokens (restricted tokens P^r and unrestricted tokens P^u) to dynamically leverage the frozen CLIP knowledge, allowing the learnable prompt to learn a robust representation for a wide range of scenarios. Specifically, the text and image are first encoded by the frozen CLIP encoder. Then, the optimal matching embeddings between the textual and visual prompts are selected by the frozen CLIP model, referred to as W^{frozen} and X^{frozen}. The resulting embeddings are projected by Dimension Transformation (DT) and further reshaped into the size of the restricted tokens $P^r \in \mathbb{R}^{C\times b\times 512}$. Note that, there are visual DT and textual DT for corresponding modality, and both of them are built with a two-layer bottleneck structure (Linear-ReLU-Linear), with the hidden layer reducing the input layer to 64 and 128 as middle dimensions for visual DT and textual DT, respectively. The output size of both DTs is raised to $b \times 512$, depending on learnable token length b.

After that, the first element-wise addition is conducted between the text embedding with a specific class, which has the highest similarity identified by the frozen CLIP, and the image embedding. Next, a second element-wise addition is performed between the modified embedding and P^r to generate the knowledge-based learnable tokens P^{kr}. Subsequently, P^{kr} and P^u are combined as $P = [P^{kr}, P^u]$. Finally, P undergoes the process outlined in Eqs. 2 and 3.

Entropy-Regulation. As discussed in [15], the direction of modality conversion in learnable tokens affects generalization performance, likely due to information loss. Therefore, we propose a module, namely Entropy-Regulation (ER), to mitigate the information loss and balance the generalization performance between the directions. In specific, both before and after the projection function, we employ the Fast Fourier Transform (FFT) on learnable tokens, Then the entropy of the learnable token is calculated as

$$\textbf{Entropy}(q_i) = -\textstyle\sum_{i=1}^{512} q_i \log_2 (q_i + \epsilon), \tag{7}$$

serving as the information recorder IR. Here, the input q_i represents the probability of occurrence of the i-th elements of each learnable token after the operation of FFT. ϵ is a small positive number. To account for the entropy gap between

before and after the conversion for every learnable token involved, including the unrestricted and restricted learnable tokens in each transformer layer, we have

$$\mathcal{L}_{ER} = IR^{before} - IR^{after}. \tag{8}$$

Self-generated Proxy Alignment. Self-generated Proxy Alignment (SPA) mechanism is introduced to generate a proxy representation by transforming the frozen CLIP knowledge into cross-modal space for the training. Specifically, the text embedding W^{frozen} from frozen CLIP is mapped into the image modality space by the Cross-modal Dimension Transformation (CDT), which has the same structure as DT with different settings in terms of the dimensions. Note that, the mapped embedding from text modality space potentially remains independent of image modality-specific information. The alignment between the mapped embedding and the prompted image embedding X^p manages to develop a robust representation for X^p, facilitating the improvement of generalization capability. This operation is referred to as Domain Resistance (DR).

In addition, SPA generates sufficient embedding samples from the frozen CLIP knowledge in image modality space to address the insufficient training process on text learnable prompts that occurred in existing methods. Such an operation is referred to as Prompt Argumentation (PA). The prompted text embedding W^p is mapped into the image modality space by the CDT. Then the alignment between the resulting embedding and image embeddings X^{frozen} from frozen CLIP provides sufficient training for W^p.

Moreover, SPA employs the Euclidean distance (i.e. non-contrastive loss) on training to reduce the intra-modal gap and consequently enhance the generalization performance further. Thus, to account for the distribution gap in the alignment of both DR and PA can be calculated as

$$\mathcal{L}_{DR} = \left\| CDT(W^{frozen}) - X^p \right\|_2^2, \tag{9}$$

$$\mathcal{L}_{PA} = \left\| CDT(W^p) - X^{frozen} \right\|_2^2, \tag{10}$$

to prompt the synergy from the marriage of transforming frozen knowledge in a cross-modal manner and non-contrastive loss. Here, $CDT(\cdot)$ performs cross-modal dimension transformation. Finally, we have the total loss as

$$\mathcal{L}_{Total} = \mathcal{L}_{CE} + \mathcal{L}_{ER} + \lambda_1 \mathcal{L}_{DR} + \lambda_2 \mathcal{L}_{PA}, \tag{11}$$

where λ_1 and λ_2 refer to the weighting factors associated with corresponding loss and are set to 5 and 7, respectively. Consequently, SPA empowers SCP to tap the potential of CLIP, conducive to superior generalization capability.

