Feature Purification Matters: Suppressing Outlier Propagation for Training-Free Open-Vocabulary Semantic Segmentation

Supplementary Material

A. Overview

In this supplementary material, we provide more analysis of CLIP's attention mechanism (Sec. B), more ablation experiments about our proposed modules (Sec. C), more qualitative segmentation results (Sec. D), and more discussions(Sec. E).

B. More analysis of CLIP's attention

We conduct an in-depth analysis of CLIP's attention distribution using ViT-B/16 as the backbone model. Different attention maps are visualized in Fig. 1, which reveals the emergence of outliers and their distinct distribution.

First, as illustrated in Fig.1 (a), CLIP's attention maps exhibit a distinctive 'band-like' stripe pattern across different layers, particularly from the 6th layer onward. This pattern indicates that regardless of which image token is selected, its attention is consistently drawn to specific tokens, *i.e.* outliers, within these stripes. To further investigate this phenomenon, we randomly select image tokens and visualize their attention maps, as shown in Fig. 1 (b) & (c). It is observed that once outliers emerge, their locations remain nearly identical across different randomly chosen image tokens. This observation strongly correlates with the 'band-like' stripe structure observed in Fig. 1 (a). Furthermore, we analyze the attention map of the class token and find that its outlier distribution aligns closely with that of the randomly selected image token, as shown in Fig. 1 (d). This phenomenon suggests that all input tokens consistently focus on the same outliers. Moreover, the class token is prone to exhibit a stronger response to the outlier than the image tokens in the attention map [1]. Based on this, we argue that the self-relevance of image tokens and the similarity between the class token and image token within the attention map can act as an effective detector for outliers. Beyond that, it's seen that outliers do not appear in each layer, yet they emerge in the deeper layers. Our proposed SOM can be deployed at each CLIP's image encoder layer to detect outliers adaptively. Notably, if SOM doesn't detect any outliers at a given layer, the original attention map and features will be kept and propagated to the following layers.

C. More Ablation Studies

Visualization of feature purification of SOM We visualize the outlier detection results on different input images. As shown in Fig. 2, our SOM adaptively identifies outliers based on the distinct patterns of different input images without any manual settings.

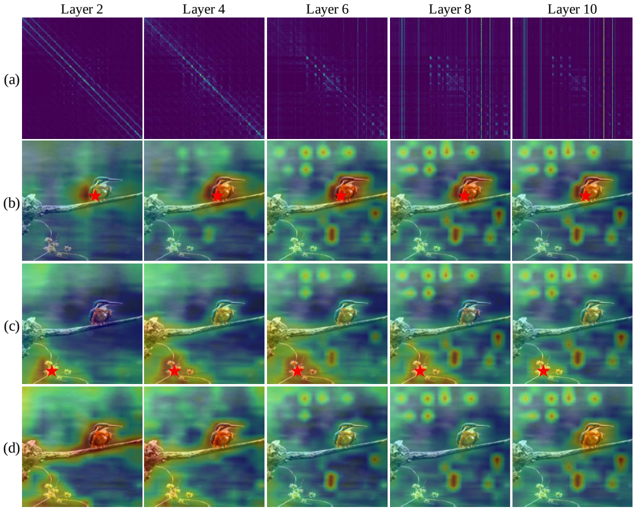

Figure 1. In-depth of CLIP's attention mechanism across various layers. (a) Attention maps of image tokens. (b) & (c) Attention maps of the selected image token, marked as ⋆, (d) Attention maps of the class token.

