ResCLIP: Residual Attention for Training-free Dense Vision-language Inference

Supplementary Material

In this supplementary document, we present additional materials not included in the main manuscript due to page limitations. The supplementary content is outlined:

- Sec. A: Additional attention comparison between the proposed method and previous works.
- Sec. B: Ablation studies on different ViT backbones.
- Sec. C: Extension on other CLIP-like models.
- Sec. D: Inference efficiency analysis of designed models.
- Sec. E: More segmentation visualization results.

Now, we will present these materials as follows.

A. Attention Comparison

To further illustrate the impact of ResCLIP on attention mechanisms beyond examples shown in Fig. 3 in main paper, we present additional attention visualizations in Fig. A1. These visualizations demonstrate how our method enhances the attention maps across different training-free open-vocabulary semantic segmentation (OVSS) models so that our method could better aggregate information from previous layers. From Fig. A1, we can observe that our ResCLIP could attend to regions sharing similar class-specific features while previous works usually exhibit spatial-invariant features or focus on the local patches.

In particular, after integrating our Residual Cross-correlation Self-attention (RCS) and Semantic Feedback Refinement (SFR) modules into existing works, the attention maps show two key improvements: 1) enhanced local patches awareness and 2) strengthened global semantic correspondence. For example, in the left part of Fig. A1, we observe that previous works fail to effectively capture features from other "sheep" instances while our method can not only capture information from semantically consistent objects but also maintain local consistency. Similar phenomena can be observed from the right example in Fig. A1. Moreover, we can see that intermediate layers (*e.g.*, layers 5 and 11) show decent class-specific feature correspondence ability, which motivates us to incorporate them to remold the attention in the last block of CLIP.

B. Ablation Studies on ViT Backbones

In the main manuscript, we demonstrate effectiveness of the proposed RCS and SFR modules on ViT-B/16 backbone. To further demonstrate their generalization on other ViT backbones, we conduct additional experiments of ablation studies across ViT-B/16, ViT-B/32, and ViT-L/14 backbones. Moreover, we also evaluate our ResCLIP method by integrating it with previous training-free counterparts, *i.e.*, SCLIP [45], ClearCLIP [27], and NACLIP [18].

The experimental results are shown in Table A1. We can see that both RCS and SFR modules contribute substantially to performance improvements across multiple backbones and baselines, demonstrating the great generalization of our proposed modules. Specifically, taking NACLIP with ViT-B/16 as an example, Our RCS improves the average mIoU from 39.4% to 40.6%, while SFR increases it to 40.7%. When combining both modules, the performance further improves to 41.4%, suggesting complementary benefits from both components. Similar patterns are observed with other baseline methods.

Notably, our method demonstrates robust performance across different backbone architectures. For instance, when applied to SCLIP with ViT-L/14, ResCLIP significantly improves the average performance from 26.2% to 37.0%, showing particular effectiveness on larger architectures. The improvement is consistent across datasets both with and without a background class. Specifically, ViT-B/16 achieves 43.2% mIoU on datasets with a background class, showing a 1.8% mIoU improvement over NACLIP baseline, and 40.3% mIoU on datasets without a background class, with a 2.1% mIoU improvement. These comprehensive results validate that our proposed modules effectively enhance dense prediction capability of CLIP across various architectures and dataset configurations, demonstrating the robustness and generalization ability of our approach.

C. Extension on other CLIP-like Models

In the main paper, we evaluate our method by integrating it with existing approaches, which are typically improved versions based on the vanilla CLIP model. To further evaluate the effectiveness of our method on other CLIP-like models, we conduct additional experiments on the OpenCLIP [10]. For a fair comparison, we first reproduce the results of SCLIP [45], ClearCLIP [27], and NACLIP [18] on OpenCLIP [10]. Then, we implement the proposed method based on the OpenCLIP [10]. As shown in Table A2, we present the comprehensive results on datasets without a background class. We can observe that our method shows consistent improvements over different baseline approaches, demonstrating its effectiveness.

