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

Shuo Jin[1,2] Siyue Yu[1*] Bingfeng Zhang[3] Mingjie Sun[4] Yi Dong[2] Jimin Xiao[1]
[1]Xi'an Jiaotong-Liverpool University [2]University of Liverpool
[3]China University of Petroleum (East China) [4]Soochow University
{shuo.jin, yi.dong}@liverpool.ac.uk, {siyue.yu02, jimin.xiao}@xjtlu.edu.cn
bingfeng.zhang@upc.edu.cn, mjsun@suda.edu.cn

Abstract

Training-free open-vocabulary semantic segmentation has advanced with vision-language models like CLIP, which exhibit strong zero-shot abilities. However, CLIP's attention mechanism often wrongly emphasises specific image tokens, namely outliers, which results in irrelevant over-activation. Existing approaches struggle with these outliers that arise in intermediate layers and propagate through the model, ultimately degrading spatial perception. In this paper, we propose a Self-adaptive Feature Purifier framework (SFP) to suppress propagated outliers and enhance semantic representations for open-vocabulary semantic segmentation. Specifically, based on an in-depth analysis of attention responses between image and class tokens, we design a self-adaptive outlier mitigator to detect and mitigate outliers at each layer for propagated feature purification. In addition, we introduce a semantic-aware attention enhancer to augment attention intensity in semantically relevant regions, which strengthens the purified feature to focus on objects. Further, we introduce a hierarchical attention integrator to aggregate multi-layer attention maps to refine spatially coherent feature representations for final segmentation. Our proposed SFP enables robust outlier suppression and object-centric feature representation, leading to a more precise segmentation. Extensive experiments show that our method achieves state-of-the-art performance and surpasses existing methods by an average of 4.6% mIoU on eight segmentation benchmarks. The code is released at: https://github.com/Kimsure/SFP.

1. Introduction

Open-vocabulary semantic segmentation (OVSS) [4, 58] seeks to partition an image into distinct regions and assign pixel-level labels to arbitrary semantic categories. Unlike traditional semantic segmentation, OVSS must handle an open set of classes, making it more challenging. Vision-language models (VLM), such as CLIP [38], demonstrate remarkable zero-shot performance by leveraging large-scale image-text pairs [39]. Thus, existing approaches can be categorized into two pipelines: fine-tuning CLIP or training-free CLIP. Although the fine-tuning pipeline yields superior performance, it risks overfitting and compromises generalization [30, 43]. In contrast, the training-free pipeline preserves CLIP's original generalization capability without additional training, which is more convenient and efficient.

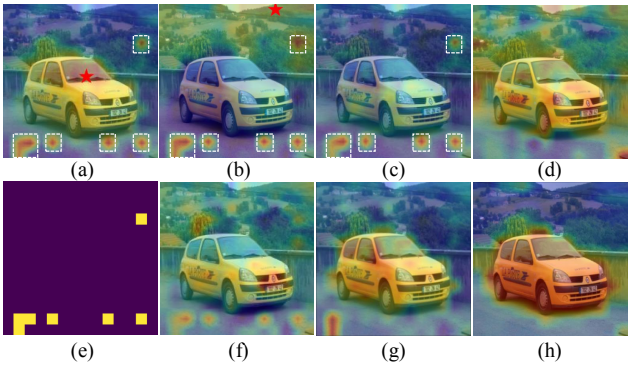

Figure 1. Visualization of attention and feature maps. (a) Attention map of the selected target object token (marked as ⋆). (b) Attention map of the selected background token (marked as ⋆). (c) Attention map of the class token. (d) Diagonal weights of the attention map of each image token. (e) Detected outliers. (f) - (h) output feature map of CLIP [38], SCLIP [45], and our SFP.

Recent training-free methods [12, 40] have demonstrated that in the original CLIP's pair-wise attention maps, there exist the same highly correlated image tokens for all input tokens, such as those marked with the white box in Fig. 1 (a) - (b). In such a situation, regardless of whether it is the target image token (Fig. 1 (a)), the background image to-

*Corresponding author: siyue.yu02@xjtlu.edu.cn

ken (Fig. 1 (b)), or the class token (Fig. 1 (c)), the pair-wise token relationships within the attention map are all over-activated for these identical tokens, *i.e.*, *outliers*. Consequently, when aggregated with outliers, the derived semantic feature representation is easily polluted to overemphasize these outliers and the interpretation of the target region is impeded, *e.g.*, the feature map of 'car' in Fig 1 (f) focuses on the outliers instead of the object itself.

To mitigate outliers and ensure more accurate semantic feature representation, many training-free approaches have presented diverse modifications to the self-attention mechanism in the last layer [14, 16, 24, 45]. While these strategies effectively restrict CLIP's attention to ignore outliers, some remain unfiltered, as shown in Fig. 1 (g). Moreover, some approaches [2, 12] rely on empirical parameter tuning to detect and remove a fixed number of outliers, which is less robust to different backbones and images. Beyond these optimization approaches, several studies [12, 40] further reveal that outliers predominantly appear in the intermediate layers. In this way, once outliers emerge, they will propagate through subsequent layers, continuously affecting semantic feature representations. Yet, previous approaches fail to analyze and address outliers in the intermediate layers so that outliers are easily propagated to pollute the final semantic feature. Thus, our objective is to adaptively eliminate outliers across multiple layers to achieve more robust semantic feature representations without training.

Building on these insights, we propose a **S**elf-adaptive **F**eature **P**urifier framework (**SFP**) to resolve outliers adaptively and enhance semantic representations for CLIP-based training-free OVSS. Firstly, we design a Self-adaptive Outlier Mitigator (SOM) to identify and mitigate outliers across multiple layers. Tokens are expected to focus primarily on their relevant regions, especially themselves, rather than irrelevant regions in the attention map. However, as shown in Fig. 1 (d), we observe that the self-response of image tokens in the attention map is not always prominent, especially outliers. The comparison of Fig. 1 (c) & (d) reveals that outliers exhibit a stronger response to the class token than their self-response. Hence, our SOM can identify outliers by simply computing the difference between the self-response values of the image tokens in the attention map (*i.e.*, the diagonal weights of the attention map) and their attention values with the class token. The detected outliers are shown in Fig. 1 (e). In this way, our parameter-free SOM enables adaptive detection across various layers to mitigate the propagating influence of outliers without manual tuning.

