Pre-trained Vision and Language Transformers Are Few-Shot Incremental Learners

Supplementary Material

The supplementary materials cover baseline information (§A), additional details on main experiments (§B), and ablation studies (§C) including pre-trained knowledge tuning (§C.1), modulation prompts (§C.2), semantic knowledge distillation loss (§C.3), and further analysis for the performance (§C.4). Also included are PKT for domain-specific knowledge (§D), analysis of transferability (§E), limitation of prefix tuning (§F), and visualization of entropy-based divergence loss effectiveness (§G).

A. Details of Baseline

CEC. In Zhang et al. [40], a graph attention network (GAN) acts as a classifier. The feature extractor is trained for strong base knowledge, and the GAN adapts to novel class knowledge in incremental sessions. GAN parameters increase with each new session. In our experiments, we partially trained the first two feature extractor layers for stable GAN training.

NC-FSCIL. Yang et al. [39] proposed the framework inspired by neural collapse (Papyan et al. [22]) which aims to align between the feature and corresponding weight of the classifier. NC-FSCIL pre-assigned the set of classifier prototypes which is formed as a simplex equiangular tight frame (ETF). NC-FSCIL proposed aligning a classifier with prototypes to enhance the performance of the classifier.

WaRP. Kim et al. [12] introduced the weight space rotation process which is called WaRP. They change the trained weight space into a new space where most of the important previous knowledge is condensed into a few parameters. It means WaRP can train the network to capture the knowledge in incremental sessions without suffering catastrophic forgetting.

L2P. In Wang et al. [37], L2P is a prompt-based framework for class incremental learning, leveraging a pre-trained vision transformer. Using the prompt pool, L2P selects a prompt based on input samples and fine-tunes it for training. We adopted a single prompt during training to avoid performance deterioration from expanding prompts and omitted the prompt selection process.

DualPrompt. In Wang et al. [36], DualPrompt excels in class incremental learning by training G-Prompt and E-Prompt separately. It dynamically expands E-Prompt to retain task-specific knowledge. Similar to L2P, in our experiment, we used a single G-Prompt and E-Prompt, respectively and omitted prompt selection during evaluation.

B. More Details for Main Experiments

We presented the average accuracy across five simulations on CUB200, CIFAR-100, and miniImageNet. The highest and second-highest performances were indicated by bold and underlined text, respectively.

As shown in Table S1. Our method outperformed others in all sessions, except the base session with ViT-B. In contrast to prior approaches suffering significant performance drops with new class arrivals, our method demonstrated a minor performance decline. Notably, prompt-based methods like L2P and DualPrompt performed less effectively than baselines. While prompt-based methods show promising performance in class incremental learning, their effectiveness diminished in FSCIL where transferability is crucial. Due to the limited trainable parameters of prompts, they struggle to capture sufficient domain-specific knowledge in the base session, impeding effective transfer to incremental sessions. Our method, PriViLege, designed to transfer diverse and domain-specific knowledge leveraging prompts, successfully mitigated catastrophic forgetting in FSCIL, aiding newly introduced classes.

In Table S2, we reported the performance on CIFAR-100. Our method, PriViLege, consistently outperformed other baselines in every incremental session. It is noteworthy that in training session 4, our method exhibited an enhanced performance of approximately $+0.28\%$ compared to the previous session. This improvement is particularly significant considering that WaRP, the second-highest performer, experienced a substantial performance decline of about -2.11% in the same session. The notable performance gain of our proposed method emphasizes its robust transferability, which not only contributes to forward transfer but also marginally contributes to positive backward transfer.

In Table S3, utilizing a network pre-trained on ImageNet-21K (Russakovsky et al. [28]), our method demonstrated the highest performance among all datasets. CEC exhibited a secondary performance, attributed to its partial network training. Surprisingly, our method reported minimal knowledge forgetting even after training all sessions. It is noteworthy that methods leveraging pre-trained knowledge, such as L2P and DualPrompt, showed competitive performance with existing FSCIL methods such as WaRP and NC-FSCIL. This observation underscores the significance of considering how to effectively leverage pre-trained knowledge when employing a ViT in FSCIL.

