Hierarchical Cross-modal Prompt Learning for Vision-Language Models

Supplementary Material

The following sections contain supplemental information and encompass the formulation of the Hierarchical Knowledge Mapper in Sec. A, more implementation details in Sec. B, and a thorough ablative analysis of HiCroPL C.

A. Formal Description of Hierarchical Knowledge Mapper

The hierarchical knowledge mapper projects multi-scale knowledge into a single prompt of another modality, which allows the prompt to adaptively absorb cross-modal information from multiple scales. Taking text-to-image mapping as an example, formally, let $P_v = \{p_v^{l,1}, p_v^{l,2}, ..., p_v^{l,m}\} \in \mathbb{R}^{k \times m \times d_v}$ denote visual prompts and $\tilde{P}_p = \{\tilde{p_p^1}, \tilde{p_p^2}, ..., \tilde{p_p^k}\}$ represent refined textual proxy tokens. The cross-modal mapping is computed as:

$$\begin{aligned} \mathbf{Q} &= P_v \mathbf{W}_q, \quad \mathbf{W}_q \in \mathbb{R}^{d_v \times d_v}, \\ \mathbf{K} &= P_p \mathbf{W}_k, \quad \mathbf{W}_k \in \mathbb{R}^{d_t \times d_v}, \\ \mathbf{V} &= P_p \mathbf{W}_v, \quad \mathbf{W}_v \in \mathbb{R}^{d_t \times d_v}, \end{aligned} \tag{7}$$

where $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v$ are learnable projection matrices. The scaled dot-product attention computes cross-modal interaction:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_v}}\right)\mathbf{V}. \tag{8}$$

Following the standard transformer architecture, we employ layer normalization and residual connections:

$$\begin{aligned} \mathbf{Q}' &= \mathbf{Q} + \text{Attention}(\text{LN}(\mathbf{Q}), \text{LN}(\mathbf{K}), \text{LN}(\mathbf{V})), \\ P_v &= \mathbf{Q}' + \text{FFN}(\text{LN}(\mathbf{Q}')), \end{aligned} \tag{9}$$

where FFN denotes the feed-forward network with GELU activation:

$$\text{FFN}(\mathbf{x}) = \mathbf{W}_2 \cdot \text{GELU}(\mathbf{W}_1 \mathbf{x} + \mathbf{b}_1) + \mathbf{b}_2, \tag{10}$$

where $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{b}_1$, and $\mathbf{b}_2$ are learnable parameters.

B. Additional Implementation Details

Additional Training details. We train HiCroPL for 5 epochs for cross-dataset evaluation and domain generalization settings. The text feature dimension d_t = 512 and the image feature dimension d_v = 768. We fix the learning rate at 0.0025, and optimization is performed using the Adam optimizer with a momentum of 0.9 and weight decay of 0.0005. The corresponding hyperparameters are fixed across all datasets in the same task. All experiments are conducted on a single NVIDIA A100 GPU.

Dataset	Class name	LLM-generated descriptions.
SUN397	airplane cabin	The cabin of an airplane typically has rows of seats on either side of a central aisle.
	bookstore	A bookstore has shelves full of books and usually has a desk where you can pay for your books.
	campus	A campus looks like a collection of buildings that are close together.

Table 9. Example of descriptive text generated by LLM.

Datasets	Classes	Training Size	Validation Size	Testing Size
ImageNet [7]	1,000	1,281,167	N/A	50,000
Caltech101 [8]	100	4,128	1,649	2,465
EuroSAT [12]	10	13,500	5,400	8,100
SUN397 [48]	397	15,880	3,970	19,850
DTD [6]	47	2,820	1,128	1,692
UCF101 [41]	101	7,639	1,808	3,783
FGVCAircraft [31]	100	3,334	3,333	3,333
OxfordPets [35]	37	2,944	736	3,669
StanfordCars [21]	196	6,509	1,635	8,041
Flowers102 [33]	102	4,093	1,633	2,463
Food101 [3]	101	50,500	20,200	30,300
ImageNet-V2 [37]	1000	N/A	N/A	10000
ImageNet-Sketch [44]	1000	N/A	N/A	50889
ImageNet-A [14]	200	N/A	N/A	7500
ImageNet-R [13]	200	N/A	N/A	30000

Table 10. Detailed statistics of the datasets.

LLM-generated category descriptions. We employ large language model (LLM) to generate detailed descriptions for each category, providing diverse frozen text features. For each category, we utilize GPT-3 [4] to generate descriptive sentences. For simplicity, we adopt the publicly available CoPrompt [39] data. However, unlike CoPrompt, we average the embeddings of all descriptions for each category to obtain the final category embedding, rather than dynamically selecting a single sentence as the category representation. Table 9 presents a sample of the LLM-generated category descriptions.