Nevertheless, outliers not only cause tokens to overemphasize them but also impede tokens' responses to semantic regions in the attention map. Therefore, merely removing the outliers can't directly increase the attention weights of the related regions. To enhance the tokens' attention responses to semantic regions, we propose a Semantic-aware Attention Enhancer (SAE). SAE incorporates the self-self attention mechanism in CLIP's last layer and applies it to the purified feature derived from SOM at the penultimate layer, enabling a better focus on semantic regions. In addition, shallow attention maps typically contain structured object information and focus on relevant semantic regions [27, 30]. To further enhance the precise spatial perception of the final feature map, we propose a Hierarchical Attention Integrator (HAI) as an assistant to the SAE, where the attention maps in previous layers are firstly processed to mitigate the influence of the outliers and then integrated to refine the final semantic feature for prediction. The refined semantic feature for the final prediction is shown in Fig. 1 (h). Our SFP can generate a more purified and complete semantic feature for segmentation.

We conduct comprehensive experiments to evaluate the proposed **SFP**. Our method outperforms existing methods and achieves state-of-the-art (SOTA) performance across eight evaluation datasets. In summary, our contributions are concluded as follows:

- We present a Self-adaptive Feature Purifier framework (SFP) for training-free OVSS, which can derive a more accurate semantic feature to strengthen the final open-vocabulary semantic segmentation.
- We propose a simple yet effective mechanism, Self-adaptive Outlier Mitigator (SOM), which adaptively detects and mitigates outliers across CLIP's diverse layers so that the influence of the propagated influence of outliers can be suppressed.
- We introduce a Semantic-aware Attention Enhancer (SAE) and a Hierarchical Attention Integrator (HAI) to further refine the semantic feature relationships on relevant regions, aggregating the more relevant semantic feature for final prediction.
- Comprehensive experiments demonstrate that SFP derives the best performance under a fair comparison *e.g.*, SFP achieves a gain of 5.6% mIoU on Cityscapes and an average gain of 4.6% mIoU across eight datasets.

2. Related Work

2.1. pre-trained Vision-Language Models

Pre-trained vision-language models [17, 26, 38] have drawn significant attention for their ability to bridge visual and textual modalities. CLIP [38] has emerged as a highly successful model trained on a large-scale dataset of image-text pairs using contrastive learning. OpenCLIP [9] builds upon CLIP by leveraging public datasets such as LAION [39]. Despite their strong zero-shot capabilities, these models are pre-trained at the image level for classification. This presents a notable limitation: the attention maps of VLMs tend to overemphasize outliers rather than semantic image tokens, leading to insufficient capture of spatial details. This draw-

back hinders the performance in dense prediction tasks like semantic segmentation, which demands precise pixel-level understanding. Our goal is to adaptively detect and mitigate the impact of these outliers and enhance the semantic feature representation for dense prediction tasks.

2.2. Open-vocabulary Segmentation

Semantic segmentation involves the pixel-wise classification of an image. Compared with the traditional segmentation methods [6, 8, 31, 49] that are trained and evaluated on the same fixed set of seen classes. OVSS methods [36, 59] aim to recognize and segment unseen objects during inference by harnessing the zero-shot ability of VLMs. Existing works can be broadly divided into two categories: training-based and training-free.

Training-based Methods Training-based methods [18, 37, 56] typically require fine-tuning on a fixed set of categories from a pixel-level annotated dataset. Some approaches follow a two-stage pipeline, where the first stage extracts the mask proposals, and the second stage assigns semantic labels to them. For instance, OVSeg [51] first trains class-agnostic mask proposals using the query-based framework Mask2Former [8] and then fine-tunes CLIP [38] to classify the cropped and masked images. Other approaches adopt a single-stage pipeline. SAN [52] introduces a side adapter to adapt CLIP [38] for both classification and segmentation. SED [48] presents a simple encoder-decoder architecture with category early rejection for fast inference. CAT-Seg [10] constructs a pixel-level cost map for segmentation. However, these training-based methods are prone to overfitting the training dataset, which limits the generalization ability due to fine-tuning CLIP.

Training-free Methods Unlike training-based methods, training-free methods [20, 41] aim to directly adapt VLMs for OVSS without extra training. Several approaches leverage the vision foundation models [5, 21, 42] to improve CLIP's spatial coherence. LaVG [19] uses DINO [5] for panoptic cuts, which iteratively discovers object masks until all image pixels are covered. ProxyCLIP [25] adjusts attention weights based on affinities learned from DINO's feature correspondence. However, these methods fail to exploit CLIP's inherent potential. Other efforts aim to modify CLIP's final layer to address its poor spatial consistency. For example, MaskCLIP [14] only utilizes the value feature of the final layer for prediction. CLIPtrase [40] integrates self-self attention with clustering for post-processing. SCLIP [45] and ClearCLIP [24] both modify the last layer self-attention mechanism to better represent the local information. Nevertheless, few training-free methods address propagated outliers, which act as noisy inputs across multiple layers and degrade spatial perception. In contrast, our approach enables adaptive detection and mitigation of propagated outliers among various layers, which improves CLIP's semantic awareness and boosts its internal potential.

3. Method

3.1. Overview

Fig. 2 depicts the whole framework of SFP, including four main modules: a frozen CLIP backbone to encode the input image and text descriptions, a SOM to detect and eliminate outliers, an SAE to augment the attention intensity on related semantic regions, and an HAI that leverages intermediate-layer attention maps to assist SAE for enhanced semantic feature representation. The complete inference process is as follows:

1) First, the image and text descriptions are input to the CLIP encoder for the image feature and the text feature. Among the CLIP image encoder, our SOM is added after each transformer layer to mitigate the outliers and generate the purified feature for propagation.
2) Then, the derived final feature map from the last SOM, F_{SOM}, is input to our SAE, which replaces the final transformer layer. SAE emphasizes the attention values of relevant semantic regions to generate the strengthened semantic feature F_{SAE}.
3) After that, our HAI extracts the multi-layer attention maps from the image encoder to further refine F_{SAE} for the final semantic feature F_{vis}.
4) Finally, the segmentation result is obtained by calculating the cosine similarity of the final semantic feature F_{vis} and text feature F_{txt}.