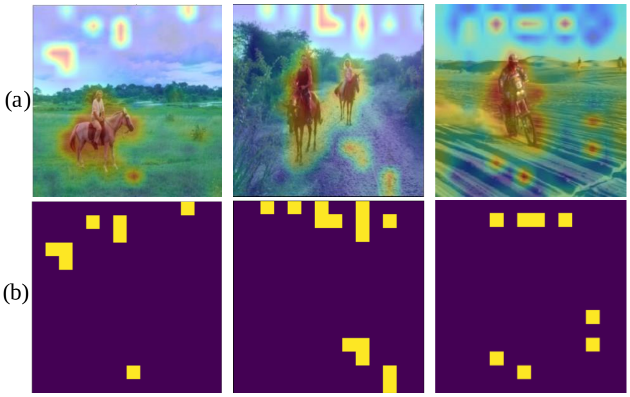

Figure 2. Visualization of outlier detection for different input images. (a) Image features with outliers. (b) Detected outliers by our SOM. It's seen that our SOM can adaptively identify different numbers of outliers for feature purification.

Effect on semantic-aware attention enhancer As we deploy self-self attention in our SAE to augment semantic coherence, we also explore different designs of self-self attention. We conduct the thorough experiment by taking Q-K

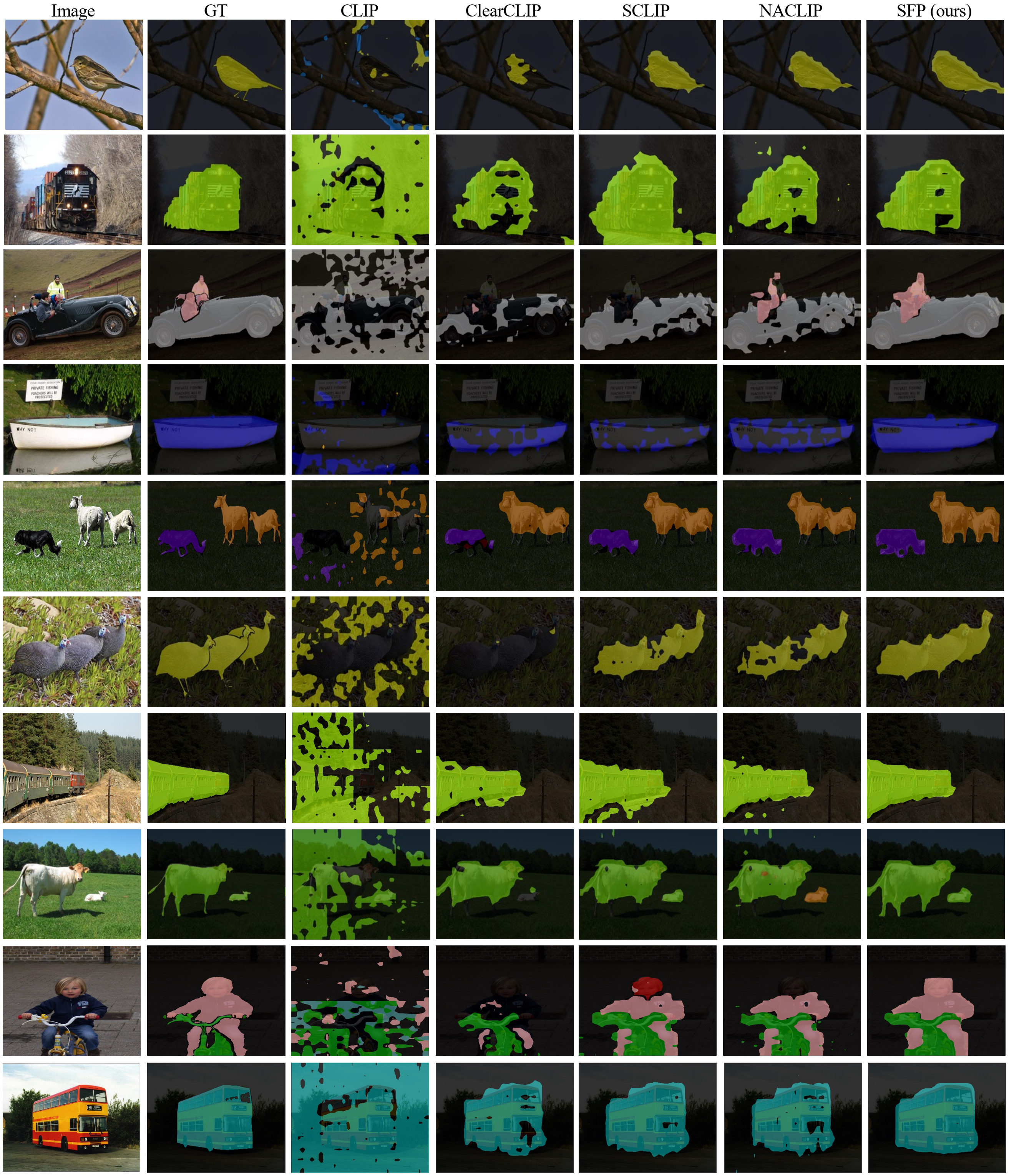

Figure 3. More Open-Vocabulary Segmentation Results. We compare our SFP with CLIP [4], ClearCLIP [3], SCLIP [5] and NACLIP [2], all without post-processing. Our SFP produces much clearer and more accurate segmentation results.

attention as the baseline without SAE and HAI in Tab. 1, where the '+' means directly adding the two kinds of attention maps. It's shown that different designs of self-self attention achieve improvements over the baseline. Among them, the 'Q-Q + K-K' combination surpasses the baseline by an average of 1.7% mIoU. This result indicates that such a combination can help our SAE provide better token relationships for feature purification.

attention mechanism	V21	ADE	Avg.
Q-K (baseline)	60.1	18.2	39.2
Q-Q	62.1	19.3	40.7
K-K	62.0	19.3	40.7
V-V	61.4	18.9	40.2
Q-Q + K-K	**62.3**	**19.5**	**40.9**
Q-Q + V-V	61.9	19.1	40.5
K-K + V-V	61.1	18.7	39.9

Table 1. Ablation results on different designs of self-self attention mechanism in our SAE.

D. More Visualization Comparison

Fig. 3 presents more segmentation visualizations. We compare our SFP with CLIP [4], ClearCLIP [3], SCLIP [5] and NACLIP [2]. These results highlight that our SFP consistently delivers higher-quality and more precise segmentation maps than other approaches.

E. More discussions

E.1. About post-processing