Datasets. We evaluate the performance of our method on 15 recognition datasets. For base-to-novel generalization and cross-dataset evaluation tasks, we evaluate our method on 11 image datasets covering various recognition tasks. These include ImageNet [7] and Caltech101 [8] for general object recognition. Five fine-grained classification datasets, OxfordPets [35], StanfordCars [21], Flowers102 [33], Food101 [3], and FGVCAircraft [31]. SUN397 [48] is used for scene recognition, UCF101 [41] for action recognition, DTD [6] for texture classification,

BKF	$\mathcal{L}_{cons}$	Base	Novel	HM
		82.15	74.07	77.90
	√	82.09	76.02	78.94
√		**85.96**	74.65	79.91
√	√	85.89	**77.99**	**81.75**

Table 11. Ablation experiments on the components of HiCroPL. BKF refers to the Bidirectional Knowledge Flow mechanism.

Frozen prompts choice	Base	Novel	HM
a photo of a {}	84.92	75.99	80.21
frozen diverse prompts	85.14	75.23	79.88
LLM (a sentence)	85.33	76.41	80.63
LLM(ensemble)	**85.89**	**77.99**	**81.75**

Table 12. Ablation on frozen prompt choices.

and EuroSat [12] for satellite image classification. For the domain generalization task, ImageNet [7] is used as the source domain dataset for training the model, and its variants ImageNet-A [14], ImageNet-R [13], ImageNet-Sketch [44] and ImageNet-V2 [37] are used for out-of-distribution dataset evaluation. The detailed statistics of the 11 datasets, as well as the four variants of ImageNet [7], are shown in Table 10.

C. Additional Experiments

Effect of consistency regularization. Table 11 provides ablation experiments on the components of HiCroPL. The bidirectional knowledge flow mechanism significantly boosts base class performance and achieves the best overall results. Additionally, by leveraging intermediate-layer features, it also improves performance on novel classes. While using the regularization term alone enhances generalization to novel classes, it does not provide gains on base classes. Ultimately, the combination of both components in HiCroPL achieves the best performance.

Effect of frozen prompts. Since different frozen prompts provide distinct knowledge to constrain prompt learning, we evaluate the effectiveness of various hand-crafted prompts. Specifically, we compare the fixed prompt "a photo of a {}" used in KgCoOp [51], the diverse textual descriptions in PromptSRC [20], the randomly sampled LLM prompts in CoPrompt [39], and the averaged LLM prompts in our HiCroPL. The results are shown in Table 12. Compared to the dynamically generated individual sentences in CoPrompt, ensemble LLM-generated prompts provide richer textual features, thereby improving performance. However, the diverse textual descriptions used in PromptSRC are based on the text templates provided by CLIP for ImageNet, which may lead to inaccurate descriptions when applied to other datasets, resulting in performance degradation.

Criterion	Base	Novel	HM
MSE	85.11	74.39	79.39
L1	85.79	77.2	81.27
Cosine	**85.89**	**77.99**	**81.75**

Table 13. Comparison of different distillation consistency criteria. Cosine similarity works best.

Influence of different consistency criteria. We evaluate the impact of different consistency criteria on constraints in Table 13. The results show that using cosine similarity as the consistency criterion provides the best performance, followed by L1, while using MSE severely degrades the performance.

Few-shot experiments. We evaluate the adaptability of HiCroPL through few-shot experiments. Table 14 provides detailed per-dataset results for various methods under the few-shot setting. Compared to previous methods, HiCroPL achieves consistent improvements.

Dataset	Method	1 shot	2 shots	4 shots	8 shots	16 shots
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
	HiCroPL	**74.67**	**76.67**	**79.01**	**80.96**	**83.30**
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
	HiCroPL	**70.54**	**70.92**	**71.99**	**72.91**	**73.87**
Caltech101	Linear probe CLIP	79.88	89.01	92.05	93.41	95.43
	CoOp	92.60	93.07	94.4	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	93.83	94.53	95.27	95.67	96.07
	HiCroPL	**94.44**	**95.33**	**95.66**	**96.23**	**96.23**
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
	HiCroPL	**92.29**	**92.50**	93.24	**93.70**	**93.81**
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
	HiCroPL	**70.64**	**74.98**	76.84	**81.03**	**84.28**
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**
	HiCroPL	**86.32**	90.78	**94.15**	95.94	97.32
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	87.50
	HiCroPL	**86.37**	86.21	**86.98**	**87.33**	**87.6**
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
	HiCroPL	**31.89**	**33.90**	**38.37**	42.72	**51.13**
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
	HiCroPL	**70.27**	**72.48**	**74.62**	**76.24**	**77.66**
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	55.5	61.00	66.50	71.33
	PromptSRC	56.23	59.97	65.53	69.87	72.73
	HiCroPL	**59.52**	**62.00**	**67.14**	**70.04**	**75.65**
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	86.90	88.80	**92.43**
	HiCroPL	**82.2**	**85.53**	**87.47**	**89.17**	92.05
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
	HiCroPL	**76.92**	**78.69**	**82.71**	**85.22**	**86.70**

Table 14. Comparison of HiCroPL performance with various methods for each dataset in few-shot setting.