Supplementary Material of

CAPT: Confusion-Aware Prompt Tuning for Reducing Vision-Language Misalignment

A. Discussion

CAPT explicitly models systematic confusion at both the semantic and sample levels, enabling the model to extract stable confusion structures from its own misclassifications. However, on relatively easier datasets such as Flowers [38] and Food101 [3], where overall accuracy is **already high** and genuinely confusable samples are **scarce**, the room for improvement is limited, resulting in modest performance gains. In contrast, CAPT yields more pronounced gains on more complex datasets such as UCF101 [50], DTD [6], with about **2.35%** improvement. Nevertheless, CAPT achieves stable superiority across all 11 benchmarks, with a 50.72% correction rate on confusing samples, demonstrating its effectiveness on real challenging classes.

B. Impact of Model Choice on Confusion Bank Construction

In the main text, we use PromptKD to construct our confusion bank. In practice, different models often exhibit similar and stable confusion patterns across many categories. However, due to differences in their feature spaces, alignment behaviors, and error distributions, confusion banks built from different models vary significantly in the coverage and representativeness of confusing instances, which in turn leads to different downstream effects. delivers the highest correction rate. Shown in Table 9, PromptKD achieves the highest overall accuracy, whereas the Confusion Bank built from CLIP delivers the highest correction rate. Thus, stronger models help construct more accurate recognition benchmarks, while simpler models tend to capture more transferable and consistent confusion patterns that are more effective for error correction.

Table 9. Impact of Model Choice on Confusion Bank Construction.

Method Choice	Base	Novel	HM	Correction Rate
CLIP	73.56	77.23	77.35	64.28
CoOp	84.19	71.88	77.55	42.71
MaPLe	85.64	79.14	82.26	39.39
TAC	84.41	78.67	81.44	49.20
PromptKD	87.41	80.90	83.90	50.72

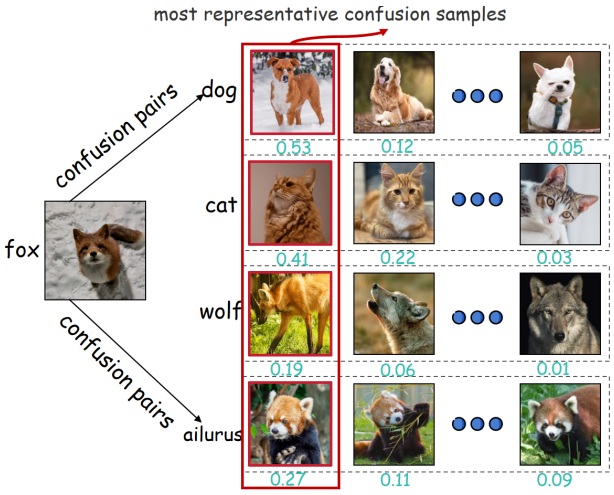

Figure 6. Instance of selecting representative confusion samples.

C. More Details about the Most Representative Confusion Samples

In the Sample Confusion Miner (SAM) module, we leverage the previously identified semantic confusion pairs and apply Eq 8 to retrieve, for each confusable class, the most representative confusing sample. Figure 6 illustrates examples of selecting representative confusion samples for specific instances. For each confusable class, only the most similar sample is retained as the representative confusing example, which helps filter out noisy candidates, sharpen the model's focus on critical confusion patterns, and ultimately enhance its ability to discriminate fine-grained differences and make more stable predictions.

D. Datasets

To obtain a comprehensive understanding of our method's capability, we evaluate it across an extensive suite of benchmarks covering standard recognition, transfer learning, and distribution shift scenarios, as summarized in Table 12. Our core evaluation is conducted on 11 widely adopted datasets that span a broad range of visual domains—including general object recognition (ImageNet [7], Caltech [11]), fine-grained categorization (OxfordPets [40], StanfordCars [29], Flowers [38], Food101 [3], FGVCAircraft [36]), satellite-based scene understanding (EuroSAT [18]), large-scale scene recognition (SUN397 [57]), human action recognition (UCF101 [50]), and texture classification (DTD [6]).

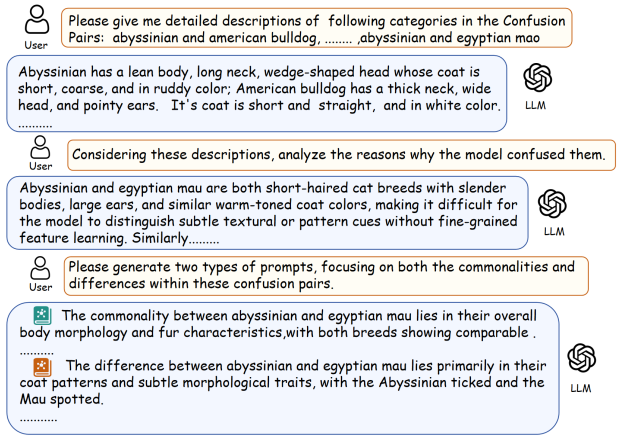

Figure 7. Procedure of generating semantic commonality and difference prompts.