Method	Sessions											A_{Avg}
	A_{Base}	1	2	3	4	5	6	7	8	9	A_{Last}	
Fine-Tuning + Proto ψ	**84.21±0.13**	66.43±3.40	25.00±14.47	25.44±6.70	16.19±12.58	4.58±3.34	1.42±1.03	1.49±0.80	3.62±3.94	5.50±5.67	3.79±1.47	21.60±1.32
CEC[CVPR'21]	75.40±8.01	73.23±8.32	72.00±8.25	68.70±8.43	69.35±8.68	67.78±7.88	67.01±7.79	66.40±8.04	65.78±8.10	65.57±7.95	65.70±8.03	72.41±1.18
L2P[CVPR'22]	44.97±2.32	30.28±6.67	27.21±6.04	24.44±5.44	22.41±4.87	20.81±4.49	19.47±4.24	18.19±4.09	17.16±3.87	16.26±3.65	15.41±3.45	24.99±4.30
DualPrompt[ECCV'22]	53.37±1.83	45.99±2.58	41.15±2.85	37.33±2.86	34.32±2.72	31.57±2.45	29.44±2.34	27.58±2.20	25.92±2.24	24.55±2.12	23.25±2.02	36.30±2.39
NC-FSCIL[ICLR'23]	78.49±2.32	71.52±2.11	65.54±1.93	60.30±1.78	55.81±1.65	51.96±1.53	48.72±1.44	45.78±1.35	43.18±1.27	40.92±1.21	38.80±1.14	57.92±1.71
WaRP[ICLR'23]	67.74±5.57	64.21±5.54	61.06±5.90	57.80±5.93	55.78±5.96	53.81±6.08	52.82±6.25	51.61±6.47	50.13±6.27	50.02±6.23	49.36±6.56	55.85±6.06
PriViLege (Ours)	82.21±0.20	**81.25±0.20**	**80.45±0.20**	**77.76±0.41**	**77.78±0.47**	**75.95±0.40**	**75.69±0.41**	**76.00±0.33**	**75.19±0.45**	**75.19±0.47**	**75.08±0.52**	**77.50±0.33**

Table S1. The performance of every session on CUB200.

Method	Sessions									A_{Avg}
	A_{Base}	1	2	3	4	5	6	7	A_{Last}	
Fine-Tuning + Proto ψ	**91.36±0.15**	73.95±1.38	41.61±12.23	40.46±10.96	41.69±9.77	13.96±8.53	16.45±10.66	8.71±5.90	5.19±0.13	37.04±1.06
CEC[CVPR'21]	74.20±2.03	71.49±2.13	70.11±2.54	67.34±2.88	65.96±2.64	65.14±3.36	64.74±3.96	63.48±4.09	61.48±3.33	67.10±2.92
L2P[CVPR'22]	83.29±0.50	76.81±0.43	71.29±0.43	66.53±0.39	62.38±0.36	58.68±0.38	55.42±0.36	52.49±0.33	49.87±0.31	64.08±0.39
DualPrompt[ECCV'22]	85.11±0.29	78.42±0.29	72.81±0.35	67.92±0.35	63.69±0.29	59.92±0.26	56.60±0.23	53.62±0.21	50.93±0.21	65.45±0.27
NC-FSCIL[ICLR'23]	89.51±0.23	82.62±0.21	76.72±0.19	71.61±0.18	67.13±0.17	63.18±0.16	59.67±0.15	56.53±0.14	53.70±0.14	68.96±0.17
WaRP[ICLR'23]	86.20±1.46	82.58±1.53	79.30±1.77	75.57±1.66	73.46±1.61	71.07±1.69	69.58±1.80	67.70±1.85	65.48±1.87	74.55±1.67
PriViLege (Ours)	90.88±0.20	**89.39±0.23**	**88.97±0.15**	**87.55±0.24**	**87.83±0.24**	**87.35±0.24**	**87.53±0.25**	**87.15±0.21**	**86.06±0.32**	**88.08±0.20**

Table S2. The performance of every session on CIFAR-100.

