Appendix for *Harnessing Vision Foundation Models for High-Performance, Training-Free Open Vocabulary Segmentation*

Supplementary Material

Setting	VOC20	Context59	Context60	Stuff	ADE
ProxyCLIP					
336-336-112	79.7	34.4	38.1	25.7	19.4
448-336-224	78.5	34.8	38.3	25.7	19.3
576-336-224	73.4	33.8	37.0	24.6	19.0
Trident w.o SAM Refine					
336-336-112	83.7	35.8	39.6	27.0	20.7
448-336-224	83.3	37.0	40.7	27.6	20.6
576-336-224	82.0	37.2	40.9	27.5	20.9

Table 1. **Ablation for Different Input Resolution.** SAM refinement is excluded here in Trident to clearly show the effect of the proposed framework. The setting is an abbreviation of **shorter side - window size - stride**.

1. More Ablations

On the Effect of Input Resolution. To verify the claim that Segment-then-Splice paradigm might diminish with increased resolution, we conducted ablation studies on more datasets for ProxyCLIP [4], and our Trident. The detailed results are presented in Table 1. ProxyCLIP exhibits deteriorating performance as resolution increases; specifically, its performance on the VOC20 [1] dataset decreases from 79.7% to 73.4% mIoU when the input resolution is increased from 336 to 576 for the shorter side of the image. This trend is consistent across other datasets, as the receptive field becomes more constrained in the Segment-then-Splice paradigm with increased resolution. In contrast, our Splice-then-Segment paradigm benefits from increased resolution, thereby obtaining better performance with increased resolution. A notable exception is the VOC20 dataset, where the performance of our method also declines due to baseline's significantly reduced effectiveness.

On SAM's Feature in Correlation Matrix. We provide a comprehensive ablation in Tab. 2 comparing different SAM [3] image encoder features (q, k, v, x) in the last transformer layer on ProxyCLIP without refinement. Results show our proposed combination of cosine similarity and attention weights (Eq.5 in main paper) performs best, surpassing individual feature's cosine similarity.

2. More details on SAM Refinement.

For SAM refinement, we only generate prompts for classes with activation scores above a preset threshold (*e.g.* 0.2). Additionally, we filter out extremely small regions (*e.g.* less than 900 pixels) before prompting SAM. These tricks significantly reduce the number of prompts—for COCO Object dataset, we only use 4.21 box/point prompts per image on average. For intuitive understanding, we present the generated prompts (Fig. 1b) for the case used in Fig.3 of main paper.

Additionally, we show why using only mask or box/point prompts degrades performance in a special case (Fig. 1a). Sub-optimal activations cause mask prompts to simultaneously cover foreground and background regions, creating ambiguity for SAM's decoder. Box/point prompts only highlight partial objects due to their sparse encoding. For a qualitative comparison of the SAM refinement, we visualize our segmentation results against the ProxyCLIP baseline in Fig. 1c.

Type	V21	C60	Obj.	V20	C59	Stf.	City	ADE	Avg.
q-q	49.8	28.8	30.3	76.4	32.7	21.9	31.9	15.9	35.9
k-k	24.4	13.8	11.6	62.7	16.8	10.4	2.7	5.3	18.5
v-v	60.6	35.0	36.7	80.6	38.2	26.2	36.7	19.5	41.7
x-x	63.4	36.8	38.1	82.3	40.2	27.1	38.8	20.2	43.4
Aff.	64.5	37.2	39.5	83.7	40.9	27.6	40.4	20.9	44.3

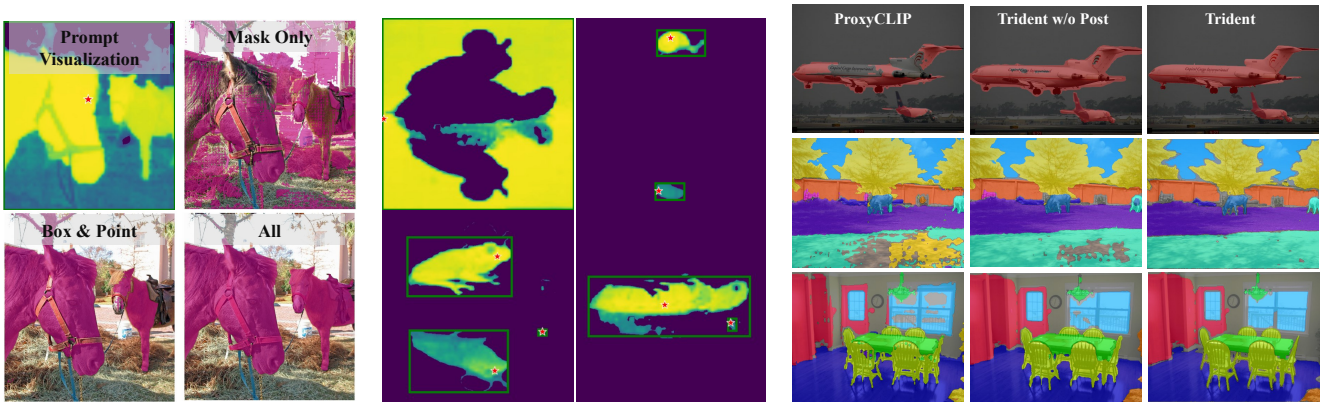
Table 2. Ablation of SAM's feature in correlation matrix.