Specifically, when integrating SCLIP [45] with our method, ResCLIP achieves significant gains across all datasets, improving the average performance by 1.6% mIoU. The improvement is particularly pronounced on VOC20, where ResCLIP enhances the mIoU from 66.6% to 71.8%. Most notably, integrating ResCLIP with NACLIP [18] yields substantial improvements across all

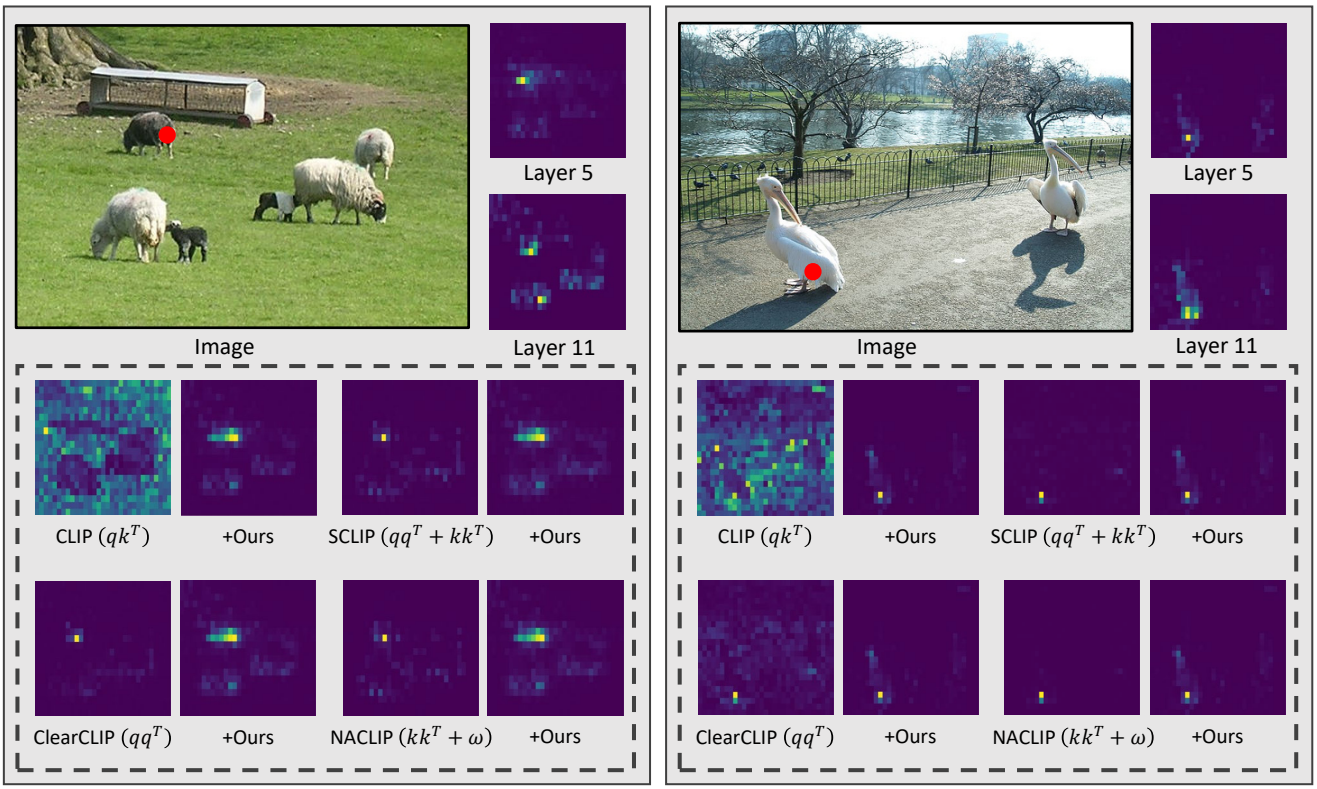

Figure A1. Additional comparison of attention maps across CLIP [37], SCLP [45], ClearCLIP [27], NACLIP [18], and ours. The attention maps of non-last layers show the localization properties and can heal the attention in the last layer. The red point serves as the source point from which the attention map is computed and visualized.

Table A1. Ablation studies of our proposed modules in ViT-B/16, ViT-B/32 and ViT-L/14 backbones. Our ResCLIP setting is marked in gray. The best result on each dataset is **bolded**. The $\text{Avg}_{w/o}$ means the average mIoU for datasets *without* a background class, $\text{Avg}_{w/}$ means the average mIoU for datasets *with* a background class, and Avg. means the average mIoU for all eight datasets.