Method	Sessions									A_{Avg}
	A_{Base}	1	2	3	4	5	6	7	A_{Last}	
Fine-Tuning + Proto ψ	93.67±0.02	87.12±5.61	73.54±15.17	50.29±16.74	26.39±17.13	7.29±0.02	23.52±18.90	29.74±4.40	9.87±5.42	44.60±0.92
CEC[CVPR'21]	87.43±5.90	85.99±6.70	84.03±7.03	83.21±7.28	83.11±7.16	81.64±7.66	80.66±7.56	80.72±7.56	80.74±7.51	83.06±7.14
L2P[CVPR'22]	94.59±0.21	87.49±0.45	81.18±0.49	75.76±0.45	71.05±0.39	66.86±.0.36	63.15±0.34	59.82±0.32	56.84±0.32	72.97±0.36
DualPrompt[ECCV'22]	95.05±0.20	87.81±0.19	81.51±0.21	76.07±0.21	71.38±0.12	67.19±0.15	63.45±0.12	60.15±0.10	57.14±0.11	73.31±0.15
NC-FSCIL[ICLR'23]	77.25±0.42	71.30±0.39	66.21±0.36	61.80±0.34	57.94±0.32	54.53±0.30	51.50±0.28	48.79±0.27	46.35±0.25	59.52±0.33
WaRP[ICLR'23]	83.30±1.06	80.53±1.48	77.22±1.01	74.99±1.50	73.64±0.97	71.52±1.07	69.16±0.84	68.79±0.79	67.97±1.28	74.13±1.08
PriViLege (Ours)	**96.68±0.06**	**96.49±0.05**	**95.65±0.15**	**95.54±0.13**	**95.54±0.13**	**94.91±0.16**	**94.33±0.15**	**94.19±0.12**	**94.10±0.13**	**95.27±0.11**

Table S3. The performance of every session on miniImageNet.

PKT Components			CUB200		
LT	Modulation	B+VL	A_{Base}	A_{Last}	A_{Avg}
			84.21±0.13	3.79±1.47	21.60±1.32
	✓		65.31±1.81	51.04±1.36	57.47±1.51
		✓	76.43±0.35	60.32±0.73	67.38±0.41
	✓	✓	76.20±0.41	61.47±0.83	67.86±0.52
✓			74.48±0.14	64.75±0.99	68.66±0.52
✓	✓		77.38±0.82	68.09±1.02	71.42±0.80
✓		✓	78.30±1.55	68.58±2.68	72.07±1.90
✓	✓	✓	79.06±0.77	**70.81±0.76**	**73.36±0.77**

Table S4. Further ablation experiment for PKT on CUB200. Modulation denotes leveraging modulation prompts and B+VL denotes prefix tuning the B-Prompt and prompt tuning the VL-Prompt.

C. Additional Ablation Studies

We conducted additional ablation studies to confirm the effectiveness of each proposed component, focusing on pre-trained knowledge tuning, modulation prompts, and semantic knowledge distillation loss.

C.1. Ablation Study for PKT

In Table S4, we conducted an ablation study on pre-trained knowledge tuning. Our baseline (row 3) utilized fine-tuning with a prototype classifier ψ. Rows 4 to 6 did not incorporate layer tuning. In Table S4, we observed a gradual performance enhancement with the proposed pre-trained knowledge tuning. Notably, employing all proposed components showed the highest performance in both A_{Last} and A_{Avg}.

We observed that the absence of layer tuning led to lower performance in the base session compared to its presence. This observation highlights the importance of layer tuning for acquiring sufficient domain-specific information, especially given the limited capacity of a fixed model. Additionally, relying solely on modulation prompts recorded lower performance than leveraging only learnable prompts like B-Prompt and VL-Prompt. Modulation prompts, designed to facilitate learnable prompt updates, struggled to provide useful knowledge for prefix tuning. Consequently, adopting only B-Prompt and VL-Prompt yielded better performance than solely relying on modulation prompts, emphasizing their additional capacity, irrespective of layer tuning.