DINO	SAM	Ref.	#Params. (M)	Mem. (MB)	Thru. (imgs/sec)
CLIP-B/16 + SAM-B/16					
			149	672	118.8
✓			234	851	68.5
✓	✓		323	2501	15.3
✓	✓	✓	364	2526	10.0
OpenCLIP-H/14 + SAM-H/16					
			986	2308	28.2
✓			1071	2484	22.4
✓	✓		1708	5804	6.4
✓	✓	✓	1749	5827	5.0

Table 3. **Efficiency Analysis for Trident Framework.** The inference latency is tested on RTX 4090 GPU with FP16 precision.

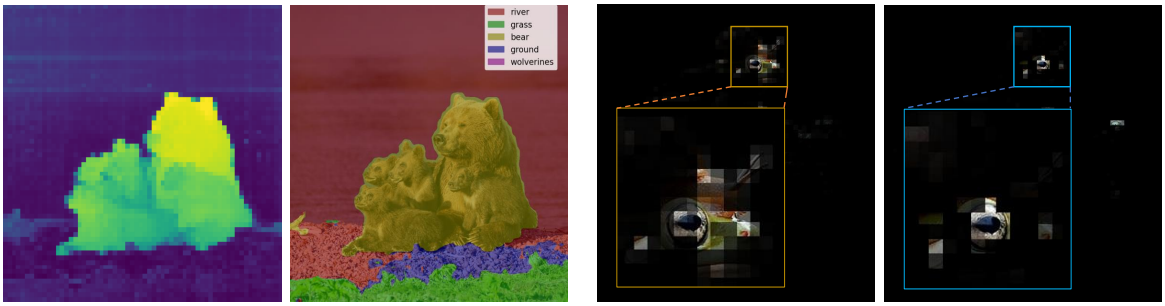
3. Efficiency Analysis

As the proposed Trident integrates three foundational models, we conducted an analysis of their impact on inference costs. We adopted an image resolution of 448 $\times$ 448 and



(a) (b) (c)

Figure 1. **Additional analysis of our SAM refinement module.** (a). Ablation study showing that a combination of box/point and mask prompts is superior to using either alone. (b). Illustration of the automatically generated prompts for the example in Fig. 3 of main paper. (c). Qualitative comparison against the ProxyCLIP baseline under with or without SAM refinement setting.



(a) (b)

Figure 2. (a). Activation map and segmentation result of Trident model for the case presented in Fig.2 of the main paper. (b). Qualitative comparison between SAM's attention(left) vs. our affinity(right).

utilized a sliding window with a size of 336 and stride of 224 for CLIP and DINO. For SAM, the input resolution was set to 1024. The throughput was tested on an RTX 4090 GPU using FP16 precision for all models. The detailed results are reported in Tab. 3, which includes results for both the base and huge versions of these models. The introduction of SAM resulted in a significant increase in GPU memory usage and processing time, primarily due to its demand for high-resolution inputs. Additionally, the incorporation of SAM refinement led to a slight increase in both GPU memory usage and time cost.

4. Additional Comparison with LVLMs

Recent advancements in Large Vision-Language Models (LVLMs) that combine Large Language Models (LLMs) with vision foundation models (e.g., SAM [3]) have demonstrated promising capabilities for open-vocabulary segmentation tasks. To provide a more comprehensive analysis of our proposed model, we conducted additional comparisons with recent LVLMs [5, 9] that are capable of open-vocabulary segmentation. The detailed results are presented in Table 4. Although all compared models integrate SAM for segmentation, our Trident framework establishes significantly better performance. Compared to Text4Seg [5], our Trident achieves an average mIoU improvement of 4.7% while being approximately 35 times faster in throughput.

Method	**VOC20**	**Context59**	**ADE150**	**Avg.**	**FPS**
LaSagnA[Arxiv24'04]	69.8	46.1	14.3	43.4	-
Text4Seg[ICLR'25]	76.5	52.5	16.5	48.5	0.14
Trident	88.7	44.4	26.7	53.2	5.0

Table 4. Comparison with recent Large Vision-Language Models (LVLMs). We report the mIoU (%) and inference latency to evaluate the performance of different models. FPS measurements were conducted on an NVIDIA RTX 4090 GPU.