3.2. Preliminary

In general, the training-free OVSS approach requires CLIP [38], which encompasses an image encoder and a text encoder to acquire multi-model features. Subsequently, these features are aligned via linear projection layers. In detail, the input image is first partitioned into a token sequence $X_{in} = [x_{cls}, x_1, ..., x_N] \in R^{(N+1)\times d}$, where x_i denotes the image token embedding, x_{cls} denotes the class token embedding, N is the image token sequence length, and d is the dimension of token embeddings, respectively. Then, these tokens are input to the image encoder, which contains a stack of vision transformer layers.

Each transformer layer consists of a multi-head self-attention (MHSA) mechanism and a feed-forward network (FFN) layer to extract the image feature. Specifically, in each MHSA, the input sequence X_{in} is mapped to query, key, and value embeddings, $Q, K, V \in R^{(N+1)\times d}$. Afterwards, Q and K are calculated to obtain the attention map $Attn \in R^{(N+1)\times(N+1)}$, which can be formulated as:

$$Attn = \text{Softmax}(\frac{QK^T}{\sqrt{d}}), \tag{1}$$

where T means matrix transposition. After that, the attention map is multiplied V to aggregate the global infor-

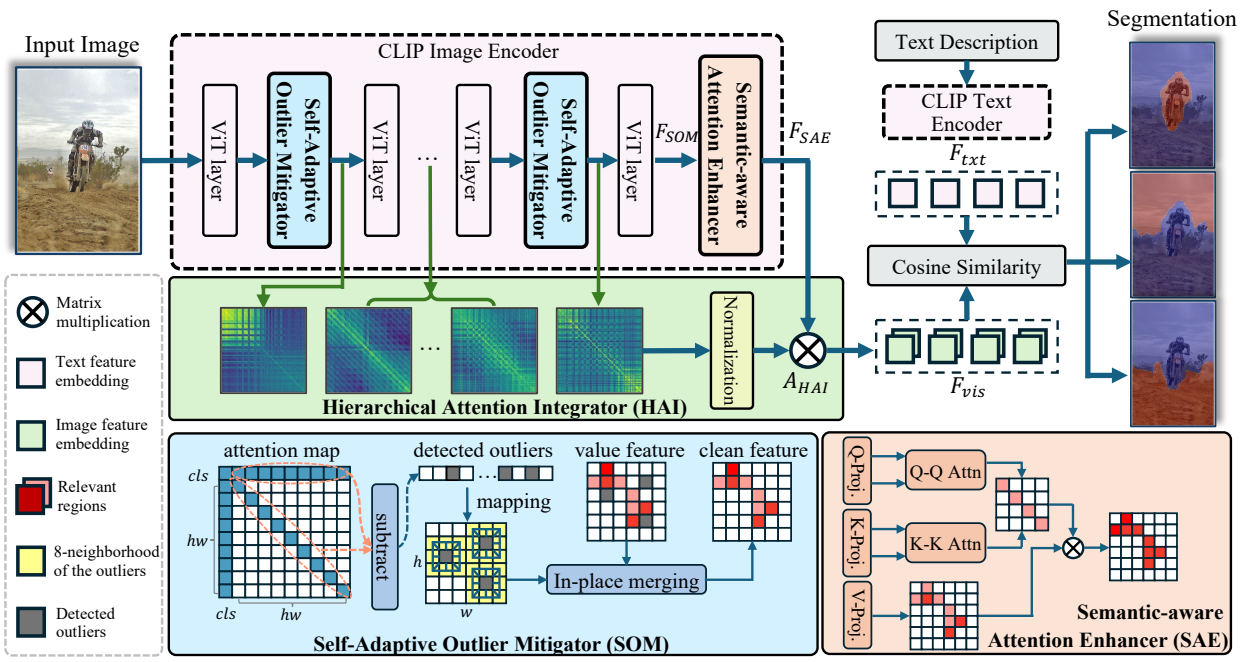

Figure 2. Framework of our SFP. It contains four main modules: a frozen CLIP backbone for feature extraction, a Self-adaptive Outlier Mitigator (SOM) at each image encoder layer for outlier detection and elimination, a Semantic-aware Attention Enhancer (SAE) at the last layer to emphasise attention values of relevant semantic regions and a Hierarchical Attention Integrator (HAI) for object-centric feature representation. The segmentation result is generated by calculating the cosine similarity between the optimized semantic feature F_{vis} and the text feature F_{txt}.

mation, and an FFN is used to obtain the image feature $F_{img} \in R^{(N+1)\times d}$, formulated as:

$$F_{img} = \mathrm{FFN}(AttnV). \tag{2}$$

Since the class embedding x_{cls} is not used to predict the segmentation result, we omit it and reshape F_{img} to obtain the visual feature $F_{vis} \in R^{h\times w\times d}$, where h and w denote the height and width of the feature map.

Meanwhile, the textual description is derived from the standard ImageNet [13] prompts and transformed to the text feature $F_{txt} \in R^{c\times d}$ with CLIP's text encoder, where c represents the number of classes. By calculating the cosine similarity between F_{vis} and F_{txt}, the segmentation map $Seg \in R^{H\times W}$ can be generated like:

$$Seg = U(\arg\max_{c} \cos(F_{vis}, F_{txt})), \tag{3}$$

where $\cos(\cdot)$ means cosine similarity calculation and $U(\cdot)$ denotes the upsampling operation to match the original input image size. H and W denote the height and width of the input image, respectively.

3.3. Self-adaptive Outlier Mitigator

CLIP [38] is pre-trained at the image level for classification, which tends to concentrate on the most discriminative feature that is important for classification. In this way, all input tokens, including both class and image tokens, will emphasize these identical tokens (outliers) in attention maps [12]. In other words, the corresponding outlier column in the attention map will be highlighted. Further, given that the classification process relies solely on the class token, the class token may exhibit a stronger response to the outlier in attention maps compared to image tokens. This prompts us to investigate whether the self-relevance of image tokens and the class token's similarity across all image tokens could serve as potential indicators to identify outliers.

Based on the above assumption, to robustly mitigate outliers, we introduce a Self-adaptive Outlier Mitigator (SOM) to detect and eliminate the outliers automatically. In particular, we use the attention map $Attn$ to obtain the self-attention weights of image tokens, denoted as $Attn_{i,i}$, and their attention weights to the class tokens, denoted as $Attn_{cls,i}$, where i denotes the i-th image token and cls denotes the class token. By comparing the difference between $Attn_{i,i}$ and $Attn_{cls,i}$, we obtain the set of outliers $\mathcal{S}$, since the self-attention weights should be higher than others. This detection process of SOM is formulated as,

$$\mathcal{S} = \{i \mid Attn_{i,i} < Attn_{cls,i}\}, \quad i \in [1, N]\,, \tag{4}$$

where $\mathcal{S}$ denotes the set of outliers.

