PromptKD: Unsupervised Prompt Distillation for Vision-Language Models

Supplementary Material

1. Experimental Settings

Dataset. We evaluate the performance of our method on 15 recognition datasets. For generalization from base-to-novel classes and cross-dataset evaluation, we evaluate the performance of our method on 11 diverse recognition datasets. Specifically, these datasets include ImageNet-1K [4] and Caltech101 [5] for generic object classification; OxfordPets [16], StanfordCars [12], Flowers102 [15], Food101 [1], and FGVCAircraft [14] for fine-grained classification, SUN397 [24] for scene recognition, UCF101 [22] for action recognition, DTD [3] for texture classification, and EuroSAT [6] for satellite imagery recognition. For domain generalization experiments, we use ImageNet-1K as the source dataset and its four variants as target datasets including ImageNet-V2 [17], ImageNet-Sketch [23], ImageNet-A [8], and ImageNet-R [7].

Training Details. For PromptKD, we follow the same settings as PromptSRC, setting the prompt depth to 9 and the vision and language prompt lengths to 4. We use the stochastic gradient descents (SGD) as the optimizer. All student models are trained for 20 epochs with a batch size of 8 and a learning rate of 0.005. We follow the standard data augmentation scheme as in PromptSRC, i.e., random resized cropping and random flipping. The temperature hyperparameter τ in the current distillation method is default set to 1. The text prompts of the first layer are initialized with the word embeddings of "a photo of a {classname}". We conduct all experiments on a single Nvidia A100 GPU.

Training Data Usage. In the initial stage of our method, we employ PromptSRC to pre-train our ViT-L/14 CLIP teacher model. During this stage, we utilize the same training data as PromptSRC for the training process. In the subsequent stage, we adopt the transductive zero-shot learning paradigm and employ the entire training dataset to train our student model. In Table 1, we provide the details of the number of images used for training on the base-to-novel generalization setting.

Dataset	Train	Test Base	Test Novel
ImageNet	1,281,167	25,000	25,000
Caltech101	4,128	1,549	916
OxfordPets	2,944	1,881	1,788
StandfordCars	6,509	4,002	4,039
Flowers102	4,093	1,053	1,410
Food101	50,500	15,300	15,000
FGVCAircraft	3,334	1,666	1,667
SUN397	15,880	9,950	9,900
DTD	2,820	864	828
EuroSAT	13,500	4,200	3,900
UCF101	7,639	1,934	1,849

Table 1. Number of images used for distillation and testing per dataset.

2. Additional Experiments

Domain Generalization. In our PromptKD, the teacher model is first pre-trained using PromptSRC [11] on the source dataset (i.e., ImageNet). Then we train student models using unlabeled target datasets and then evaluate their performance after training.

In Table 2, we present the results of PromptKD and other state-of-the-art methods (i.e., CoOp [26], CoCoOp [25], MaPLe [10], PromptSRC [11], TPT [21], PromptAlign [18]) on four different datasets. On the target dataset, our method shows a clear performance advantage compared to other methods.

		Target Dataset				
ZSL	ViT-B/16	-V2	-S	-A	-R	Avg.
Inductive	CLIP	60.83	46.15	47.77	73.96	57.18
	CoOp	64.20	47.99	49.71	75.21	59.28
	CoCoOp	64.07	48.75	50.63	76.18	59.91
	MaPLe	64.07	49.15	50.90	76.98	60.27
	PromptSRC	64.35	49.55	50.90	77.80	60.65
Transductive	TPT	63.45	47.94	54.77	77.06	60.81
	CoOp+TPT	66.83	49.29	57.95	77.27	62.83
	CoCoOp+TPT	64.85	48.47	58.47	78.65	62.61
	PromptAlign	65.29	50.23	59.37	79.33	63.55
	PromptKD	**69.77**	**58.72**	**70.36**	**87.01**	**71.47**
	Δ	+4.48	+8.49	+10.99	+7.68	+7.92

Table 2. Comparison of PromptKD with existing advanced approaches on domain generalization setting. Based on our pipeline, we perform unsupervised prompt distillation using the unlabeled domain data respectively (i.e., the transductive setting). The source model is training from ImageNet [4]. "ZSL" denotes the setting type for Zero-Shot Learning. PromptKD achieves consistent improvement on all target datasets.

Teacher Accuracy. In Table 3 and Table 4, we present the pre-trained ViT-L/14 based CLIP teacher model accuracy on the base-to-novel and cross dataset experiments.

Layer of Projector. Table 5 presents the distillation performance of different MLP layers used in the projector. The results show that two layers of MLP are effective enough to achieve feature alignment. More or fewer MLP layers will cause over-fitting or under-fitting problems in training.

Distillation with Different Students. To verify the effectiveness of PromtpKD on student models with different capacities, as shown in Table 6, we further conduct experiments on the CLIP models with ViT-B/32 image encoder.

Dataset	Base	Novel	HM
ImageNet	83.24	76.83	79.91
Caltech101	98.71	98.03	98.37
OxfordPets	96.86	98.82	97.83
StandfordCars	84.53	84.25	84.39
Flowers102	99.05	82.60	90.08
Food101	94.56	95.15	94.85
FGVCAircraft	54.44	43.07	48.09
SUN397	84.97	81.09	82.98
DTD	85.76	70.65	77.48
EuroSAT	94.79	83.15	88.59
UCF101	89.50	82.26	85.73

Table 3. Pre-trained ViT-L/14 CLIP teacher accuracy on base-to-novel generalization experiments.