Methods	Module		ViT-B/16			ViT-B/32			ViT-L/14		
	RCS	*SFR*	$\text{Avg}_{w/o}$	$\text{Avg}_{w/}$	Avg.	$\text{Avg}_{w/o}$	$\text{Avg}_{w/}$	Avg.	$\text{Avg}_{w/o}$	$\text{Avg}_{w/}$	Avg.
SCLIP [45]	-	-	37.1	40.0	38.2	32.1	36.2	33.6	23.6	30.5	26.2
+ResCLIP(Ours)	✓		38.8	42.4	40.2	34.6	36.9	35.4	36.6	36.9	36.7
		✓	37.9	42.0	39.4	32.2	36.4	33.8	28.9	30.5	29.5
	✓	✓	**39.3**	**42.7**	**40.5**	**34.8**	**37.1**	**35.7**	**36.7**	**37.4**	**37.0**
ClearCLIP [27]	-	-	37.5	39.1	38.1	34.8	35.6	35.1	34.5	35.5	34.9
+ResCLIP(Ours)	✓		39.7	41.6	40.4	35.3	35.7	35.4	38.3	36.7	37.7
		✓	39.4	41.7	40.2	35.1	35.8	35.3	37.0	36.2	36.7
	✓	✓	**40.0**	**42.0**	**40.7**	**35.5**	**35.9**	**35.6**	**38.4**	**37.2**	**37.9**
NACLIP [18]	-	-	38.2	41.4	39.4	34.4	37.0	35.4	36.2	36.9	36.5
+ResCLIP(Ours)	✓		39.7	42.2	40.6	35.7	37.3	36.3	38.4	38.2	38.3
		✓	39.3	42.9	40.7	35.7	37.4	36.3	37.4	38.4	37.8
	✓	✓	**40.3**	**43.2**	**41.4**	**36.2**	**37.5**	**36.7**	**39.1**	**39.2**	**39.1**

Table A2. Quantitative comparison on datasets *without* a background class based on OpenCLIP [10] with ViT-B/16 architecture. Our results are marked in gray. The best results on each dataset are **bolded**. Results show that our method is also effective on other VLMs.

Methods	VOC20	Context59	Stuff	Cityscape	ADE20k	Avg.
OpenCLIP [10]	47.2	9.0	5.0	5.1	2.9	13.84
SCLIP [45]	66.6	31.7	21.2	31.4	18.5	33.9
+ResCLIP(ours)	**71.8**	**32.9**	**21.9**	**31.9**	**18.8**	**35.5** (+1.6)
ClearCLIP [27]	81.4	34.1	23.1	31.8	18.9	37.9
+ResCLIP(ours)	**83.3**	**34.3**	**23.1**	**32.3**	**19.1**	**38.4** (+0.5)
NACLIP [18]	76.2	30.3	20.3	32.3	17.6	35.3
+ResCLIP(ours)	**82.5**	**33.0**	**22.2**	**32.9**	**19.0**	**37.9** (+2.6)

datasets, with an impressive average gain of 2.6% mIoU, including a remarkable 6.3% improvement on VOC20 datasets from 76.2% to 82.5%. These consistent improvements across different CLIP models and datasets demonstrate the generalization of our approach. The results also validate that the observation of our proposed method is effective on other CLIP-like models.

D. Inference Efficiency Analysis

The additional inference time introduced by our method is limited, as all operations only adjust attention in the final layer. Specifically, RCS computes the average of intermediate-layer attentions already generated during inference, while SFR performs lightweight mask adjustments in the final layer of CLIP. Using a single RTX 3090 GPU with batch size 1, input resolution of 336×336, and fp16 half precision, our experimental evaluation shows negligible impact on inference speed across all models. As shown in Table A3, our enhancements increase total FLOPs by less than 7%. Moreover, RCS demonstrates negligible overhead, while implementation of SFR can be further optimized to improve efficiency.

Table A3. The Speed and FLOPs comparison of different methods on VOC 20 using CLIP-ViT-B/16 backbone. All the experiments are conducted on a single RTX 3090 GPU. IPS: Image Per Second.

Metrics	CLIP	NACLIP	+*RCS*	+*SFR*	+ResCLIP
Speed (IPS) ↑	32.8	32.3	30.5	29.3	28.9
FLOPs (G) ↓	41.7	41.8	42.0	44.6	44.8