After recognizing outliers, we propose eliminating them at the feature level rather than modifying the attention map

to prevent potential model collapse. Inspired by previous work [2], it's crucial to ensure the semantic consistency of the feature map, and directly masking outliers at the feature level is not an appropriate approach. Thus, SOM can eliminate outliers through in-place merging, where the detected outliers are replaced with the average value of its 8-neighbours. That is to say, the derived F_{img} in the original attention mechanism will be updated through our SOM as

$$\hat{F}_{img}[u] = \begin{cases} \frac{1}{8} \sum_{v \in \mathcal{N}_8(u)} F_{img}[v], & u \in \mathcal{S} \\ F_{img}[u], & \text{otherwsie} \end{cases}, \quad (5)$$

where $\mathcal{N}_8(u)$ means the 8-neighbour non-outlier tokens of the u-th point and $\hat{F}_{img}$ denotes the updated feature. In this way, our SOM initially identifies the outlier and then leverages surrounding token features to replace the detected one. This process can mitigate the adverse influence exerted by the outliers and maintain the feature distribution that is crucial for CLIP's image encoder. By doing so, the purified and superior semantic feature representation is generated.

Furthermore, our SOM exhibits remarkable versatility in that it can be applied to each individual layer of the image encoder. This allows for the detection of outliers and the generation of purified features at multiple levels. In this manner, our parameter-free SOM establishes an efficient and precise mechanism for detecting and eliminating outliers that propagate across various layers of the image encoder. This mechanism is instrumental in preserving accurate semantic information and feature representations throughout the entire image encoder, thereby enhancing its overall performance and reliability.

(a) (b) (c)

Figure 3. Comparison of different feature visualization. (a) Noisy feature with outliers. (b) Purified feature with SOM. (c) Semantic feature with SAE. It's seen that the proposed SAE strengthens the semantic regions with higher object activation.

3.4. Semantic-aware Attention Enhancer

The comparison of Fig. 3 (a) & (b) indicates that the purified feature F_{SOM} generated by the last SOM effectively ignores the irrelevant regions of outliers, marked as white boxes. However, it's also observed from Fig. 3 (b) & (c) that the object activation of F_{SOM} is not salient enough, which indicates that eliminating outliers can't contribute to the attention response intensity.

To address this issue, we propose a Semantic-aware Attention Enhancer (SAE) to strengthen the object activation in CLIP's last layer, where the self-self attention mechanism is used to derive better feature relationships,

$$Attn_{SAE} = \text{Softmax}(\lambda(\frac{QQ^T}{\sqrt{d}} + \frac{KK^T}{\sqrt{d}})), \quad (6)$$

where $Attn_{SAE}$ represents the semantic-aware attention map and λ denotes the logit scale used to sharpen attention scores, set to 0.5. Then, the semantic feature F_{SAE} is obtained through the residual connection with the purified feature F_{SOM} as follows:

$$F_{SAE} = F_{SOM} + \text{FFN}(Attn_{SAE} F_{SOM}). \quad (7)$$

The self-self attention mechanism has shown a strong ability to construct feature relationships [16, 24, 45]. Therefore, applying self-self attention to our purified feature enables the establishment of more precise spatial correlations. These accurate spatial correlations play a crucial role in enabling each point to aggregate the relevant semantic feature effectively. As illustrated in Fig. 3 (b) & (c), SAE significantly augments attention intensity in semantic regions compared to the feature solely processed by SOM.

3.5. Hierarchical Attention Integrator

To further refine the generated semantic feature F_{SAE}, we leverage attention maps from shallow layers to capture discriminative pair-wise feature relationships since shallow layers typically focus on object structures [27]. To this end, we propose a Hierarchical Attention Integrator (HAI), which integrates multi-layer attention maps to refine F_{SAE}.

Instead of using the original attention maps with outliers, HAI integrates attention maps from each transformer layer while excluding the detected outlier weights to avoid information pollution. Firstly, HAI employs the outlier set $\mathcal{S}$ to select the non-outlier weights of each attention map, which can be formulated as,

$$\overline{Attn}^l_{j,i} = \begin{cases} Attn_{j,i}, & \text{if } i \notin \mathcal{S}^l \\ 0 & \text{otherwise} \end{cases}, \quad (8)$$

where $\overline{Attn}^l$ and $\mathcal{S}^l$ denote the filtered attention map and the outlier set at the l-th layer. j and i correspond to the row and column indices of the attention map. Next, HAI integrates $\overline{Attn}^l$ from each layer, formulated as,

$$A = \sum_{l=1}^{L-1} \overline{Attn}^l, \quad (9)$$

where A denotes the integrated attention map, and L denotes the total number of transformer layers.

Methods	Pub. & Year	Extra backbone	Training free	with background			without background					Avg.
				V21	PC60	Object	V20	PC59	Stuff	ADE	City	
GroupVit [50]	CVPR'22	✘	✘	50.4	18.7	27.5	79.7	23.4	15.3	9.2	11.1	29.4
SegCLIP [32]	ICML'23	✘	✘	52.6	24.7	26.5	-	-	-	-	-	-
TCL [7]	CVPR'23	✘	✘	55.0	30.4	31.6	83.2	33.9	22.4	17.1	24.0	37.2
DINOiser† [47]	ECCV'24	DINO	✘	62.1	32.4	34.8	80.9	35.9	24.6	20.0	31.7	40.3
SAM-CLIP† [46]	ECCV'24	SAM	✘	60.6	29.2	-	-	-	-	17.1	-	-
PnP-OVSS† [33]	CVPR'24	BLIP	✔	-	-	36.2	51.3	28.0	17.9	14.2	-	-
LaVG† [19]	ECCV'24	DINO	✔	62.1	31.6	34.2	82.5	34.7	23.2	15.8	26.2	38.8
ProxyCLIP† [25]	ECCV'24	DINO	✔	61.3	35.3	37.5	80.3	39.1	26.5	20.2	38.1	42.3
CLIP [38]	ICML'21	✘	✔	18.6	9.9	8.1	49.4	11.1	5.7	3.1	6.5	14.1
MaskCLIP [14]	ECCV'22	✘	✔	38.8	23.6	20.6	74.9	26.4	16.4	9.8	12.6	27.9
CLIPSurgery [28]	PR'25	✘	✔	-	29.3	-	-	-	21.9	-	31.4	-
CaR [44]	CVPR'24	✘	✔	48.6	30.5	36.6	73.7	39.5	-	17.7	-	-
GEM [3]	CVPR'24	✘	✔	46.2	-	-	-	32.6	15.7	-	-	-
CLIPtrase [40]	ECCV'24	✘	✔	50.9	29.9	**43.6**	81.0	33.8	22.8	16.4	21.3	37.5
ClearCLIP* [24]	ECCV'24	✘	✔	57.0	32.6	33.0	80.9	35.9	23.9	16.7	30.0	38.8
SCLIP* [45]	ECCV'24	✘	✔	59.7	31.7	33.5	81.5	34.5	22.7	16.5	32.3	39.1
NACLIP* [16]	WACV'25	✘	✔	58.9	32.2	33.2	79.7	35.2	23.3	17.4	35.5	39.4
SFP (Ours)		✘	✔	**63.9**	**37.2**	37.9	**84.5**	**39.9**	**26.4**	**20.8**	**41.1**	**44.0**