Additionally, we observed that adopting modulation prompts with additional learnable prompts showed performance enhancement in A_{Last} and A_{Avg} regardless of layer tuning. Since the modulation prompts can contribute to the update of the learnable prompts, additional prompts, especially B-Prompt, can capture more effective domain-specific knowledge via the modulation prompts. Lastly, we observed the most promising performance when we adopted all the proposed components. The proposed pre-trained knowledge tuning can contribute to capturing effective domain-specific knowledge at the base session due to the additional B-Prompt and VL-Prompt assisted by the modulation prompts while preserving pre-trained knowledge through partial layer tuning.

Prefix Tuning		CUB200		
Key	Value	A_{Base}	A_{Last}	A_{Avg}
P_M^G		78.29±0.27	69.14±0.63	72.34±0.30
P_M^S		78.68±0.17	69.94±1.14	72.91±0.25
P_M^S	P_M^G	**79.06±0.77**	**70.81±0.76**	**73.36±0.77**

Table S5. Further ablation experiment for the modulation prompts.

$\mathcal{L}_{SKD}$		CUB200		
$\mathcal{L}_{KD}$	$\mathcal{L}_{CE}$	A_{Base}	A_{Last}	A_{Avg}
		79.06±0.77	70.81±0.76	73.36±0.77
✓		80.24±0.59	71.59±0.58	74.51±0.13
✓	✓	**82.10±0.57**	**73.44±0.40**	**76.27±0.30**

Table S6. Further ablation experiment for semantic knowledge distillation on CUB200.

C.2. Ablation Study for Modulation

For an analysis of the modulation prompt, we conducted an additional ablation study on CUB200. We assessed the effectiveness of the head-specific prompt P_M^S and the generic prompt P_M^G by separately incorporating each prompt in prefix tuning. As shown in Table S5, relying solely on each head-specific prompt or generic prompt resulted in lower performance compared to leveraging both prompts simultaneously. Since the modulation prompts are constructed in different layers, with the head-specific prompt originating from the MSA layer and the generic prompt from the MLP layer, the head-specific prompt can contribute to scaling the attention score of the B-Prompt, capturing additional relationships between key vectors. Meanwhile, the generic prompt affords the incorporated knowledge with the B-Prompt through the scaling value vectors. Thus, we demonstrated that utilizing modulation prompts is highly beneficial to assist B-Prompt in prefix tuning.

C.3. Ablation Study for $\mathcal{L}_{SKD}$

The proposed semantic knowledge distillation loss comprises knowledge distillation loss and cross-entropy loss. To thoroughly assess its effectiveness, we conducted an additional ablation study using only PKT as the baseline. As indicated in Table S6, our proposed semantic knowledge distillation loss exhibited a gradual improvement in performance. Especially, it is noteworthy that the cross-entropy loss significantly contributes to performance enhancement across all metrics, due to the reduction of heterogeneity be-

CIFAR-100	A_{Base}	A_{Last}	A_{Avg}	Fgt
CLIP-FT	79.43	33.64	50.65	45.79
CLIP-LP	82.20	48.51	58.95	33.69
LP-DiF*-CLIP	80.23	72.02	75.12	8.21
PriViLege-CLIP	**84.25**	**78.35**	**77.16**	**5.90**

Table S7. Experiments of fine-tuning (FT) and linear proving (LP). The performance of LP-DiF* comes from the original paper.

CUB200	A_{Base}	A_{Last}	A_{Avg}
CEC (ViT-S)	78.51	70.10	72.93
WaRP (ViT-S)	72.56	56.96	62.54
PriViLege (ViT-S)	**80.25**	**72.24**	**74.89**
CEC (ViT-L)	76.78	69.45	71.82
WaRP (ViT-L)	78.83	62.64	68.38
PriViLege (ViT-L)	**83.79**	**76.43**	**79.20**

Table S8. Experiments of adopting ViT-S and ViT-L on CUB200.

tween two different spaces.

