Cascade Prompt Learning for Vision-Language Model Adaptation
Supplementary Material

Ge Wu[1†], Xin Zhang[1†], Zheng Li[1], Zhaowei Chen[3],
Jiajun Liang[3], Jian Yang[1], Xiang Li[1,2*]

[1] VCIP, CS, Nankai University
[2] NKIARI, Shenzhen Futian
[3] Megvii Technology
gewu.nku@gmail.com, {zhasion, zhengli97}@mail.nankai.edu.cn,
{csjyang, xiang.li.implus}@nankai.edu.cn,
{chenzhaowei, liangjiajun}@megvii.com

A Additional ablation studies

Impact of training epoch for the first phase: Fig. 1 (left) shows the impact of training epochs in the first stage on CasPL performance with the DTD dataset. The accuracy of the base class remains stable with increasing epochs, while the accuracy of the novel class decreases after 20 epochs.
Distillation temperature of learning boosting prompts: The temperature hyperparameter regulates the softness of the distributions. Therefore, in Fig. 1 right, we examine the influence of employing different temperature hyperparameters to train boosting prompts in the first stage and then fine-tuning adapter prompts in the second stage, specifically on the DTD dataset. According to the results, HM demonstrates the best performance when the temperature is set to 1. Hence, the temperature hyperparameter is default set to 1.

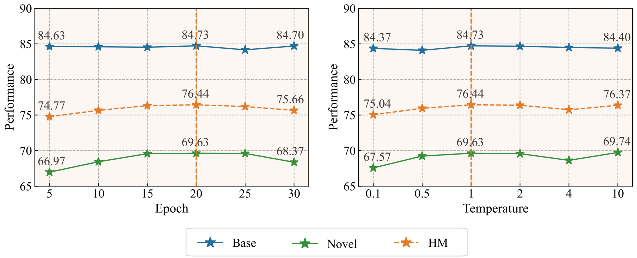


Fig. 1: Ablation study on the number of training epochs for the first phase (left) and the choice of temperature hyperparameter in Eq. (2) (right), based on the DTD dataset.

[†] Equal contributions. Work is done when Ge Wu is an intern at Megvii Technology.
[*] Corresponding author.

Table 1: The accuracy of zero-shot inference on domain generalization by CLIP (ViT-B/16) with adding boosting prompts. Boosting prompts can assist CLIP in enhancing domain generalization performance.

Method	Source	Target		
	ImageNet	ImageNet-V2	ImageNet-S	ImageNet-R
CLIP	66.73	60.83	46.15	73.96
+ **boosting**	70.40 **(+3.67)**	63.30 **(+2.47)**	47.70 **(+1.55)**	75.30 **(+1.34)**

Table 2: Ablation study on the HM results of boosting prompts trained with varying amounts of unlabeled images per class from the DTD dataset. ("Full" indicates the utilization of the entire unlabeled dataset.) Utilizing more unlabeled data enables the boosting prompt to acquire more domain-general knowledge.

Number	1	2	4	8	16	32	Full
HM	62.68	70.40	72.31	74.35	74.91	75.40	76.44

CLIP with boosting prompts for zero-shot inference: Table 1 investigates the efficacy of integrating boosting prompts into CLIP for zero-shot inference. It demonstrates the accuracy improvement in domain generalization for CLIP (ViT-B/16)+ boosting prompt on both source and target datasets. However, solely using boosting prompts is less effective compared to our two-stage CasPL, as shown by the comparison with Table 1. This highlights the distinct roles played by the boosting prompts and the adapting prompts in our proposed framework.

Unsupervised training of boosting prompts using partial data: This section investigates the impact of training boosting prompts with varying quantities of data on the outcomes of CasPL. Table 2 presents the DTD dataset's corresponding HM values for different quantities. It is observed that, with an increase in the number of instances per category, the performance metric exhibits an overall upward trend, and PromptSRC +CasPL outperforms best through training on the entire dataset. Notably, when the instances per class are four or more, the HM of PromptSRC +CasPL (≥72.31%) exceeds that of PromptSRC HM (71.75%), underscoring the effectiveness of boosting prompts.

B Additional implementation details

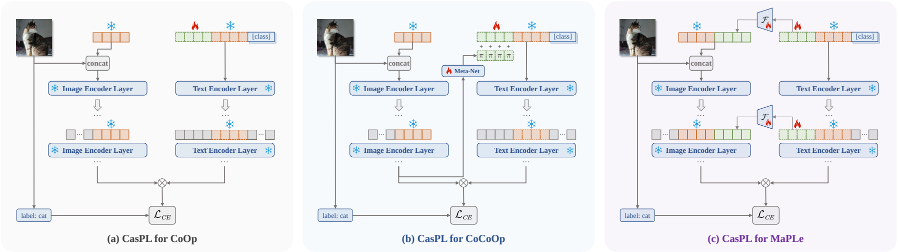
B.1 Boosting prompt phase

General training details: For the first phase of CasPL, we train the boosting prompts with a layer depth of 12, prompt length of 8, and a learning rate of 0.0025 using the SGD optimizer for 20 epochs. All learnable prompts are initialized with a normal distribution. To streamline the training of the boosting prompts on ImageNet, we utilize 8 NVIDIA 3090 GPUs, while all other experiments are conducted on a single NVIDIA 3090.