These datasets collectively cover a wide spectrum of visually confusable scenarios, ranging from highly similar fine-grained categories to semantically overlapping scene types, making them well-suited for evaluating confusion-aware modeling. To further assess the adaptability of our approach, we also incorporate three widely used transfer-learning datasets, which provide additional settings where category-level similarity and domain-specific ambiguity introduce new confusion patterns that challenge representation robustness. Beyond standard benchmarks, we additionally examine robustness under real-world distribution shifts using four challenging ImageNet variants: ImageNet-V2 [45], ImageNet-Sketch [53], ImageNet-A [20], and ImageNet-R [19]. These variants introduce complementary perturbations—such as re-sampled images, sketch abstractions, natural adversarial examples, and artistic renditions—that often amplify inherent confusability among visually similar categories. Collectively, this diverse suite of benchmarks offers a rigorous and multi-dimensional testbed for evaluating not only the overall effectiveness and generalization of our method but also its ability to remain robust under semantically and visually confusing conditions.

E. Details of Semantic Prompt Generation in Semantic Confusion Miner.

We leverage confusion pairs generated by SEM to construct semantic level commonality and difference prompts, which are then used to initialize the expert layers in MGDE, effectively integrating multi-class confusion information. Specifically, as illustrated in the Figure 7, we employ a CoT-based [34, 55] procedure to guide prompt generation: the model first produces detailed feature descriptions for the categories involved in the confusion pairs, then uses these descriptions to reason about the underlying causes of misalignment, and finally generates the corresponding difference and commonality prompts. By progressively guiding the model through this process, we can more accurately capture the confusion characteristics within the pairs, thereby enhancing the model's understanding and discrimination of inter-class confusion relationships.

F. Details of Dynamic Weight in Diff-Manner Adapter

In Equal 10, the hyper-parameter α plays a critical role in learning sample confusion features. On the one hand, when α is too small, the model focuses only on coarse, global confusion structures and fails to align finer-grained confusion cues. On the other hand, an excessively large α leads the model to overemphasize local details, reducing generalization ability and causing severe overfitting. To balance global confusion patterns and instance-level cues, we introduce an adaptive α scheduling mechanism that enables the model to automatically determine its level of attention based on the confusion intensity of each sample. We first get the confusion intensity c_i of each representative confusion sample based on Equal 8. Then, we compute a dynamic α:

$$\alpha_i = s \cdot c_i^{\gamma}, \tag{13}$$

where s and γ are new introduced hyper-parameters, with $s > 1$ and $\gamma > 0$ to control the sensitivity of α to confusion strength. In this work, we set $s = 5$ and $\gamma = 0.5$ and the results of different s and γ is shown in the Figure 10.

G. Additional Materials on the Diff-Manner Adapter in SAM

We introduce the Diff-Manner Adapter in the Sample Confusion Manner (SAM), which adaptively captures typical confusion features from representative confusion samples from both global and local perspectives. As shown in Figure 5, Grad-CAM visualizations are provided for global-only, local-only, and fused feature representations, with corresponding accuracy results summarized in the accompanying table 10. It is noteworthy that when only global features are used, the accuracy of base classes drops significantly, while using only local features leads to a clear decline in novel class accuracy. This indicates that global and local features play complementary roles in supporting the classification decisions for base and novel classes. Framework of Equal 9,Equal 10 when capturing local sample confusion feature is shown in Figure 8.

H. Effects of Different Confusion Pair and Representative Sample Number.

Figure 9 illustrates the influence of the number of representative samples in the Sample Confusion Miner and confusion pairs in the Semantic Confusion Miner. We observe

Table 10. Impact of global and local part in Diff-Manner Adapter.

Method	Base	Novel	HM
global	81.22	76.23	78.65
local	85.19	72.88	78.56
global+local	87.41	80.90	83.90

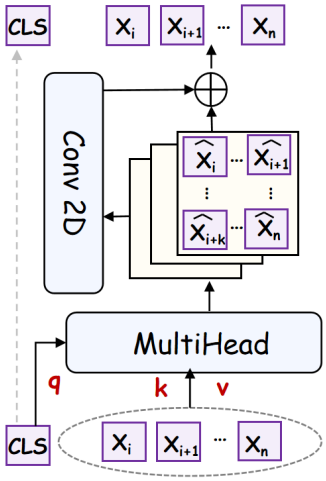

Figure 8. Framework of Equal 9,Equal 10 in Diff-Manner Adapter.

that retaining one representative sample per confusing category yields the best performance, as excessive samples introduce redundant confusion information, making it harder for the model to learn effectively. For confusion pairs, the best results are achieved when five pairs are retained; too few pairs fail to capture sufficient confusion information, while too many introduce unnecessary noise that hinders the model's confusion learning.