E. Additional Visualization Results

We present additional qualitative comparisons across ADE20K [58], COCO Object [7], and PASCAL VOC [16] datasets in Fig. A2, Fig. A3, and Fig. A4 to further demonstrate the effectiveness of our ResCLIP, respectively. Compared to existing methods, our approach usually presents better quality in terms of the semantic segmentation masks. From these qualitative results, we can have the following observations: 1) Our method generates significantly cleaner segmentation masks with reduced noise artifacts. This improvement is particularly evident in complex scenes from ADE20K, where ResCLIP maintains coherent building segmentation without the internal hollows or fragmentations commonly seen in other baselines (*i.e.*, the 1-st *col.* in Fig. A2). The enhanced segmentation quality extends to diverse scenarios, such as the precise delineation of vehicles in parking lots and the clear separation of multiple instances in crowded scenes (*i.e.*, the 2-nd and 4-th *col.* in Fig. A2). 2) ResCLIP presents superior performance in handling multiple object instances, demonstrating its enhanced spatial-semantic understanding. For example, in the COCO Object dataset (see Fig. A3), our method accurately segments groups of animals while maintaining clear boundaries between individuals(*i.e.*, the 4-th and 5-th *col.* in Fig. A3). This capability stems from the improved attention mechanism of our ResCLIP, which better captures both global spatial relationships and local feature consistency. 3) Our method handles varying scales and perspectives better. As shown in Fig. A4, our method produces consistent segmentation quality across both indoor and outdoor scenes. These qualitative results validate the effectiveness of our proposed RCS and SFR modules in enhancing dense prediction capabilities of CLIP.

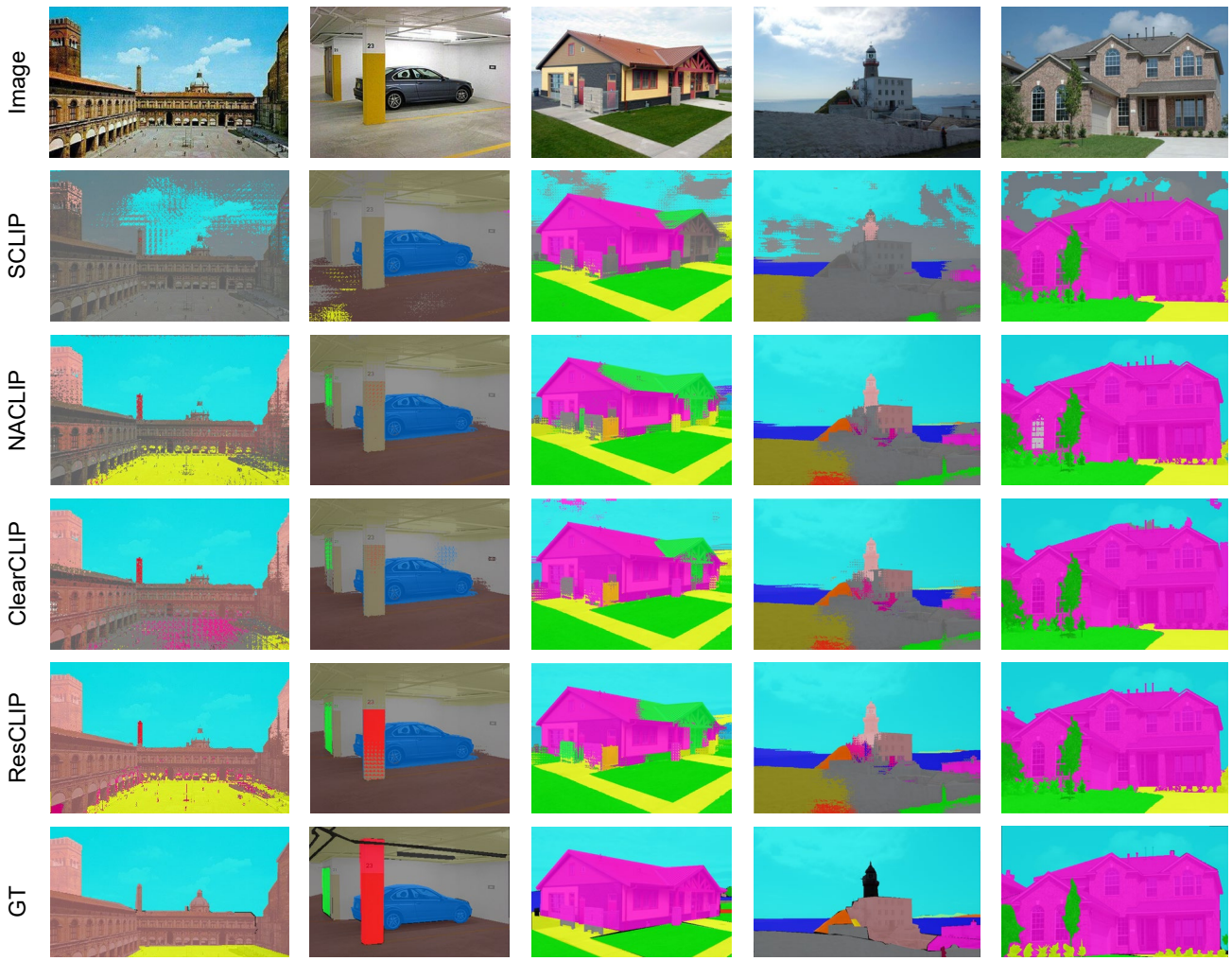

Figure A2. Additional qualitative visualization results among different CLIP-based training-free segmentation methods on ADE20K [58] dataset.