Table 1. Quantitative comparison of our SFP against other approaches across eight segmentation benchmark datasets. The best and second-best results are marked with **bold** and underline, respectively. † indicates that the additional backbone settings follow the original paper. * indicates the results are re-evaluated using the official implementation.

After integrating multi-layer attention maps, we deploy a normalization strategy to optimize A instead of simply averaging it. Inspired by previous works [22, 35], we iteratively apply row-wise and column-wise normalization to A for a more object-centric attention distribution, leading to a stronger focus on relevant regions. This process can be formulated as,

$$A_{HAI} = norm_c(norm_r(A)), \tag{10}$$

where A_{HAI} denotes the normalized attention map, $norm_c$ and $norm_r$ denote the column-wise and row-wise normalization respectively.

Then, A_{HAI} is multiplied with the semantic feature F_{SAE} to obtain the final semantic feature F_{vis} for segmentation, which is formulated as follows:

$$F_{vis} = A_{HAI} F_{SAE}. \tag{11}$$

Finally, we compute the cosine similarity between F_{vis} and the text feature F_{txt} extracted from the CLIP text encoder to predict the segmentation map, as formulated in Eq. 3.

4. Experiment

4.1. Experimental Settings

Datasets We conduct comprehensive experiments to verify the effectiveness of our method on diverse benchmark datasets: PASCAL VOC 2012 [15], PASCAL Context [34], and COCO [29]. These datasets are categorized into two settings: 1) with a background class, namely V20, PC59, and Stuff; 2) without the background class, namely V21, PC60, and Object. We also report results on the ADE20K (ADE) [57] and Cityscapes (City) [11] datasets.

Implementation Details Similar to existing training-free methods, we adopt CLIP ViT-B/16 with the original pre-trained weight [38]. Following previous works [2, 45], we combine standard ImageNet prompts [13] with the class names to construct text descriptions. For a fair comparison with previous works [25, 45], we resize images with a short side of 336 pixels (or 560 pixels for the high-resolution Cityscapes dataset) and apply sliding window inference using a 224 × 224 window with a 112 × 112 stride. Our experiments are conducted using one NVIDIA 3090 GPU.

4.2. Comparison with State-of-the-art Methods

Our SFP is evaluated against existing SOTA methods, including both training-based and training-free approaches. The post-processing mask refinement techniques, such as PAMR [1] or DenseCRF [23], are not included for a fair comparison. The standard mean intersection over union (mIoU) is used as the metric for evaluation.

Quantitative Evaluation The main results are summarized in Tab. 1, where all evaluations are conducted using CLIP ViT-B/16 as the backbone to ensure a fair comparison. We compare SFP against multiple methods, including training-based methods, training-free methods with extra backbones, and training-free methods without extra backbones.

As shown in Tab. 1, our SFP outperforms existing SOTA methods, regardless of whether an additional backbone is introduced or not. In particular, compared with existing

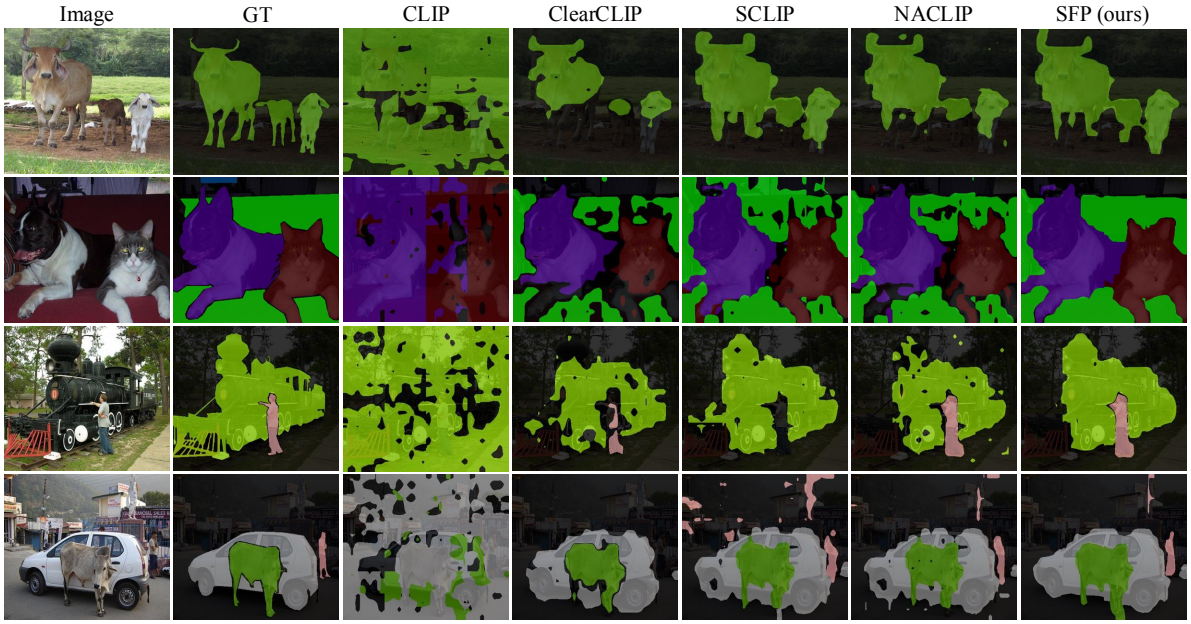

Figure 4. Qualitative Results Comparison. We compare our SFP with CLIP [38], ClearCLIP [24], SCLIP [45] and NACLIP [16] without post-processing. It's observed that SFP generates more precise segmentation results.