I. Further Evidence of the Fixed Confusion Pattern

We observe a fixed confusion pattern in which certain categories are consistently misaligned into specific target classes with significantly higher probability. This observation provides a foundation for CAPT to address the alignment problem. Figure 1 illustrates the confusion pattern observed on the Oxford dataset. This finding is further substantiated by the misalignment analysis across multiple datasets in Figure 11. For instance, in the dataset Caltech101, 'cougar_body' is misclassified as 'cougar_face' 8 times with almost no other errors; in the dataset Food101, 'cake' is misclassified as 'chocolate' 41 times exclusively. This pattern is further underscored by a striking example from the dataset FGVCAircraft, where 'Permanent Crop Land' is misclassified as 'Pasture Land' 230 times—a frequency orders of magnitude higher than other single-digit

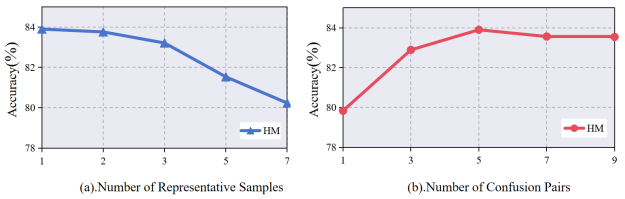

Figure 9. Effects of Different Confusion Pair in SEM and Representative Sample Number in SAM.

error rates. These results collectively demonstrate that the fixed confusion pattern is not an isolated occurrence, but a pervasive systemic bias, revealing a fundamental limitation in current models for fine-grained recognition.

J. Impact of Confusing Sample Quality

To further evaluate the model's sensitivity to misclassified samples, we randomly injected 5% and 10% Gaussian noise into the selected most representative confusion samples. Results are reported in the Table 11. We observe that introducing a small amount of noise leads to only marginal performance degradation. This not only demonstrates the robustness of our method to perturbations, but also indicates that our statistics-based design effectively mitigates the impact of sample noise.

Table 11. Effects of Noise Add.

Method	Base	Novel	HM
Random Add 5% Gaussian Noise	79.87	74.49	77.09
Random Add 10% Gaussian Noise	72.41	69.56	70.96

Table 12. Statistics of the datasets used in our experiments

Dataset	Classes	Train	Val	Test	Task	Hand-crafted Prompt
ImageNet	1,000	1.28M	N/A	50,000	General object recognition	"a photo of a [CLASS]."
Caltech101	100	4,128	1,649	2,465	General object recognition	"a photo of a [CLASS]."
EuroSAT	10	13,500	5,400	8,100	Remote sensing classification	"a satellite image of [CLASS]."
SUN397	397	15,880	3,970	19,850	Scene classification	"a photo of a [CLASS]."
DTD	47	2,820	1,128	1,692	Texture classification	"a [CLASS] texture."
UCF101	101	7,639	1,808	3,783	Human action recognition	"a photo of a person doing [CLASS]."
FGVCAircraft	100	3,334	3,333	3,333	Fine-grained aircraft recognition	"a photo of a [CLASS], an aircraft type."
OxfordPets	37	2,944	736	3,669	Fine-grained pet recognition	"a photo of a [CLASS], a pet breed."
StanfordCars	196	6,509	1,635	8,041	Fine-grained car recognition	"a photo of a [CLASS], a car model."
Flowers102	102	4,093	1,633	2,463	Fine-grained flower recognition	"a photo of a [CLASS]."
Food101	101	50,500	20,200	30,300	Food image classification	"a photo of a [CLASS], a type of food."
ImageNet-V2	1,000	N/A	N/A	10,000	ImageNet distribution shift	"a photo of a [CLASS]."
ImageNet-Sketch	1,000	N/A	N/A	50,899	ImageNet distribution shift	"a sketch of a [CLASS]."
ImageNet-A	1,000	N/A	N/A	7,500	Natural adversarial examples	"a photo of a [CLASS]."
ImageNet-R	1,000	N/A	N/A	30,000	Artistic rendition recognition	"a rendition of a [CLASS]."