C.4. Further Analysis for the Performance

We conducted further experiment to compare our method with the LP-DiF [8] which is based on CLIP. Table S7 shows that PriViLege recorded better performance and lower forgetting than LP-DiF. We also analyzed the scalability of our proposed method. In Table S8, our method showed improved performance when the capacity of the base model is increased. Experimental results shows that our method can expect performance enhancement in a more strong base model.

C.5. Considering Pre-trained Base Model

The pre-trained dataset, ImageNet-21K, includes almost all classes in the dataset used in experiments. For example, CIFAR-100 and CUB200 include 12 and 150 exclusive classes, respectively. We conducted further experiments to prove that the performance enhancement stems from the proposed method, PriViLege. Table S9 shows our superior performance on FGVC-aircraft, a dataset non-overlapped with ImageNet-21K. Moreover, Table S10 presents our remarkable performance even when training from scratch. These results support the superiority of our method regardless of the base model.

D. PKT for Domain-Specific Knowledge

To confirm the effectiveness of pre-trained knowledge tuning in capturing domain-specific knowledge, we conducted additional analysis on CUB200. This aimed to clarify the reasons for performance enhancement through the proposed PKT. Figure S1 displays the attention map of FR-B-Prompt, representing the use of a learnable prompt via prefix tuning on the frozen ViT, alongside the map of our proposed

FGVC-aircraft	A_{Base}	A_{Last}	A_{Avg}
CEC	23.05	16.85	19.46
WaRP	24.85	15.69	19.67
PriViLege (Ours)	**58.30**	**45.55**	**50.87**

Table S9. Experiments on FGVC-aircraft.

CIFAR-100	A_{Base}	A_{Last}	A_{Avg}
CEC	8.30	4.76	6.09
WaRP	35.82	23.30	28.85
PriViLege	**50.37**	**30.83**	**39.05**

Table S10. Experiments of ViT-B scratch.

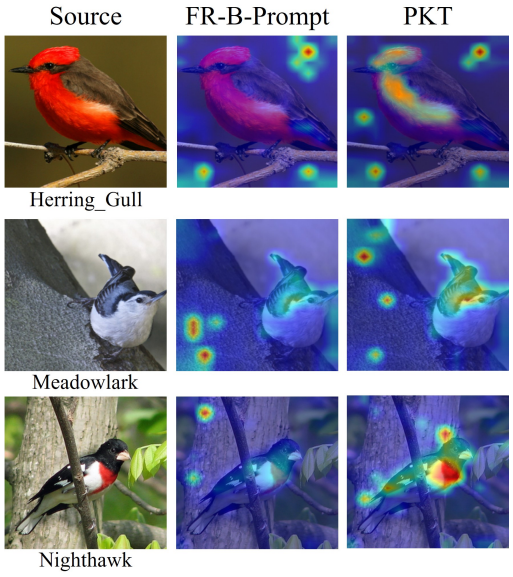
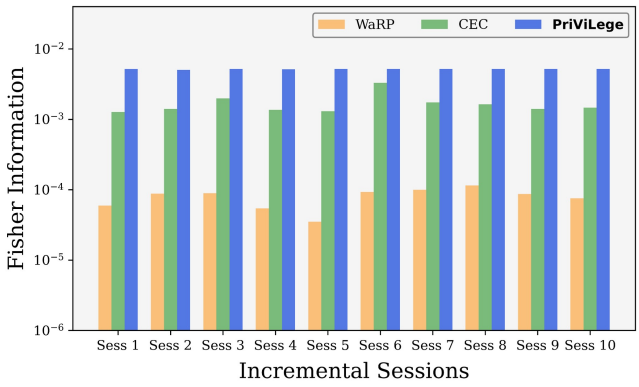
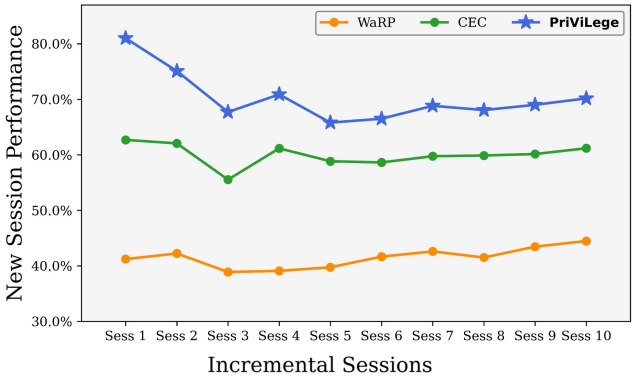