4 Experiments

In this section, we first assess our SCP effectiveness in generalization capability. In specific, we compare our proposed SCP with thirteen well-known prompt tuning methods on four benchmarks widely used in previous studies, including

Base-to-New class generalization, Cross-Dataset generalization, Domain-Shift generalization and Few-Shot classification. Then, the extended experiments are conducted to analyse the effectiveness of modules used in our SCP.

Table 1. Comparison of the Base-to-New generalization with 13 existing approaches. 'Uni', 'Multi', 'FC' denotes the 'uni-modal prompt tuning', 'multi-modal prompt tuning' and 'frozen CLIP knowledge', respectively. Δ refers to the gap between our proposed SCP and TCP.

Datasets	Sets	CLIP (IJCV22)	CoOp (CVPR22)	CoCoOp (ICCV23)	DAPT (ICCV23)	ProGrad (CVPR22)	ProDA (CVPR23)	KgCoOp (ICCV23)	RPO (ICLR23)	PLOT (ICCV23)	MaPLe (CVPR23)	DePT (CVPR24)	PromptSRC (ICCV23)	TCP (CVPR24)	SCP	Δ
		-	Uni	Uni+FC	Multi	Uni+FC	Uni	Uni+FC	Multi	Uni	Multi	Uni	Multi+FC	Uni+FC	Multi+FC	
Average	Base	69.34	82.38	80.47	83.18	82.48	81.56	80.73	81.13	83.98	82.28	83.62	84.12	84.13	**86.48**	+2.35
	New	74.22	67.96	71.69	69.27	70.75	72.30	73.60	75.00	71.72	75.14	75.04	75.02	75.36	**76.11**	+0.75
	HM	71.70	74.48	75.83	75.59	76.16	76.65	77.00	77.78	77.37	78.55	79.10	79.31	79.51	**80.97**	+1.46
ImageNet	Base	72.43	76.46	75.98	76.83	77.02	75.40	75.83	76.60	77.30	76.66	77.03	77.75	77.27	**82.34**	+5.07
	New	68.14	66.31	70.43	69.27	66.66	70.23	69.96	71.57	69.87	70.54	70.13	70.70	69.87	**76.91**	+7.04
	HM	70.22	71.02	73.10	72.85	71.46	72.72	72.78	74.00	73.40	73.47	73.42	74.06	73.38	**79.53**	+6.15
Caltech101	Base	96.84	97.80	97.96	97.83	98.02	98.27	97.72	97.97	**98.53**	97.74	98.30	98.13	98.23	97.42	-0.81
	New	94.00	93.27	93.81	93.07	93.89	93.23	94.39	94.37	92.80	94.36	94.60	93.90	**94.67**	89.85	-4.82
	HM	95.40	95.48	95.84	95.39	95.91	95.68	96.03	96.03	95.58	96.02	96.41	95.97	**96.42**	93.48	-2.94
Oxford Pets	Base	91.17	94.47	95.20	95.00	95.07	95.43	94.65	94.63	94.50	95.43	94.33	95.50	94.67	**98.11**	+3.44
	New	97.26	96.00	97.69	95.83	97.63	97.83	97.76	97.50	96.83	97.76	97.23	97.40	97.20	**99.40**	+2.20
	HM	94.12	95.23	96.43	95.41	96.33	96.62	96.18	96.05	95.65	96.58	95.76	96.44	95.92	**98.75**	+2.83
Standford Cars	Base	63.37	75.67	70.49	75.80	77.68	74.70	71.76	73.87	79.07	72.94	79.13	78.40	**80.80**	78.72	-2.08