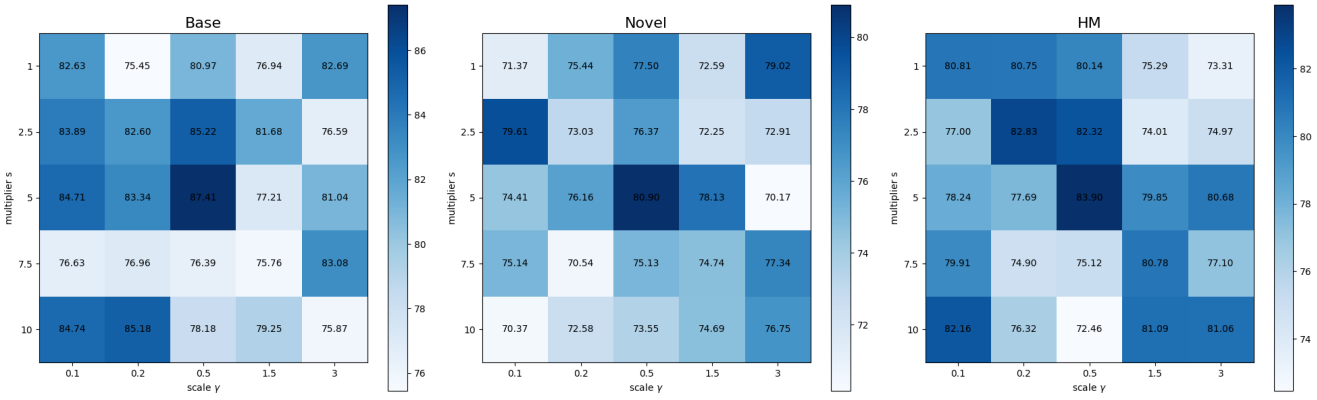

Figure 10. Effects of different s and γ in choosing adaptive α.

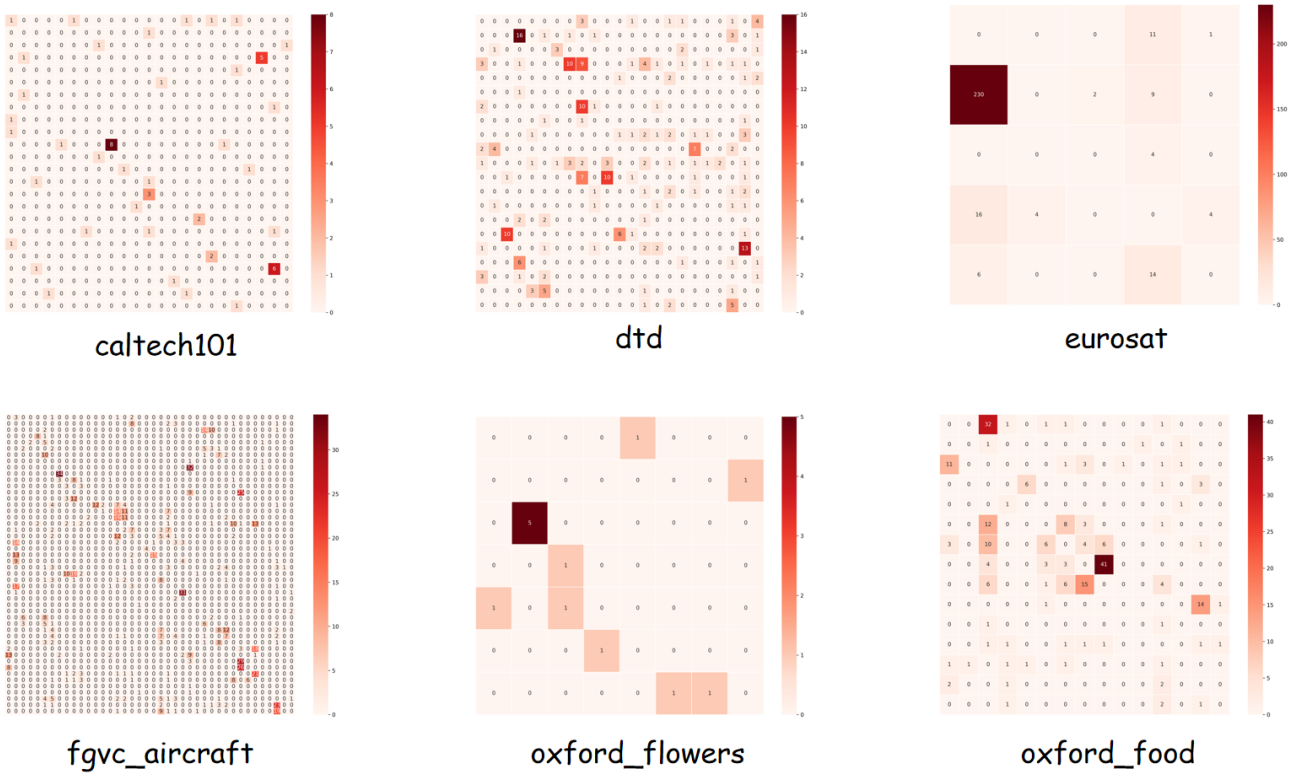

Figure 11. Heatmaps of model misclassifications across more datasets prove the fixed confusion pattern.