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

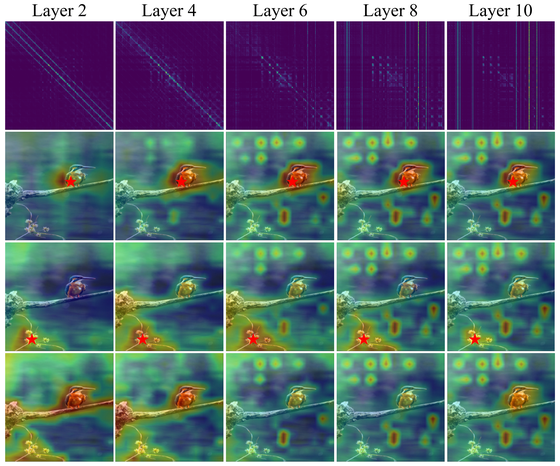
Shuo Jin, Siyue Yu*, Bingfeng Zhang, Mingjie Sun, Yi Dong, Jimin Xiao

XJTLU XI'AN JIAOTONG-LIVERPOOL UNIVERSITY · UNIVERSITY OF LIVERPOOL

ICCV OCT 19-23, 2025 HONOLULU HAWAII

Read paper

## Motivation & Introduction



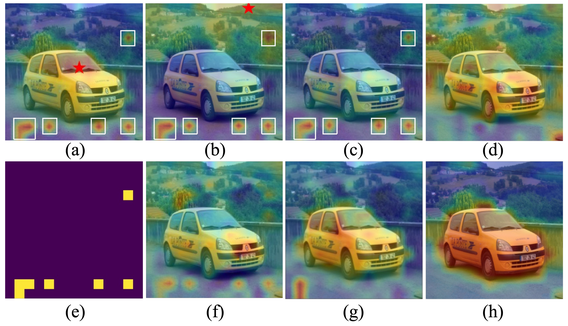

### Limitations of CLIP

- The in-depth analysis of CLIP's attention reveals the emergence of outliers and their distinct distribution.
- Outliers tend to appear in deep layers.
- 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.

### Research Goal

- To resolve outliers adaptively and enhance semantic representations for CLIP-based training-free open-vocabulary semantic segmentation (OVSS).

## Main Contributions

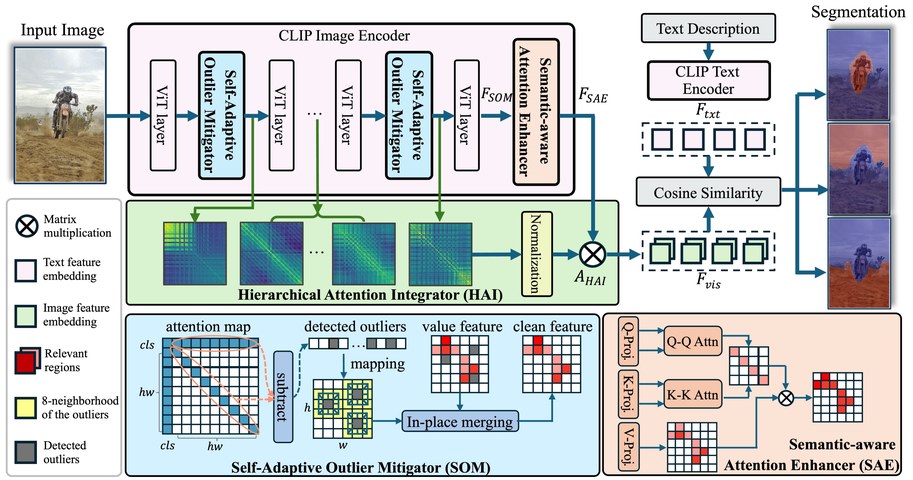



Visualization of our **S**elf-adaptive **O**utlier **M**itigator (SOM)

### Self-adaptive Outlier Mitigator

- In the attention map, simply compare the difference between the self-response values of the image tokens (i.e., the diagonal weights) and their attention values with the class token.

## Methods




### Self-adaptive Outlier Mitigator

- To detect and eliminate the outliers automatically. Based on the comparison between self-response and class-response within the CLIP's self-attention map.

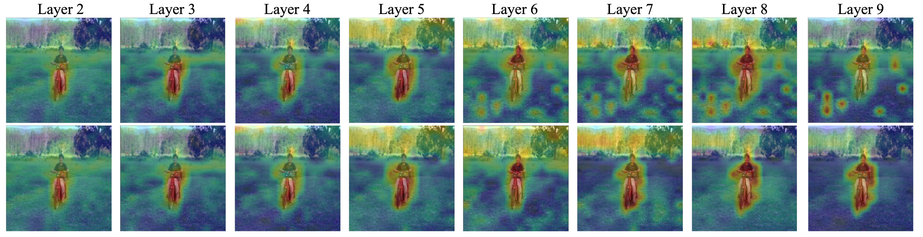
### Semantic-aware Attention Enhancer

- To strengthen the object activation in CLIP's last image encoder layer using the self-self attention mechanism, which is not considered in outlier mitigator.

### Hierarchical Attention Integrator

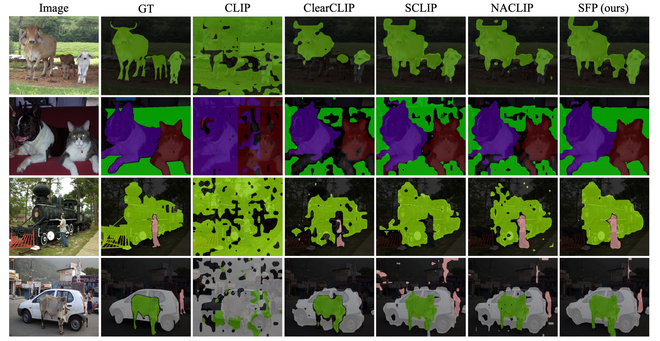
- To further refine the generated feature. Leverage attention maps from shallow layers to capture discriminative pair-wise feature relationships.

## Outlier Mitigation Result




## Experiments

### Qualitative Evaluation




### Quantitative Evaluation

*No extra training & backbone !*

| Methods | Pub. & Year | Extra backbone | Training free | with background V21 | with background PC60 | with background Object | without background V20 | without background PC59 | without background Stuff | without background ADE | without background City | without background Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 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** |

### Ablation Studies

**Ablation of proposed modules**

| | 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** |

**Ablation of different backbones**

| 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] | 57.9 | 16.7 | 37.3 |
| | 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] | 60.6 | **22.6** | 41.6 |
| | NACLIP [16] | 52.2 | 17.3 | 34.8 |
| | SFP | **64.7** | 22.1 | **43.4** |

**Ablation of aggregated attention layers**

| $A_{l=l_0}^{l=L-1}$ | 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 |

**Ablation of different self-self attention**

| 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 |