training-free methods without extra backbones [16, 24, 45] or cluster methods [53–55], our SFP achieves an average gain of 4.6% mIoU over previous SOTA NACLIP [16]. Specifically, On the V20 and V21 dataset [15], SFP surpasses NACLIP [16] 4.8% and 5% mIoU, respectively. These improvements indicate that our SFP can generate more purified feature maps for more precise segmentation. Beyond that, SFP outperforms SOTA training-free methods with extra backbones, *e.g.*, ProxyCLIP [25], with a gain of 3.0% mIoU on the Cityscape [11] dataset and an average gain of 1.7% mIoU on eight datasets. Our SFP can even beat DINOiser [47], which requires both training and an extra backbone, with an average improvement of 3.7% mIoU. These results highlight our SFP boosts CLIP's internal potential and reveal that feature purification does matter in training-free OVSS, where more purified features can predict better segmentation masks.

Qualitative Evaluation In Fig. 4, we compare the visualization results of various training-free methods without extra backbones, including CLIP [38], ClearCLIP [24], SCLIP [45], NACLIP [16] and our SFP. In particular, the segmentation maps obtained by vanilla CLIP [38] exhibit significant noises, unveiling its limitations in spatial localization. Besides, in complex scenarios, like 'sofa' in the second row, SCLIP [45] and NACLIP [16] fail to capture the complete segmentation mask. In contrast, our SFP can predict more accurate and complete segmentation maps. These results indicate that our SFP has a good ability to predict comprehensive segmentation masks.

4.3. Ablation Study

Comprehensive ablation studies are performed to analyze the effectiveness of our proposed mechanisms. Unless otherwise specified, we utilize CLIP with ViT-B/16 backbone and evaluate on V21 [15] and ADE [57] datasets for background and non-background settings, respectively.

Effect on each proposed module We first conduct several experiments to validate the effectiveness of our proposed modules of our SFP: SOM, SAE, and HAI by adding each module step by step, where ClearCLIP [24] is used as the baseline. As shown in Tab. 2, our SOM achieves a gain of 3.1% and 1.5% mIoU on V21 and ADE datasets, respectively. Further, SAE enhances attention intensity in relevant regions, leading to an average gain of 1.7% mIoU. Finally, HAI helps our SFP reach 63.9% and 20.8% mIoU on V21 and ADE datasets, which set the new SOTA performance. Due to space constraints, more ablation studies about SOM and SAE are summarized in the supplementary material.

	SOM	SAE	HAI	V21	ADE	Avg.
0				57.0	16.7	36.9
1	✔			60.1	18.2	39.2
2	✔	✔		62.3	19.5	40.9
3	✔	✔	✔	**63.9**	**20.8**	**42.4**

Table 2. Ablation results on each proposed module to validate the impact on performance across V21 [15] and ADE [57] datasets.

More visualization on the ability of SOM to suppress outlier propagation We visualize purified feature maps

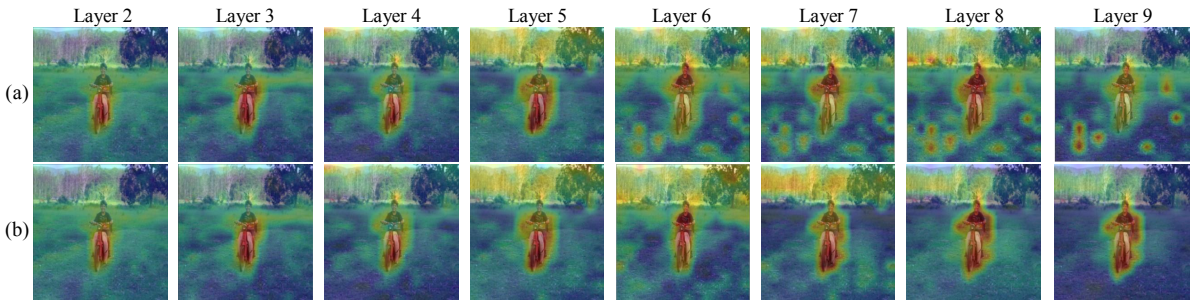

Figure 5. Visualization of the feature purification across CLIP's image encoder layers. (a) Feature maps without SOM purification. (b) Feature maps purified by SOM.

$A_{l=l_0}^{l=L-1}$	1	5	8	10	11
V21	**63.9**	63.4	63.0	62.6	62.5
ADE	**20.8**	20.5	20.3	19.9	19.7
Avg.	**42.4**	42.0	41.7	41.3	41.1

Table 3. Ablation results on different selection strategies in HAI, where $A_{l=l_0}^{l=L-1}$ represents the integrated attention map from l_0 to the $L-1$ layer. '1,5,8,10,11' indicates the value of l_0.

across CLIP's layers, from the 2nd layer to the 9th layer, which is shown in Fig. 5. It's found that the outliers gradually affect the aggregated features, especially from layer 6 to layer 9 in Fig. 5 (a). Then, with the help of our SOM, the influence of outliers is mitigated layer by layer, like in Fig. 5 (b). Thus, our SOM can suppress the propagated outlier features and gradually purify the overall semantic features.

Effect on hierarchical attention integrator A series of ablation studies is conducted to investigate the optimal way to leverage attention maps from different layers. We omit HAI from our SFP to establish the baseline model for this experiment. As summarized in Tab. 3, we take l_0 to represent the starting layer. When $l_0 = 1$, our HAI performs best with all attention maps selected. Besides, increasing l_0 from 1 to 11 decreases the segmentation performance, which demonstrates the effectiveness of the shallow attention maps for semantic feature refinement.