	New	74.89	67.53	73.59	63.93	68.63	71.20	75.04	**75.53**	74.80	74.00	75.47	74.73	74.13	68.83	-5.30
	HM	68.65	71.37	72.01	69.36	72.88	72.91	73.36	74.69	76.88	73.47	77.26	75.52	**77.32**	73.44	-3.88
Flowers	Base	72.08	97.27	94.87	96.97	95.54	97.70	95.00	94.13	97.93	95.92	**98.00**	97.90	97.73	95.35	-2.38
	New	77.80	67.13	71.75	60.90	71.87	68.68	74.73	76.67	73.53	72.46	76.37	76.77	75.57	**77.49**	+1.92
	HM	74.83	79.44	81.71	74.81	82.03	80.66	83.65	84.50	83.99	82.56	85.84	**86.06**	85.23	85.50	+0.27
Food101	Base	90.10	89.37	90.70	90.37	90.37	90.30	90.50	90.33	89.80	90.71	90.50	90.63	90.57	**98.93**	+8.36
	New	91.22	88.77	91.29	91.30	89.59	88.57	91.70	90.83	91.37	92.05	91.60	91.50	91.37	**98.43**	+7.06
	HM	90.66	89.07	90.99	90.83	89.98	89.43	91.09	90.58	90.58	91.38	91.05	91.06	90.97	**98.68**	+7.71
FGVC Aircraft	Base	27.19	39.67	33.41	39.97	40.54	36.90	36.21	37.33	42.13	37.44	**43.20**	42.30	41.97	41.40	-0.57
	New	36.29	31.23	23.71	29.80	27.57	34.13	33.55	34.20	33.73	35.61	34.83	**36.97**	34.43	34.87	+0.44
	HM	31.09	34.95	27.74	34.14	32.82	35.46	34.83	35.70	37.46	36.50	38.57	**39.46**	37.83	37.86	+0.03
SUN397	Base	69.36	80.85	79.74	80.97	81.26	78.67	80.29	80.60	82.20	80.82	82.33	82.83	82.63	**90.73**	+8.10
	New	75.35	68.34	76.86	76.97	74.17	76.93	76.53	77.80	73.63	78.70	77.80	79.00	78.20	**85.43**	+7.23
	HM	72.23	74.07	78.27	78.92	77.55	77.79	78.36	79.18	77.68	79.75	80.00	80.87	80.35	**88.00**	+7.65
DTD	Base	53.24	79.97	77.01	82.23	77.35	80.67	77.55	76.70	81.97	80.36	82.20	82.60	82.77	**84.76**	+1.99
	New	59.90	48.60	56.00	54.23	52.35	56.48	54.99	62.13	43.80	59.18	59.13	57.50	58.07	**64.41**	+6.34
	HM	56.37	60.46	64.85	65.36	62.45	66.44	64.35	68.61	57.09	68.16	68.78	67.80	68.25	**73.20**	+4.95
EuroSAT	Base	56.48	90.10	87.49	94.73	90.11	83.90	85.64	86.63	93.70	94.07	89.03	92.40	91.63	**95.35**	+3.72
	New	64.05	53.00	60.04	50.33	60.89	66.00	64.34	68.97	62.67	73.23	71.07	68.43	**74.73**	62.20	-12.53
	HM	60.03	66.74	71.21	65.74	72.67	73.88	73.48	76.79	75.11	82.30	79.04	78.63	**82.32**	75.29	-7.03
UCF101	Base	70.53	84.53	82.33	84.3	84.33	85.23	82.89	83.67	86.60	83.00	85.80	86.93	87.13	**88.20**	+1.07
	New	77.50	67.37	73.45	76.33	74.94	71.97	76.67	75.43	75.90	78.66	77.23	78.33	**80.77**	79.39	-1.38
	HM	73.85	74.98	77.67	80.12	79.35	78.04	79.65	79.34	80.90	80.77	81.29	82.41	**83.83**	83.56	-0.27

