Supplementary Material of Manuscript ID FCS-250515

A. Ablation studies

As the instance-dependent visual fusion module in INFUSE integrates visual features from two sources, the original images and the CLIP image encoder, we conduct the ablation studies to asses the individual contributions of each component, offering insights into the effects of integrating textual prompts with different types of visual features. In addition, we also evaluate performance both with and without the fusion of visual features into textual prompts.

Table 1 presents the results of the ablation study conducted on the CIFAR-10 and TinyImageNet datasets under open-set scenarios. In this setup, models were trained on known classes (6 for CIFAR-10, 20 for TinyImageNet) and evaluated on both known and unknown classes to thoroughly assess open-set recognition performance. The experiment using learnable textual prompts without visual feature interaction serves as the baseline. Our experimental results demonstrate that the interaction with both types of visual features consistently enhances open-set recognition performance. Specifically, incorporating CLIP visual features improves AUROC by 2.1% on CIFAR10 but decreases by 1.1% on TinyImageNet. Adding original image features yields consistent improvements of 2.4% on CIFAR10 and 0.3% on TinyImageNet. When combining both visual feature types, AUROC improves by 3.5% on CIFAR10 and 0.8% on TinyImageNet. These results indicate that original image features provide more consistent improvements across datasets, while the combined approach achieves optimal performance, demonstrating the effectiveness of our framework. This superior performance can be attributed to how INFUSE addresses fundamental limitations in existing approaches.

Previous methods often focus on adapting learnable textual prompts for known classes, but there's an inherent gap between textual prompts and visual features. Additionally, the textual prompts in VLMs are primarily biased towards their training data, which is typically inaccessible, potentially widening this gap. INFUSE introduces a novel attention fusion module that integrates visual features from CLIP as well as additional features directly extracted from original images, which helps mitigate these biases. By infusing these enriched visual features with textual prompts through a lightweight network, INFUSE effectively reduces the gap between textual prompts and visual features, leading to better open-set recognition performance.

Table 1: Ablation studies of visual features from two sources in INFUSE. 'Textual' refers to the textual prompts, $\boldsymbol{v}^{clip}$ denotes the features extracted by the CLIP image encoder, and $\boldsymbol{v}^{orig}$ represents visual features extracted from the original images in the instance-dependent visual feature module.

Textual	$\boldsymbol{v}^{clip}$	$\boldsymbol{v}^{orig}$	CIFAR10		TinyImageNet	
			AUROC	OSCR	AUROC	OSCR
✓			92.7	84.1	88.4	81.3
✓	✓		94.8	91.3	87.3	81.7
✓		✓	95.1	91.4	89.5	83.7
✓	✓	✓	97.1	94.7	91.6	86.6

B. Experiments on computational efficiency

To address the computational efficiency concerns, we analyze our method across three metrics: total parameters, trainable parameters, and inference time, as shown in Table 2. The parameter increase originates from four components: the learnable text prompts, the original image feature extraction, the attention fusion mechanisms, and the lightweight adaptation networks, with feature extraction contributing the majority. Our ablation study shows that adding feature extraction leads to a significant performance boost (1.4% on TinyImageNet, 4.3% on CIFAR10). Importantly, the added model capacity enhances recognition performance while maintaining real-time applicability, thereby achieving meaningful gains without compromising deployment feasibility.

C. CAM visualization

We employ Grad-CAM heatmap visualization to examine how instance-dependent visual fusion enhances model performance from an interpretability perspective. Fig. 1 illustrates the comparative analysis of CAM activation patterns between our INFUSE approach and the

Table 2: The total number of parameters, the trainable parameters and the test time on TinyImageNet compared with different methods based on the CLIP model.

Methods	Params (M)	Trainable Params(M)	Test Time (s/image)
CoOp	125.99	0.0081	0.01
CoCoOp	126.02	0.0415	0.02
A^2Pt	130.22	4.2388	0.01
INFUSE(Ours)	143.50	17.5248	0.02

baseline method without instance-dependent visual fusion. As shown in the figure, our method achieves more precise and focused activation regions. In the bee example, while the baseline method incorrectly attends to an irrelevant flower in the background, our model specifically focuses on the bee itself, demonstrating improved target localization. This enhanced localization capability may stem from the instance-dependent visual fusion mechanism, which dynamically extracts comprehensive, target-aware visual features tailored to each specific input instance.

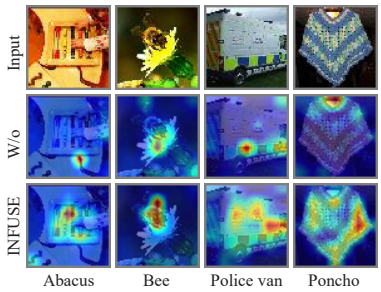

Figure 1. : Comparison of CAM activation effects. Top: Input image; Middle: Without instance-dependent visual fusion; Bottom: Our proposed INFUSE model with instance-dependent visual fusion. The results demonstrate improved activation accuracy and coverage for target regions.