Backbone	Methods	V21	ADE	Avg.
ViT-B/32	SCLIP [45]	50.6	14.8	32.7
	LaVG [19]	54.8	15.5	35.2
	ProxyCLIP [25]	<u>57.9</u>	<u>16.7</u>	<u>37.3</u>
	NACLIP [16]	51.1	14.9	33.0
	SFP	**60.1**	**17.1**	**38.6**
ViT-L/14	SCLIP [45]	44.4	10.9	27.7
	LaVG [19]	52.1	17.3	34.7
	ProxyCLIP [25]	<u>60.6</u>	**22.6**	<u>41.6</u>
	NACLIP [16]	52.2	17.3	34.8
	SFP	**64.7**	<u>22.1</u>	**43.4**

Table 4. Comparison with other SOTA methods across different CLIP backbones.

Effect on different CLIP backbones To validate the robustness, we further evaluate our SFP with other methods using different CLIP backbones, such as ViT-B/32 and ViT-L/14. As summarized in Tab. 4, some methods, like SCLIP [45] and LaVG [19], suffer from performance degradation when applied to the larger backbone. In contrast, SFP maintains superior performance across different backbones and achieves an improvement of 1.3% and 1.8% mIoU over ProxyCLIP [25] with ViT-B/32 and ViT-L/14 backbone, respectively. These results indicate that SFP can robustly handle outliers derived from different backbones and generate purified features for precise segmentation.

5. Conclusion

In this paper, we present a Self-adaptive Feature Purifier framework (SFP) to mitigate outlier propagation for training-free open-vocabulary semantic segmentation. Specifically, we propose SOM deployed at each layer of the image encoder to identify and mitigate outliers by comparing attention weights between image and class tokens, which generates purified features for propagation. Next, to recover the semantic information affected by outliers, we introduce SAE to strengthen object activation in the purified feature. Beyond that, we design HAI to utilize multilayer attention maps for refined object-centric representations. Extensive experiments show that SFP achieves superior segmentation performance compared to existing SOTA methods on various datasets. Our SFP shows that feature purification does matter for OVSS, and cleaner features can predict more complete and accurate segmentation masks.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (No. 62471405, 62331003, 62301451,62301613), Basic Research Program of Jiangsu (BK20241814), Suzhou Basic Research Program (SYG202316) and XJTLU REF-22-01-010 and RDF-22-02-066, XJTLU AI University Research Centre, Jiangsu Province Engineering Research Centre of Data Science and Cognitive Computation at XJTLU.

References

[1] Nikita Araslanov and Stefan Roth. Single-stage semantic segmentation from image labels. In *CVPR*, pages 4253–4262, 2020. 6

[2] Sule Bai, Yong Liu, Yifei Han, Haoji Zhang, and Yansong Tang. Self-calibrated clip for training-free open-vocabulary segmentation. *arXiv preprint arXiv:2411.15869*, 2024. 2, 5, 6

[3] Walid Bousselham, Felix Petersen, Vittorio Ferrari, and Hilde Kuehne. Grounding everything: Emerging localization properties in vision-language transformers. In *CVPR*, pages 3828–3837, 2024. 6

[4] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. In *NeurIPS*, pages 468–479, 2019. 1

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. 3

[6] Niccolo Cavagnero, Gabriele Rosi, Claudia Cuttano, Francesca Pistilli, Marco Ciccone, Giuseppe Averta, and Fabio Cermelli. Pem: Prototype-based efficient maskformer for image segmentation. In *CVPR*, pages 15804–15813, 2024. 3

[7] Junbum Cha, Jonghwan Mun, and Byungseok Roh. Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In *CVPR*, pages 11165–11174, 2023. 6

[8] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, pages 1290–1299, 2022. 3

[9] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, pages 2818–2829, 2023. 2

[10] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *CVPR*, pages 4113–4123, 2024. 3

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, pages 3213–3223, 2016. 6, 7

[12] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *ICLR*, pages 1–11, 2024. 1, 2, 4

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 4, 6

[14] Zheng Ding, Jieke Wang, and Zhuowen Tu. Open-vocabulary panoptic segmentation with maskclip. *arXiv preprint arXiv:2208.08984*, 2022. 2, 3, 6

[15] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 111:98–136, 2015. 6, 7

[16] 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, 5, 6, 7, 8

[17] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916, 2021. 2

[18] Siyu Jiao, Yunchao Wei, Yaowei Wang, Yao Zhao, and Humphrey Shi. Learning mask-aware clip representations for zero-shot segmentation. In *NeurIPS*, pages 35631–35653, 2022. 3

[19] Dahyun Kang and Minsu Cho. In defense of lazy visual grounding for open-vocabulary semantic segmentation. In *ECCV*, pages 143–164, 2024. 3, 6, 8

[20] Laurynas Karazija, Iro Laina, Andrea Vedaldi, and Christian Rupprecht. Diffusion models for zero-shot open-vocabulary segmentation. In *ECCV*, pages 299–317, 2024. 3

[21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023. 3

[22] Philip A Knight. The sinkhorn–knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, 2008. 6

[23] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NeurIPS*, pages 109–117, 2011. 6

[24] 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, 5, 6, 7

[25] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Proxyclip: Proxy attention improves clip for open-vocabulary segmentation. In *ECCV*, pages 70–88, 2024. 3, 6, 7, 8

[26] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900, 2022. 2

[27] Yunheng Li, Zhongyu Li, Quansheng Zeng, Qibin Hou, and Ming-Ming Cheng. Cascade-clip: Cascaded vision-language embeddings alignment for zero-shot semantic segmentation. In *ICML*, pages 28243–28258, 2024. 2, 5

[28] Yi Li, Hualiang Wang, Yiqun Duan, Jiheng Zhang, and Xiaomeng Li. A closer look at the explainability of contrastive language-image pre-training. *Pattern Recognition*, 162:111409, 2025. 6

[29] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 6

[30] Yong Liu, Sule Bai, Guanbin Li, Yitong Wang, and Yansong Tang. Open-vocabulary segmentation with semantic-assisted calibration. In *CVPR*, pages 3491–3500, 2024. 1, 2

[31] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 3

[32] Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *ICML*, pages 23033–23044, 2023. 6

[33] Jiayun Luo, Siddhesh Khandelwal, Leonid Sigal, and Boyang Li. Emergent open-vocabulary semantic segmentation from off-the-shelf vision-language models. In *CVPR*, pages 4029–4040, 2024. 6

[34] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, pages 891–898, 2014. 6

[35] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 6

[36] Zelin Peng, Zhengqin Xu, Zhilin Zeng, Yu Huang, Yaoming Wang, and Wei Shen. Parameter-efficient fine-tuning in hyperspherical space for open-vocabulary semantic segmentation. In *CVPR*, pages 15009–15020, 2025. 3