ViT-L/14	Dataset	Accuracy
Source	ImageNet	78.12
Target	Caltech101	95.61
	OxfordPets	94.19
	StandfordCars	84.53
	Flowers102	99.05
	Food101	94.56
	FGVCAircraft	54.44
	SUN397	84.97
	DTD	85.76
	EuroSAT	94.79
	UCF101	89.50

Table 4. Pre-trained ViT-L/14 CLIP teacher accuracy on cross-dataset generalization experiments.

MLP Layer	Base	Novel	HM
1	78.97	72.90	75.81
2	**79.27**	**73.39**	**76.22**
3	79.10	72.72	75.78

Table 5. Number of Projector layers. 2-layer MLP works best.

Role	Img Backbone	Base	Novel	HM
Teacher	ViT-L/14	83.24	76.83	79.91
Baseline	ViT-B/32	67.52	64.04	65.73
Student		74.29	69.29	71.70
Δ		+6.77	+5.25	+5.97
Baseline	ViT-B/16	72.43	68.14	70.22
Student		80.83	74.66	77.62
Δ		+8.40	+6.52	+7.40

Table 6. Prompt distillation with different student CLIP models. Δ denotes the performance improvement compared to the baseline result. Student models of different capacities achieved consistent improvements.

The results show that the student models achieve consistent improvements through the PromptKD method.

Temperature Hyperparameter. The temperature parameter controls the softness of probability distribution [9] and the learning difficulty of the distillation process [13]. In traditional distillation approaches, a common practice is to set the temperature parameter τ to 4 for most teacher-student pairs and datasets. In Fig. 1, we evaluate the impact of different temperature values on our proposed prompt distillation method. The results indicate that the traditional temperature setting of τ=4 is not suitable for our current task. Increasing the temperature value leads to a rapid decrease in model performance. Interestingly, the best performance is achieved when τ=1.

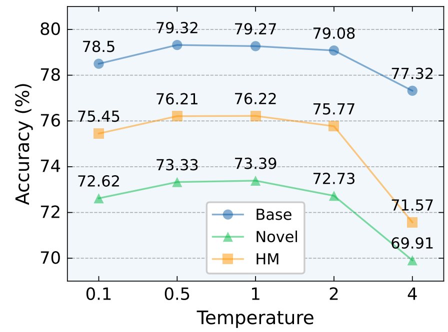

Figure 1. Choice of temperature hyperparameter. The best performance is achieved when τ=1.

Distillation with Longer Schedules. In PromptKD, for fair comparison, we adopt the same training schedule as PromptSRC, which is 20 epochs. In this part, we examine whether the student model can benefit from longer training schedules. As shown in Table 7, we conduct experiments using 20, 40, and 60 training epochs respectively. The results show that the longer the training time, the higher the student performance.

Train Epoch	Base	Novel	HM
20	79.27	73.39	76.22
40	79.75	73.65	76.58
60	**79.89**	**73.68**	**76.66**

Table 7. Distillation with longer schedules. The longer the training time, the higher the student performance.

3. Discussion

Experimental results of full fine-tune. In Table 5 of the main paper, we notice that the results of the full fine-tune method are lower than that of other distillation methods by a large margin (>2%). There are two reasons for this. The first one is due to the limited size of the dataset we used in training. It is much smaller than the CC3M [20], CC12M [2], or LAION-400M [19] datasets commonly used to train CLIP. The second reason is that the training time is short. To align with other experimental settings, we only train the student model for 20 epochs. In total, the full fine-tuning method will improve if larger data sets are used and longer training schedules are adopted.

Distillation with bad teachers. In Figure 5 of the main paper, when a weaker teacher (ViT-B/32) is chosen compared to the student (ViT-B/16), the student trained using PromptKD demonstrates superior performance compared to the baseline method (71.87%>70.22%). This situation differs from traditional distillation methods, where poor teachers often lead to a significant decline in student performance. The distinction arises due to the prompt learning method's focus on training only learnable prompts while keeping the original CLIP model weights frozen. The frozen CLIP model remains influential in the prediction process, where the trained prompts do not substantially bias the model inference.

References

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, pages 446–461. Springer, 2014. 1

[2] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, pages 3558–3568, 2021. 2

[3] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, pages 3606–3613, 2014. 1

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 1

[5] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR workshop*, pages 178–178. IEEE, 2004. 1

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

[7] 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*, pages 8340–8349, 2021. 1

[8] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, pages 15262–15271, 2021. 1

[9] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2

[10] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *CVPR*, pages 19113–19122, 2023. 1

[11] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *ICCV*, pages 15190–15200, 2023. 1

[12] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV workshop*, pages 554–561, 2013. 1

[13] Zheng Li, Xiang Li, Lingfeng Yang, Borui Zhao, Renjie Song, Lei Luo, Jun Li, and Jian Yang. Curriculum temperature for knowledge distillation. In *AAAI*, pages 1504–1512, 2023. 2

[14] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 1

[15] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008. 1

[16] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, pages 3498–3505. IEEE, 2012. 1

[17] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, pages 5389–5400. PMLR, 2019. 1

[18] Jameel Hassan Abdul Samadh, Hanan Gani, Noor Hazim Hussein, Muhammad Uzair Khattak, Muzammal Naseer, Fahad Khan, and Salman Khan. Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. In *NeurIPS*, 2023. 1

[19] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 2

[20] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pages 2556–2565, 2018. 2

[21] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *NeurIPS*, 35:14274–14289, 2022. 1

[22] 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. 1

[23] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *NeurIPS*, 32, 2019. 1

[24] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492. IEEE, 2010. 1

[25] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16816–16825, 2022. 1

[26] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022. 1