Tab. 2 lists the post-process with PAMR results. PAMR consistently improves performance, though the smaller gain in our method suggests SFP can achieve good performance without post-processing.

Methods	ClearCLIP		SCLIP		SFP	
post-process	✗	✓	✗	✓	✗	✓
voc20	80.9	81.5	81.5	83.1	84.5	84.9

Table 2. Comparison with SOTA methods via post-processing.

E.2. Performance effect on image-level global tasks

We compare our method with the original CLIP using the class token and a single prompt 'a photo of {}' (Tab. 3). As expected, our method shows limited classification performance, likely due to the removal of outliers that disrupt class token representation. This result aligns with the previous work [1], which shows that outlier tokens contribute significantly to classification. Our training-free method might degrade the class token, yet we argue that mitigating outliers during training is a more promising direction. However, our task depends on patch tokens while outliers impair segmentation. Removing them enhances patch token quality and improves performance.

Benchmarks	Cifar100		Flowers102		OxfordPets	
Accuracy	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
CLIP	66.6	88.5	67.7	84.5	89.1	99.5
Ours	56.7	80.8	49.6	74.4	82.4	91.6

Table 3. Zero-shot classification performance with the class token.

References

[1] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *ICLR*, pages 1–11, 2024. 1, 3

[2] Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. Pay attention to your neighbours: Training-free open-vocabulary semantic segmentation. In *WACV*, pages 5061–5071, 2025. 2, 3

[3] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Clearclip: Decomposing clip representations for dense vision-language inference. In *ECCV*, pages 143–160, 2024. 2, 3

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 2, 3

[5] Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. In *ECCV*, pages 315–332, 2024. 2, 3