4.1 Experiment Setup

Datasets. To implement the comparisons, as same as the well-known methods [43, 42, 15, 9, 41], eleven image datasets are used in Base-to-New class generalization, Cross-Dataset generalization and Few-Shot classification. Specifically, FGVCAircraft [24], Flowers102 [25], Food101 [1], OxfordPets [26] and StanfordCars [18] refer to fine-grained image datasets; Caltech101 [7] and ImageNet [5] refer to generic-object; DTD [4] refers to texture and EuroSAT [11] refers to satellite. While SUN397 [35] refers to the scene recognition task and UCF101 [31] refers to the action recognition task. As for Domain-Shift generalization, another four datasets from the ImageNet family with the shifted domain are included, i.e. ImageNet-Sketch [33], ImageNet-V2 [30], ImageNet-A [13] and ImageNet-R [12].
Baseline. In this paper, thirteen well-known prompt tuning methods are involved as the baseline, all of which do not leverage the external model but the

original CLIP as frozen knowledge. The thirteen well-known baselines with validated results are CLIP, CoOp, CoCoOp, DAPT, ProGrad, ProDA, KgCoOp, RPO, PLOT, MaPLe, DePT, PromptSRC, and TCP.

Training and Evaluation. In the comparisons, we use the CLIP with the backbone of ViT-B/16 in a 16-shot learning manner. The length b of both P^r and P^u learnable tokens is set to 2. Both P^r and P^u are first initialized with "a photo of Class" in the text modality space. The deep transformer layer depth J is set to 8. Besides, the SGD optimizer is used for training with a batch size of 4, a learning rate of 3.5e-3, and 30 training epochs. The evaluation metric is the average accuracy over 3 runs (random seeds 1, 2, and 3). Note that, in Cross-Dataset and Domain-Shift generalization, we train SCP with J modified to 3, training epoch set to 2, and a learning rate of 2.6e-3 to reduce computational cost. Similarly, in ablation experiments, we optimize the training epoch to 5.

4.2 Base-to-New Class Generalization

In the Base-to-New class generalization, every single dataset is divided into two groups (base and new). Each group contains mutually exclusive classes for training and evaluation, respectively. Then, this evaluation could assess if, after learning from the base group, the model could transfer the knowledge to the new group within one dataset. Thus, the evaluation metrics for generalization performance include the accuracies of both groups and their Harmonic Mean (HM). Note that, it can be seen from Tab. 1 that the recent significant progress in generalization performance occurred in methods that apply multi-modal prompt tuning and frozen CLIP knowledge in the training stage (i.e. PromptSRC and TCP). Thus, we compare our proposed SCP with the well-known baselines from the perspectives regarding these two prompt tuning mechanisms in the sequel.

Multi-Modal Prompt Tuning. It can be observed in Tab. 1 that the best average performances among the 4 multi-modal prompt tuning baselines are 84.12% (PromptSRC), 75.12% (MaPLe), and 79.31% (PromptSRC) for Base, New, and HM, respectively. In comparison, our proposed SCP achieves 86.48%, 76.11%, and 80.97%, respectively. The result represents a significant improvement of 2.36%, 0.99%, and 1.66%, respectively, surpassing the best generalization performance provided by the multi-modal prompt tuning baselines.

Prompt Tuning with frozen CLIP knowledge. For prompt tuning that leverages frozen CLIP knowledge, Tab. 1 clearly demonstrates that the TCP performs the best among the 5 baselines (CoCoOp, ProGrad, KgCoOp, PromptSRC, and TCP), achieving 84.13%, 75.36% and 79.51% for Base, New and HM metrics, respectively. In fact, it can be clearly seen that TCP outperforms all the other 12 listed baselines. In contrast, our proposed SCP achieves 86.48%, 76.11%, and 80.97%, respectively, representing an improvement of 2.35%, 0.75% and 1.46%, respectively. Note that, for all three metrics, our proposed SCP obtains the best generalization performance in 5/11 datasets (ImageNet, DTD, Food101, OxfordPets and SUN397). Furthermore, in Food101 and OxfordPets datasets, SCP achieves outstanding generalization performance. For example, in terms of *HM*, Food101 and OxfordPets achieve 98.68% and 98.75%, respectively,

indicating a significant improvement of 7.30% and 2.13% over the best baseline performance of 91.38% (MaPLe) and 96.62% (ProDA) for these specific datasets.

Overall, through the extensive comparison of the Base-to-New class generalization in Tab. 1, it is clearly illustrated our SCP consistently outperforms the thirteen baselines regarding the generalization performance. It shows that, with the aid of frozen CLIP knowledge in a cross-modal manner, SCP shows superior generalization capability in a wide range of image recognition downstream tasks.

4.3 Cross-Dataset Generalization

Table 2. Comparison of Cross-Dataset generalization. Δ refers to the gap between our proposed SCP and TCP.

	Source	Target										
	ImageNet	Caltech	Pets	Cars	Flowers	Food	Aircraft	SUN	DTD	EuroSAT	UCF	Avg.
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
MaPLe	70.72	93.53	90.49	65.57	**72.23**	86.20	24.74	67.01	46.49	48.06	68.69	66.30
PromptSRC	71.27	93.60	90.25	**65.70**	70.25	86.15	23.90	67.10	46.87	45.50	68.75	65.81
TCP	69.88	94.25	90.46	65.24	71.88	86.78	24.99	67.12	45.00	44.67	68.10	65.85
SCP	**87.97**	93.67	**93.21**	60.74	70.51	**96.93**	**27.75**	**82.98**	**52.88**	**50.05**	**70.52**	**69.93**
Δ	+18.09	-0.58	+2.75	-4.50	-1.37	+10.15	+2.76	+15.86	+7.88	+5.38	+2.42	+4.08