Figure S1. Attention map to assess the effectiveness of the proposed PKT on CUB200. FR-B-Prompt denotes prefix tuning learnable prompt with frozen ViT.

PKT. While FR-B-Prompt trained learnable parameters like B-Prompt using prefix tuning through the frozen ViT, the proposed PKT trained B-Prompt and modulation prompt with prefix tuning and also fine-tuned partial layers. For visualization, we trained a learnable B-Prompt but constructed the attention mask using only image tokens. As illustrated in Figure S1, we observed that the attention map of our proposed PKT exhibited greater activation towards the object compared to FR-B-Prompt. Unlike the attention map of FR-B-Prompt, the attention map of our proposed PKT was more focused on the object rather than the background. Through this observation, we demonstrate that our proposed PKT primarily aims to extract knowledge from the object. Since our PKT fine-tuned some pre-trained layers and trained B-prompt using the modulation prompt, which facilitated prefix tuning, our PKT can effectively capture domain-specific knowledge through the fine-tuned layers and more efficiently through the B-Prompt facilitated by the modulation prompts. Thus, our PKT can capture more



(a) Comparison of the fisher information to assess the transferability.



(b) Comparison of the new task performance on CUB200.

Figure S2. Comparison of the fisher information and new task performance on CUB200.

domain-specific knowledge focused on the class object.

E. Further Analysis for Transferability

To compare the transferability of our method with state-of-the-art FSCIL methods, we conducted additional analyses, considering Fisher information (Figure S2a) and new task performance (Figure S2b). Fisher information is widely used in continual learning as a metric to estimate how important the trained parameters are for the training of a given task. We assessed transferability through the Fisher information of the parameters trained at the base session. If the parameters trained at the base session have high value of the Fisher information in the incremental session, it indicates their importance for the incremental session. Through this, we evaluated the transferability of our method compared to other baselines. Additionally, by analyzing new task performance, we demonstrate effective incremental session learning through the transferred knowledge.

As illustrated in Figure S2a, our method, PriViLege, achieved the highest value of Fisher information compared to other baselines. This observation indicates that our method can effectively transfer useful domain-specific knowledge to incremental sessions. We demonstrated that our method captures transferable knowledge at the base session, consistently utilized as valuable knowledge for the incremental sessions. Furthermore, as shown in Figure S2b, our method also reported the most promising new task performance. Given that new task performance measures the accuracy of each session under all seen classes, PriViLege demonstrated remarkable performance despite the few-shot data given at the incremental sessions. Through this observation, we also validated that our proposed method, which captures transferable and useful domain-specific knowledge, exhibits outstanding transferability to facilitate the learning of incremental sessions.