Text templates for senior teacher CLIP Drawing from previous findings [25], we utilize diverse prompt templates tailored to different datasets, aiming to aug-

Table 3: Text template utilized by senior CLIP teacher for different datasets.

Dataset	Text template
OxfordPets	" a photo of a [class], a type of pet. "
Flowers102	" a photo of a [class], a type of flower. "
Food101	" a photo of [class], a type of food. "
FGVC Aircraft	" a photo of a [class], a type of aircraft. "
DTD	" [class] texture. "
EuroSAT	" a centered satellite photo of [class]. "
UCF101	" a photo of a person doing [class]. "
other datasets	" a photo of a [class]. "



(a) CasPL for CoOp

(b) CasPL for CoCoOp

(c) CasPL for MaPLe

Fig. 2: The detail of CasPL for previous methods. **(a)** CoOp [65] employs multiple layers of text-image boosting prompts and a single layer of text adapting prompts. **(b)** CoCoOp [64] utilizes multiple layers of text-image boosting prompts and a single layer of modal blending adapting prompts. **(c)** MaPLe [24] uses multiple layers of text-image boosting prompts and multiple layers of multi-modal adapting prompts.

ment the senior CLIP's text representation ability and enhance the boosting prompts' distillation effect. Table 3 presents the templates for each dataset.

B.2 Adapting prompt phase

Base-to-Novel generalization: The training details of each of the previous methods on this task are shown in Table 4. Various prior approaches based on prompt learning exhibit differences in the specifics of their implementation. CoOp [65] employs single learnable text prompts (see Fig. 2 (a)). CoCoOp [64] combines image features with learnable text prompts to obtain multi-modal information (see Fig. 2 (b)). MaPLe [24] utilizes multi-layer learnable text prompts and image prompts generated from the text prompts (see Fig. 2 (c)). Specific details of PromptSRC are elaborated in Fig. 2.
Domain generalization: Following the previous method in this task [24], we adjust the parameters for MaPLe, specifically setting its optimizer's learning rate to 0.0026 and establishing the training epoch at 2.
Few-shot experiments: Following the methodology from previous work [25], we set the training epoch for PromptSRC at 50, keeping the other configurations consistent with those outlined in Table 4. The main text features comparison curves, while additional numerical results are available in Table 5.

Table 4: Training settings for base-to-novel generalization task.

	CoOp	CoCoOp	MaPLe	PromptSRC
Vision Prompt Length	-	-	8	8
Text Prompt Length	8	8	8	8
Prompt Layer	1	1	12	12
Optimizer	SGD	SGD	SGD	SGD
Learning Rate	0.002	0.002	0.0035	0.0025
Epoch	50	10	5	20

Compare with un-/weakly-supervised methods: In this experiment, the CLIP zero-shot method utilizes simple templates as the text input, and the numerical results are derived from the official UPL [18] code. To ensure a fair comparison, the three training strategies in ENCLIP [40] are implemented based on the PromptSRC pipeline [25]. Few-pseudo labels (FPL) utilizes 16 pseudo labels per novel class and 16 labeled data per base class. Iterative Refinement of FPL (IFPL) utilizes the same training data as FPL but involves multiple iterations. The labels are recalculated in each iteration, and the prompt is reinitialized. Grow and Refine Iteratively Pseudolabels (GRIP) gradually increases the number of unlabeled datasets compared to IFPL (with a maximum limit of 16 per class in our implementation).

Table 5: The performance of CasPL (built on PromptSRC) compared to other methods in the few-shot setting. Results across various few-shot setups demonstrate CasPL's ability to enhance model performance.