For Cross-Dataset generalization, the model is trained on ImageNet as the source dataset and evaluated on 10 target datasets that potentially contain heterogeneous images not presented in ImageNet. Compared to the Base-to-New class generalization, this evaluation imposes stricter standards for the generalization capability regarding domain generalization. However, it can be seen from Tab.2 that SCP consistently outperforms the baselines regarding the generalization performance. Note that these baselines excel in the Base-to-New class generalization. Specifically, significant performance advantages are achieved on the source dataset (ImageNet) and 7 out of 10 target datasets. Notably, it can be calculated that SCP obtains significant improvements of 16.46%, 10.15%, 15.62%, and 6.01% over existing best results on the source dataset and target datasets (Food, SUN and DTD), respectively. Clearly, this analysis indicates that SCP achieves a significant generalization performance improvement for the scenario containing heterogeneous images that are not present in the training.

4.4 Domain-Shift Generalization

In the Domain-Shift generalization, in contrast to Cross-Dataset generalization, the model is still trained on ImageNet but evaluated on the dataset with the shifted domain from ImageNet. Note that, this evaluation focuses on assessing the domain generalization capability of the model while all classes have been included in the training. As shown in Tab.3, SCP significantly outperforms the

Table 3. Comparison of Domain-Shift generalization. Δ refers to the performance gap between our proposed SCP and TCP.

	Source	Target				
	ImageNet	-V2	-Sketch	-A	-R	Average
CoOp	71.51	64.20	47.99	49.71	75.21	59.28
CoCoOp	71.02	64.07	48.75	50.63	76.18	59.91
MaPLe	70.72	64.07	49.15	50.90	76.98	60.28
PromptSRC	71.27	**64.35**	49.55	50.90	77.80	60.65
TCP	69.88	63.14	48.39	50.22	76.22	59.49
SCP	**87.97**	64.25	**61.73**	**70.28**	**93.18**	**72.36**
Δ	+18.10	+1.12	+13.34	+20.07	+16.95	+12.87

existing best baseline in 3 out of 4 target datasets and is only marginally inferior to the best result in the remaining one (ImageNet-V2). Table. 3 clearly shows the superiority of the SCP on Domain-Shift generalization. Significantly, the striking improvement suggests that SCP offer a novel pathway to advance prompt tuning for recognizing heterogeneous images in the downstream application.

Table 4. Comparison of Few-Shot classification with 4-shot samples. Δ refers to the gap between our proposed SCP and TCP.

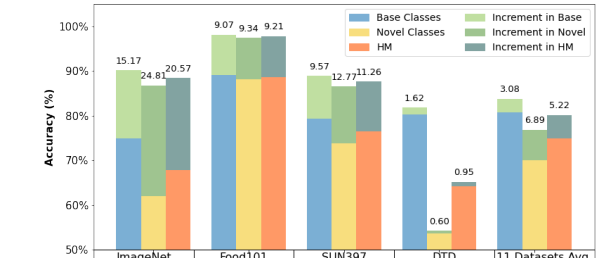
Datasets	CoOp	CoCoOp	ProGrad	KgCoOp	MaPLe	DAPT	PLOT	PromptSRC	TCP	SCP	Δ
ImageNet	69.37	70.55	70.21	70.19	70.67	70.80	70.40	70.80	70.48	**84.32**	+13.84
Caltech101	94.44	94.98	94.93	94.65	94.30	94.23	**95.13**	94.77	95.00	94.62	-0.38
OxfordPets	91.30	93.01	93.21	93.20	92.05	92.17	92.55	93.23	91.90	**96.38**	+4.48
StandfordCars	72.73	69.10	71.75	71.98	68.70	74.40	74.93	71.83	**76.30**	74.58	-1.72
Flowers	91.14	82.56	89.98	90.69	80.80	92.37	92.93	91.31	**94.40**	84.46	-9.94
Food101	82.58	86.64	85.77	86.59	86.90	83.60	86.46	86.06	85.30	**98.49**	+13.19
FGVCAircraft	33.18	30.87	32.93	32.47	29.03	32.47	35.29	32.80	36.20	**38.64**	+2.44
SUN397	70.13	70.50	71.17	71.79	71.47	72.20	70.42	72.80	72.11	**85.84**	+13.73
DTD	58.57	54.79	57.72	58.31	54.73	61.37	62.43	60.64	63.97	**66.19**	+2.22
EuroSAT	68.62	63.83	70.84	71.06	54.87	72.73	**80.70**	75.02	77.43	80.40	+2.97
UCF101	77.41	74.99	77.82	78.40	73.70	79.40	79.76	79.35	80.83	**83.12**	+2.29
Avg.	73.59	71.98	74.21	74.48	70.66	75.07	76.45	75.33	76.72	**80.64**	+3.92