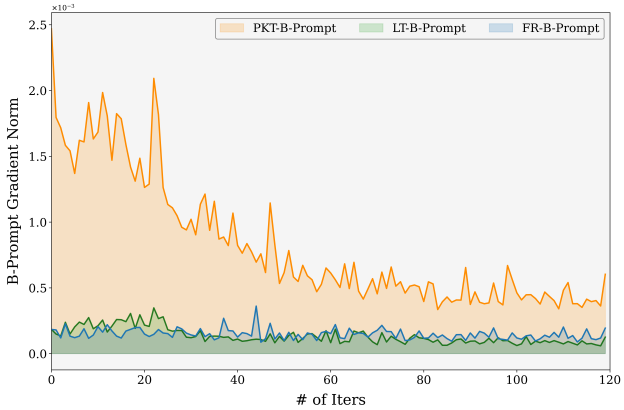
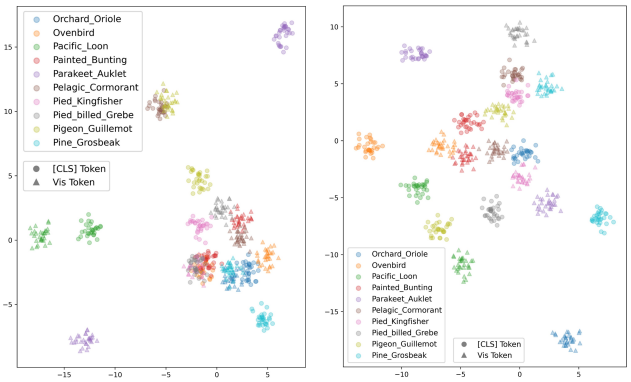


Figure S3. B-Prompt gradient magnitudes on CUB200. PKT-B-Prompt, LT-B-Prompt, and FR-B-Prompt denote adopting PKT, only layer tuning, and fixed ViT for the B-Prompt.

F. Analysis for the Limitation of Prefix Tuning

As mentioned in Section 3.1, prefix tuning has a limitation in updating B-Prompt due to its slow adaptation speed. To overcome this limitation, we proposed modulation prompts. We conducted further analysis to validate that our proposed modulation prompts can effectively enhance the update of the B-Prompt. We calculated the norm of gradient vectors of B-Prompt at every iteration.

As illustrated in Figure S3, relying solely on layer tuning with B-Prompt or leveraging the frozen ViT showed a small norm of gradient vectors due to the slow adaptation speed of prefix tuning. This is because the feature vectors from the B-Prompt tokens are overwhelmed by feature vectors from the input tokens, causing the B-Prompt to struggle to contribute to capturing knowledge and suffer slow adaptation via prefix tuning. However, utilizing PKT, including the modulation prompts, demonstrated a promising increase in the gradient norm of B-Prompt. Since the modulation prompts can scale the key and value of the B-Prompt, it promotes the update of the B-Prompt effectively.



(a) Without $\mathcal{L}_{ED}$. (b) With $\mathcal{L}_{ED}$.

Figure S4. Feature space visualization of [CLS] and vision token on CUB200. The circle and triangle denote the [CLS] token and the vision token, respectively. Each color represents the classes.

G. Further Analysis the Effectiveness of $\mathcal{L}_{ED}$

As mentioned in Section 3.2, the average pooling of [CLS] and vision token results in sharing similar feature knowledge between [CLS] and vision token, hindering effective learning of the vision token. We further analyzed the problem of average pooling and the effectiveness of entropy-based divergence loss ($\mathcal{L}_{ED}$) in the perspective of feature vectors from [CLS] and vision token, respectively. As shown in Figure S4, we visualized the feature space that includes [CLS] and vision token. Figure S4a showed the feature space without applying entropy-based divergence loss, and Figure S4b illustrated the feature space applying entropy-based divergence loss.

As shown in Figure S4a, [CLS] and vision token are located closely in the feature space or even overlap with other classes. Since [CLS] and vision token share the same objective for the classification task due to average pooling, they struggle to capture discriminative knowledge to distinguish each other. This problem hinders the vision token from capturing effective knowledge and learning discriminative features for the classification task.

However, as illustrated in Figure S4b, applying entropy-based divergence loss can effectively mitigate the problem of proximity or overlap. It is noteworthy that entropy-based divergence loss can also help [CLS] feature vector and vision feature vector become discriminative not only when they belong to different classes but also when they belong to the same class. Through this observation, we demonstrated the problem of average pooling and validated the effectiveness of the proposed entropy-based divergence loss. Our entropy-based divergence loss helps mitigate the sharing of knowledge between [CLS] and vision token and enhances discriminative ability, even when classifying [CLS] and vision token that belong to the same class.