[37] Zelin Peng, Zhengqin Xu, Zhilin Zeng, Changsong Wen, Yu Huang, Menglin Yang, Feilong Tang, and Wei Shen. Understanding fine-tuning clip for open-vocabulary semantic segmentation in hyperbolic space. In *CVPR*, pages 4562–4572, 2025. 3

[38] 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. 1, 2, 3, 4, 6, 7

[39] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, pages 25278–25294, 2022. 1, 2

[40] Tong Shao, Zhuotao Tian, Hang Zhao, and Jingyong Su. Explore the potential of clip for training-free open vocabulary semantic segmentation. In *ECCV*, pages 139–156, 2024. 1, 2, 3, 6

[41] Gyungin Shin, Weidi Xie, and Samuel Albanie. Reco: Retrieve and co-segment for zero-shot transfer. In *NeurIPS*, pages 33754–33767, 2022. 3

[42] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, pages 2256–2265, 2015. 3

[43] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *ICCV*, pages 7262–7272, 2021. 1

[44] Shuyang Sun, Runjia Li, Philip Torr, Xiuye Gu, and Siyang Li. Clip as rnn: Segment countless visual concepts without training endeavor. In *CVPR*, pages 13171–13182, 2024. 6

[45] Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. In *ECCV*, pages 315–332, 2024. 1, 2, 3, 5, 6, 7, 8

[46] Haoxiang Wang, Pavan Kumar Anasosalu Vasu, Fartash Faghri, Raviteja Vemulapalli, Mehrdad Farajtabar, Sachin Mehta, Mohammad Rastegari, Oncel Tuzel, and Hadi Pouransari. Sam-clip: Merging vision foundation models towards semantic and spatial understanding. In *CVPR*, pages 3635–3647, 2024. 6

[47] Monika Wysoczańska, Oriane Siméoni, Michaël Ramamonjisoa, Andrei Bursuc, Tomasz Trzciński, and Patrick Pérez. Clip-dinoiser: Teaching clip a few dino tricks. In *ECCV*, pages 320–337, 2024. 6, 7

[48] Bin Xie, Jiale Cao, Jin Xie, Fahad Shahbaz Khan, and Yanwei Pang. Sed: A simple encoder-decoder for open-vocabulary semantic segmentation. In *CVPR*, pages 3426–3436, 2024. 3

[49] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, pages 12077–12090, 2021. 3

[50] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *CVPR*, pages 18134–18144, 2022. 6

[51] Jilan Xu, Junlin Hou, Yuejie Zhang, Rui Feng, Yi Wang, Yu Qiao, and Weidi Xie. Learning open-vocabulary semantic segmentation models from natural language supervision. In *CVPR*, pages 2935–2944, 2023. 3

[52] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *CVPR*, pages 2945–2954, 2023. 3

[53] Xihong Yang, Xiaochang Hu, Sihang Zhou, Xinwang Liu, and En Zhu. Interpolation-based contrastive learning for few-label semi-supervised learning. *TNNLS*, 35(2):2054–2065, 2022. 7

[54] Xihong Yang, Jin Jiaqi, Siwei Wang, Ke Liang, Yue Liu, Yi Wen, Suyuan Liu, Sihang Zhou, Xinwang Liu, and En Zhu. Dealmvc: Dual contrastive calibration for multi-view clustering. In *ACMMM*, pages 337–346, 2023.

[55] Xihong Yang, Siwei Wang, Fangdi Wang, Jiaqi Jin, Suyuan Liu, Yue Liu, En Zhu, Xinwang Liu, and Yueming Jin. Automatically identify and rectify: Robust deep contrastive multi-view clustering in noisy scenarios. In *ICML*. PMLR, 2025. 7

[56] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. In *NeurIPS*, pages 32215–32234, 2023. 3

[57] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *CVPR*, pages 5122–5130, 2017. 6, 7

[58] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *ECCV*, pages 696–712, 2022. 1

[59] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *CVPR*, pages 11175–11185, 2023. 3

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.

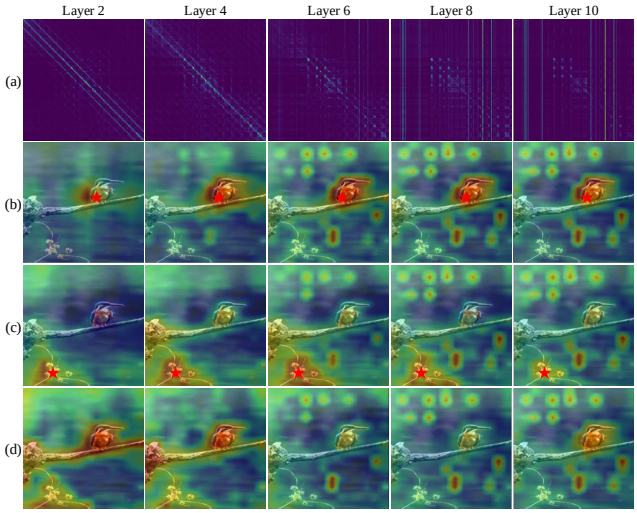

Figure 1. In-depth of CLIP's attention mechanism across various layers. (a) Attention maps of image tokens. (b) & (c) Attention maps of the selected image token, marked as ⋆, (d) Attention maps of the class token.

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.

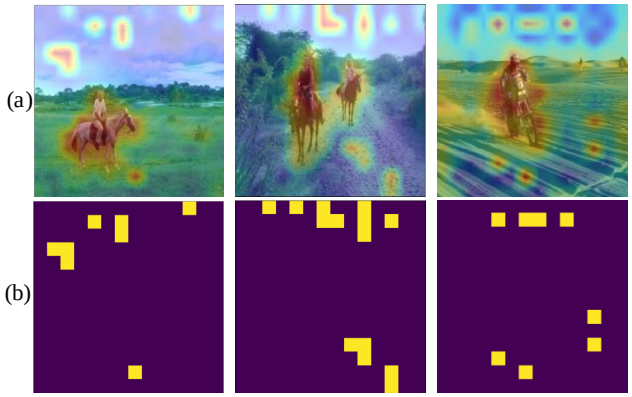

Figure 2. Visualization of outlier detection for different input images. (a) Image features with outliers. (b) Detected outliers by our SOM. It's seen that our SOM can adaptively identify different numbers of outliers for feature purification.

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