4.5 Few-shot Classification

Having evaluated the generalization capability of our proposed SCP with 16-shot training, Tab. 4 presents a similar evaluation for a more challenging scenario where fewer samples are available. In specific, the models are trained on 11 datasets with a 4-shot labelled source image and evaluation is conducted within the same class space. It can be observed that the proposed SCP outperforms all the baselines in average performance, and significantly so in 7 out of 11 datasets. This clearly indicates that SCP also enjoys superb generalization capability for image recognition downstream tasks with few-shot learning conditions.

4.6 Ablation analysis

In ablation analysis, we evaluate the effectiveness of various modules introduced by the proposed SCP regarding the performance in Base-to-New class generalization.



	ImageNet	Food101	SUN397	DTD	11 Datasets Avg.
HM w/o CPPG	67.88%	88.62%	76.45%	64.29%	74.97%
HM	88.45%	97.82%	87.70%	65.24%	80.19%

Fig. 4. Analysis of Cascade Propagation Prompt Initialization.

Effect of Cascade Propagation Prompt Initialization (CPPI). Fig. 4 examines the impact of CPPI on 4 datasets (ImageNet, Food101, SUN397 and DTD) and average performance across 11 datasets. For instance, the average Base, New and HM results with CPPI module improve the performance from 80.73%, 69.98% and 74.97% to 83.81%, 76.87% and 80.19%, respectively. This result clearly suggests that CPPI effectively embeds the frozen CLIP knowledge into the learnable prompts, addressing overfitting to the training data.

Table 5. Analysis on Entropy-Regulation. Δ refers to the gap between different directions of modality conversion within the same model.

Model	Prompt Proj.	ImageNet	Caltech101	OxfordPets	StandfordCars	Flowers	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	Average
1. MaPLe	$P \rightarrow \widetilde{P}$	72.63	96.12	96.67	72.82	81.55	90.83	29.79	79.03	61.90	80.93	80.40	
	$\widetilde{P} \rightarrow P$	73.23	96.01	96.68	72.87	82.45	91.08	35.96	79.21	65.74	63.30	81.16	
	$\lvert\Delta\rvert$	0.60	0.11	0.01	0.05	0.90	0.25	6.17	0.18	3.84	17.63	0.76	2.77
2. MaPLe w/ ER	$P \rightarrow \widetilde{P}$	72.58	96.58	96.67	72.31	81.98	91.13	29.16	78.90	61.42	80.60	80.59	
	$\widetilde{P} \rightarrow P$	73.30	95.95	96.57	72.59	82.00	91.25	29.46	79.46	66.87	61.07	80.18	
	$\lvert\Delta\rvert$	0.72	0.63	0.10	0.28	0.02	0.12	0.30	0.56	5.45	19.53	0.41	2.56
3. SCP w/o ER	$P \rightarrow \widetilde{P}$	90.53	90.12	97.82	70.23	80.57	98.80	37.81	89.66	67.86	72.70	82.96	
	$\widetilde{P} \rightarrow P$	89.26	92.35	97.03	70.61	84.32	96.13	37.28	90.87	70.30	75.32	81.62	
	$\lvert\Delta\rvert$	1.27	2.23	0.79	0.38	3.75	2.67	0.53	1.21	2.44	2.62	1.34	1.75
4. SCP	$P \rightarrow \widetilde{P}$	90.02	91.49	97.53	70.70	82.05	99.21	37.81	89.82	68.41	74.46	82.74	
	$\widetilde{P} \rightarrow P$	90.13	91.70	97.70	70.99	81.63	98.51	34.92	90.92	71.50	72.16	80.93	
	$\lvert\Delta\rvert$	0.11	0.21	0.17	0.29	0.42	0.70	2.89	1.10	3.09	2.30	1.81	1.19

Effect of Entropy-Regulation Module. To verify the effectiveness of the Entropy-Regulation (ER) for modality conversion, we disentangle the ER in our SCP and leverage it on an existing multi-modal method, MaPLe. A modified SCP and MaPLe are employed in this study to minimize the effect of the script. In Tab. 5, The absolute value, $|\Delta|$, demonstrates the performance gap between

converting text to image and the reverse process. The average value across 11 datasets indicates that the groups employing the ER exhibit a reduced disparity across different directions of modality conversion for both the SCP and MaPLe. Besides, we compare ER with a non-ER group and a modified energy-based regulation version instead of entropy-based as illustrated in Tab. 6. The results shows that entropy-based ER effectively enhances generalization performance.