Figure A3. Additional qualitative visualization results among different CLIP-based training-free segmentation methods on COCO Object [7] dataset.

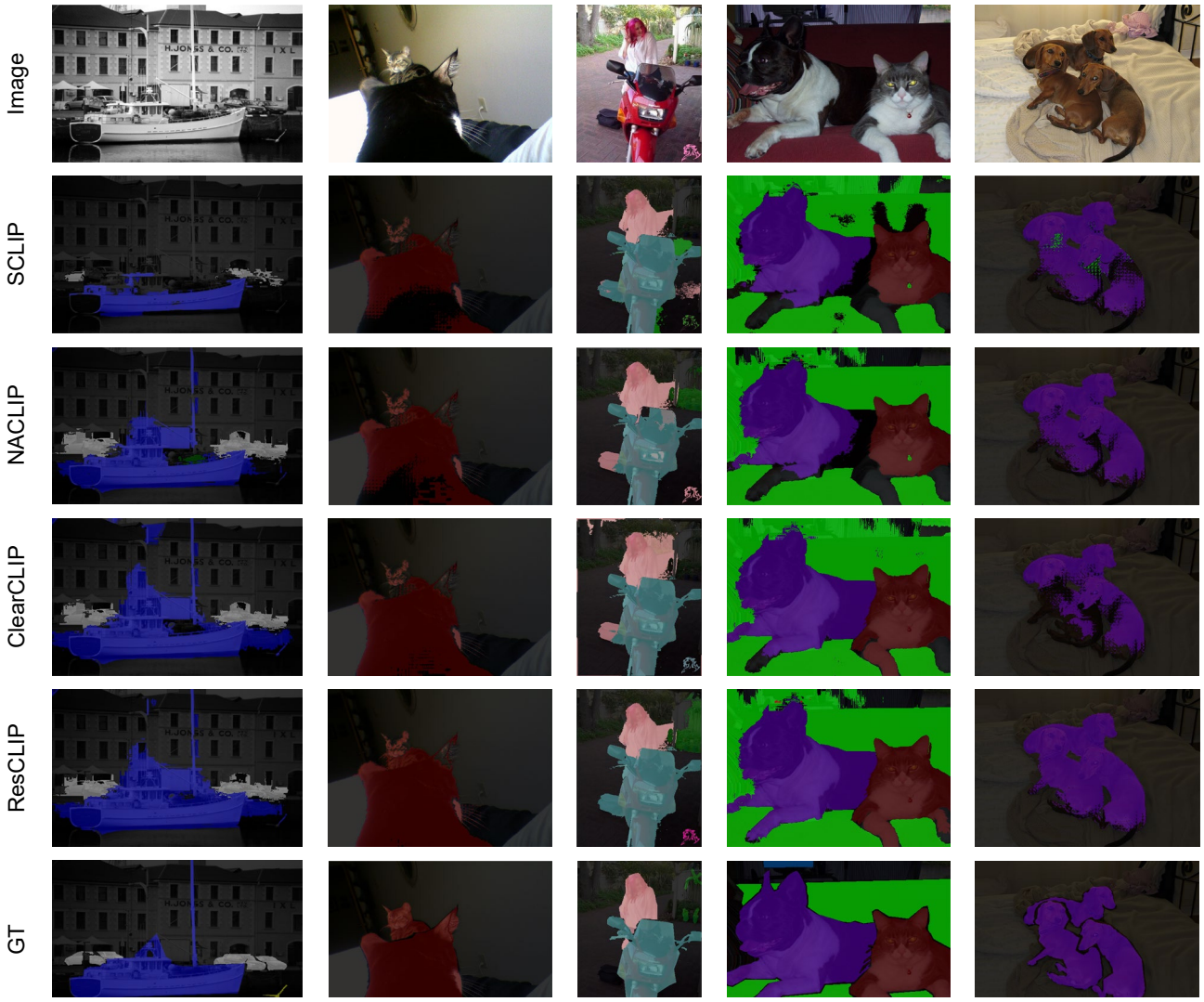

Figure A4. Additional qualitative visualization results among different CLIP-based training-free segmentation methods on PASCAL VOC [16] dataset.