Dataset	Method	1 shot	2 shots	4 shots	8 shots	16 shots
ImageNet	Linear probe CLIP	32.13	44.88	54.85	62.23	67.31
	CoOp	66.33	67.07	68.73	70.63	71.87
	CoCoOp	69.43	69.78	70.39	70.63	70.83
	MaPLe	62.67	65.10	67.70	70.30	72.33
	PromptSRC	68.13	69.77	71.07	72.33	73.17
	CasPL **(Ours)**	68.73	70.07	71.43	72.87	74.20
Caltech101	Linear probe CLIP	79.88	89.01	92.05	93.41	95.43
	CoOp	92.60	93.07	94.40	94.37	95.57
	CoCoOp	93.83	94.82	94.98	95.04	95.16
	MaPLe	92.57	93.97	94.43	95.20	96.00
	PromptSRC	93.67	94.53	95.27	95.67	96.07
	CasPL **(Ours)**	93.97	95.20	96.10	96.23	96.80
DTD	Linear probe CLIP	34.59	40.76	55.71	63.46	69.96
	CoOp	50.23	53.60	58.70	64.77	69.87
	CoCoOp	48.54	52.17	55.04	58.89	63.04
	MaPLe	52.13	55.50	61.00	66.50	71.33
	PromptSRC	56.23	59.97	65.53	69.87	72.73
	CasPL **(Ours)**	62.63	63.67	69.07	71.00	75.13
EuroSAT	Linear probe CLIP	49.23	61.98	77.09	84.43	87.21
	CoOp	54.93	65.17	70.80	78.07	84.93
	CoCoOp	55.33	46.74	65.56	68.21	73.32
	MaPLe	71.80	78.30	84.50	87.73	92.33
	PromptSRC	73.13	79.37	86.30	88.80	92.43
	CasPL **(Ours)**	83.40	86.53	91.07	91.07	94.17
StanfordCars	Linear probe CLIP	35.66	50.28	63.38	73.67	80.44
	CoOp	67.43	70.50	74.47	79.30	83.07
	CoCoOp	67.22	68.37	69.39	70.44	71.57
	MaPLe	66.60	71.60	75.30	79.47	83.57
	PromptSRC	69.40	73.40	77.13	80.97	83.83
	CasPL **(Ours)**	72.80	77.23	80.03	83.30	86.70
Flowers102	Linear probe CLIP	69.74	85.07	92.02	96.10	97.37
	CoOp	77.53	87.33	92.17	94.97	97.07
	CoCoOp	72.08	75.79	78.40	84.30	87.84
	MaPLe	83.30	88.93	92.67	95.80	97.00
	PromptSRC	85.93	91.17	93.87	96.27	97.60
	CasPL **(Ours)**	90.33	94.17	95.53	97.20	98.30
FGVCAircraft	Linear probe CLIP	19.61	26.41	32.33	39.35	45.36
	CoOp	21.37	26.20	30.83	39.00	43.40
	CoCoOp	12.68	15.06	24.79	26.61	31.21
	MaPLe	26.73	30.90	34.87	42.00	48.40
	PromptSRC	27.67	31.70	37.47	43.27	50.83
	CasPL **(Ours)**	32.80	35.20	41.03	48.03	55.37
SUN397	Linear probe CLIP	41.58	53.70	63.00	69.08	73.28
	CoOp	66.77	66.53	69.97	71.53	74.67
	CoCoOp	68.33	69.03	70.21	70.84	72.15
	MaPLe	64.77	67.10	70.67	73.23	75.53
	PromptSRC	69.67	71.60	74.00	75.73	77.23
	CasPL **(Ours)**	71.03	72.70	74.53	76.33	77.70
OxfordPets	Linear probe CLIP	44.06	58.37	71.17	78.36	85.34
	CoOp	90.37	89.80	92.57	91.27	91.87
	CoCoOp	91.27	92.64	92.81	93.45	93.34
	MaPLe	89.10	90.87	91.90	92.57	92.83
	PromptSRC	92.00	92.50	93.43	93.50	93.67
	CasPL **(Ours)**	92.97	93.37	93.97	93.93	94.13
UCF101	Linear probe CLIP	53.66	65.78	73.28	79.34	82.11
	CoOp	71.23	73.43	77.10	80.20	82.23
	CoCoOp	70.30	73.51	74.82	77.14	78.14
	MaPLe	71.83	74.60	78.47	81.37	85.03
	PromptSRC	74.80	78.50	81.57	84.30	86.47
	CasPL **(Ours)**	79.53	82.03	84.77	86.70	88.47
Food101	Linear probe CLIP	43.96	61.51	73.19	79.79	82.90
	CoOp	84.33	84.40	84.47	82.67	84.20
	CoCoOp	85.65	86.22	86.88	86.97	87.25
	MaPLe	80.50	81.47	81.77	83.60	85.33
	PromptSRC	84.87	85.70	86.17	86.90	87.5
	CasPL **(Ours)**	86.80	87.20	87.40	87.80	88.40
Average	Linear probe CLIP	45.83	57.98	68.01	74.47	78.79
	CoOp	67.56	70.65	74.02	76.98	79.89
	CoCoOp	66.79	67.65	71.21	72.96	74.90
	MaPLe	69.27	72.58	75.37	78.89	81.79
	PromptSRC	72.32	75.29	78.35	80.69	82.87
	CasPL **(Ours)**	75.91	77.94	80.45	82.22	84.49