Table 6: Analysis on ER for various mechanisms.

Regulation	Base	New	HM
w/o ER	83.91	76.95	80.28
Energy-based	54.13	52.73	53.42
Entropy-based	84.41	77.45	80.78

Table 7: Analysis on Self-generated Proxy Alignment.

Domain Resist.	Prompt Augment.	Base	New	HM
-	-	81.91	73.04	77.22
✓	-	83.46	77.11	80.16
-	✓	81.19	72.08	76.37
✓	✓	83.81	76.87	80.19

Effect of Self-generated Proxy Alignment. Table. 7 explore all four combinations in terms of the adoption of two modules (Domain Resistance and Prompt Augmentation). It can be clearly seen that solely employing the DR appears to enhance generalization capability, in contrast to the less favourable performance of only using PA. However, the combination of both DR and PA achieves higher generalization performance in the Base and HM than others.

Table 8. Analysis on the initialization templates.

Templates	Base	New	HM
“a photo of a"	83.81	76.87	80.19
“this is a picture of”	83.89	76.94	80.26
“X X X X"	84.79	78.67	81.62

Effect of different initialization templates. We analyze the effect of the initialization template for learnable tokens by comparing three different templates, “a photo of a {}", “this is a picture of {}" and random initialization template “X X X X {}". Table. 8 shows a marginal difference between the two handcrafted templates. More importantly, the random initialization template further empowers SCP to achieve superior generalizability, although it necessitates much stricter standards for the training of learnable prompts. This analysis indicates that SCP is robust against the effect associated with the initialization template.

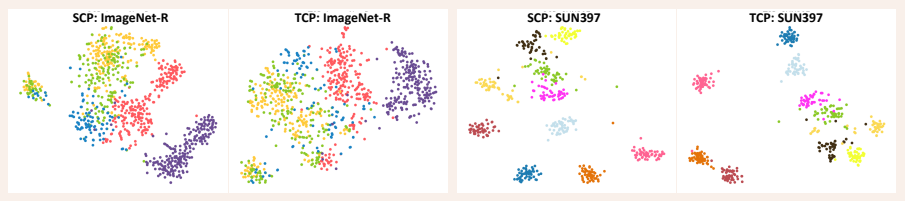


Fig. 5. t-SNE plots of image embeddings in SCP and TCP.

5 Visualization

To facilitate understanding of the superior generalization performance achieved by SCP, in Fig. 5, the image embeddings prepended with the learnable token from our proposed SCP and TCP are visualized in t-SNE, while colours differentiate classes in the dataset. SCP display more distinct discriminative clustering than those from TCP in both Cross-Dataset and Domain-Shift Generalization.

6 Conclusion

Prompt tuning effectively adapts fundamental vision-language models to the image recognition downstream tasks. However, prompt tuning methods suffer from overfitting to training data, limiting the potential improvement of generalization capability, significantly so for heterogeneous images that differ from the training data. In this study, we address this challenge from the perspective of finding a robust representation of learnable prompts. To obtain such a representation, we propose to train the learnable prompt by aligning pseudo prompts generated from self-knowledge in a cross-modal manner. The resulting representation is able to adapt to various textual classes and visual characteristics, consequently enhancing the generalization capability. Extensive experiments on four benchmarks clearly show that SCP outperforms well-known baselines in generalization performance. In particular, our proposed SCP achieves a striking improvement of generalization performance on Cross-Dataset and Domain-Shift generalization, revealing its superiority of generalization capability across a wide range of scenarios where the heterogeneous images are not present in the training data. Crucially, this paper provides a new avenue for expanding the applicability of prompt tuning to a broader spectrum of downstream applications.

References

1. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101 – mining discriminative components with random forests. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. pp. 446–461. Springer International Publishing, Cham (2014)
2. Chen, Y.C., Li, L., Yu, L., Kholy, A.E., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning (2020)

3. Cho, E., Kim, J., Kim, H.J.: Distribution-aware prompt tuning for vision-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 22004–22013 (October 2023)
4. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild (2013)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009)
6. Ding, J., Xue, N., Xia, G.S., Dai, D.: Decoupling zero-shot semantic segmentation (2022)
7. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In: 2004 Conference on Computer Vision and Pattern Recognition Workshop. pp. 178–178 (2004). https://doi.org/10.1109/CVPR.2004.383
8. Feng, C., Zhong, Y., Jie, Z., Chu, X., Ren, H., Wei, X., Xie, W., Ma, L.: Promptdet: Towards open-vocabulary detection using uncurated images (2022)
9. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters (2021)
10. Gao, P., Lu, J., Li, H., Mottaghi, R., Kembhavi, A.: Container: Context aggregation network (2021)
11. Helber, P., Bischke, B., Dengel, A., Borth, D.: Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification (2019)
12. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., Gilmer, J.: The many faces of robustness: A critical analysis of out-of-distribution generalization (2021)
13. Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.: Natural adversarial examples (2021)
14. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q.V., Sung, Y., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision (2021)
15. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19113–19122 (2023)
16. Khattak, M.U., Wasim, S.T., Naseer, M., Khan, S., Yang, M.H., Khan, F.S.: Self-regulating prompts: Foundational model adaptation without forgetting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15190–15200 (October 2023)
17. Kim, J.H., Jun, J., Zhang, B.T.: Bilinear attention networks (2018)
18. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: 2013 IEEE International Conference on Computer Vision Workshops. pp. 554–561 (2013). https://doi.org/10.1109/ICCVW.2013.77
19. Lee, D., Song, S., Suh, J., Choi, J., Lee, S., Kim, H.J.: Read-only prompt optimization for vision-language few-shot learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1401–1411 (2023)
20. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven semantic segmentation (2022)
21. Li, Z., Li, X., Fu, X., Zhang, X., Wang, W., Chen, S., Yang, J.: Promptkd: Unsupervised prompt distillation for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26617–26626 (2024)

22. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks (2019)
23. Maaz, M., Rasheed, H., Khan, S., Khan, F.S., Anwer, R.M., Yang, M.H.: Class-agnostic object detection with multi-modal transformer (2022)
24. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft (2013)
25. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing. pp. 722–729 (2008). https://doi.org/10.1109/ICVGIP.2008.47
26. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.V.: Cats and dogs. 2012 IEEE Conference on Computer Vision and Pattern Recognition pp. 3498–3505 (2012), https://api.semanticscholar.org/CorpusID:383200
27. Qian, Q., Xu, Y., Hu, J.: Intra-modal proxy learning for zero-shot visual categorization with clip. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 25461–25474. Curran Associates, Inc. (2023)
28. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021)
29. Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., Lu, J.: Denseclip: Language-guided dense prediction with context-aware prompting (2022)
30. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do imagenet classifiers generalize to imagenet? (2019)
31. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild (2012)
32. Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers (2019)
33. Wang, H., Ge, S., Xing, E.P., Lipton, Z.C.: Learning robust global representations by penalizing local predictive power (2019)
34. Wang, Y., Jiang, X., Cheng, D., Li, D., Zhao, C.: Learning hierarchical prompt with structured linguistic knowledge for vision-language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 5749–5757 (2024)
35. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 3485–3492 (2010)
36. Yao, H., Zhang, R., Xu, C.: Visual-language prompt tuning with knowledge-guided context optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6757–6767 (June 2023)
37. Yao, H., Zhang, R., Xu, C.: Tcp:textual-based class-aware prompt tuning for visual-language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 23438–23448 (June 2024)
38. Yu, Z., Yu, J., Cui, Y., Tao, D., Tian, Q.: Deep modular co-attention networks for visual question answering (2019)
39. Zang, Y., Li, W., Zhou, K., Huang, C., Loy, C.C.: Open-Vocabulary DETR with Conditional Matching, p. 106–122. Springer Nature Switzerland (2022)
40. Zhang, J., Wu, S., Gao, L., Shen, H.T., Song, J.: Dept: Decoupled prompt tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12924–12933 (June 2024)
41. Zhang, R., Fang, R., Zhang, W., Gao, P., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-adapter: Training-free clip-adapter for better vision-language modeling (2021)

42. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models (2022)
43. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (Jul 2022)
44. Zhu, B., Niu, Y., Han, Y., Wu, Y., Zhang, H.: Prompt-aligned gradient for prompt tuning (2024)