5. Qualitative Comparison

In Fig. 2a, we present qualitative results of our method for the case depicted in Fig. 2 of the main paper. Compared to the traditional splice-then-segment paradigm, our approach eliminates the window panel effect and yields more fine-grained segmentation results. Fig. 2b compares SAM's at-

tention(left) vs. our affinity(right) for the given case in Fig. 3 of the main paper. The affinity matrix only preserve sematic-consistent regions in attention matrix (e.g. attention weights include the snail region in left image while affinity weights not), domentrating better semantic consistancy in feature aggragation procedure.

Fig. 3 presents qualitative comparisons between Trident and previous SOTA methods, all utilizing the CLIP-B/16 architecture. The visual results highlight Trident's superior performance in two aspects: improved semantic consistency in object recognition and more precise segmentation boundary delineation. These advantages are particularly pronounced in complex scenes from Context60 and Cityscapes benchmarks. In Fig. 4, we present additional qualitative comparisons with previous SOTA methods on the VOC21, Context60, COCO Object, and Cityscapes benchmarks, all under the same ViT-B/16 setting. Our Trident framework demonstrates improved semantic consistency, although some masks yield incorrect classifications.

6. Details of Application to LVLMs

Recent advancements in API [11] development demonstrate that vision-language models, such as CLIP, can serve as auxiliary models to enhance the performance of Large Vision Language Models (LVLMs). Specifically, these models generate text-guided attention heatmaps by computing the similarity between textual descriptions and dense image feature maps, subsequently modulating the pixel values of the original image. To derive the similarity map between text and the dense image feature map, sophisticated decomposition mechanisms are employed. Initially, the similarity is calculated between the text embeddings and the classification token within CLIP's image encoder. This similarity is then propagated to other visual tokens using the attention scores between the classification token and the visual tokens. We observe that the image feature map can be effectively substituted with the dense feature map extracted by our Trident. This substitution results in enhanced performance.

Following API, We employ the LLaVA-1.5-13B [8] across five multi-modality datasets: MMMU [13], MM-Vet [12], MME [2], LLaVA-Wild [7], and POPE [6] to substantiate our claims. For MM-Vet and LLaVA-Wild, we utilize a GPT-based evaluation tool to score the responses. For MMMU and MME, we report accuracy based on the matching accuracy between LVLM's responses and the ground truth. For POPE, we present the F1-Score using its random split setting. Due to the unavailability of official API implementations for MMMU, MME, and POPE, we provide results based on our re-implementation.

7. Limitations

Although Trident achieves superior performance, it requires three models during inference, which compromises efficiency compared to previous methods that utilize only CLIP. We believe unsupervised learning methods, such as CLIP-DINOiser [10], could mitigate this efficiency issue while retaining the strengths of these three models.

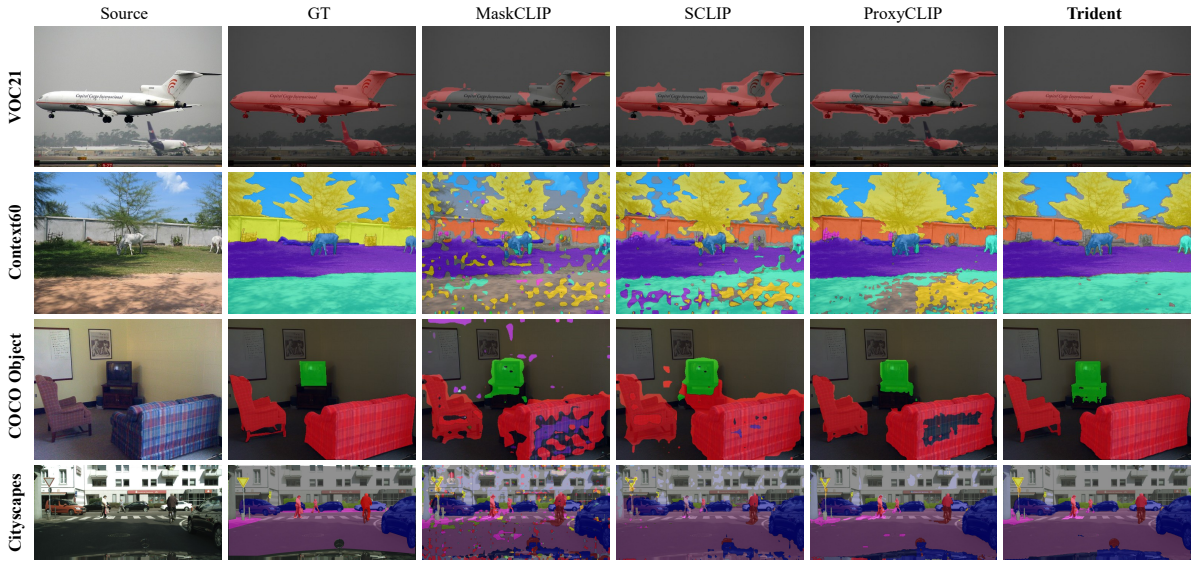


Figure 3. **Qualitative comparison with previous training-free open vocabulary segmentation methods.**

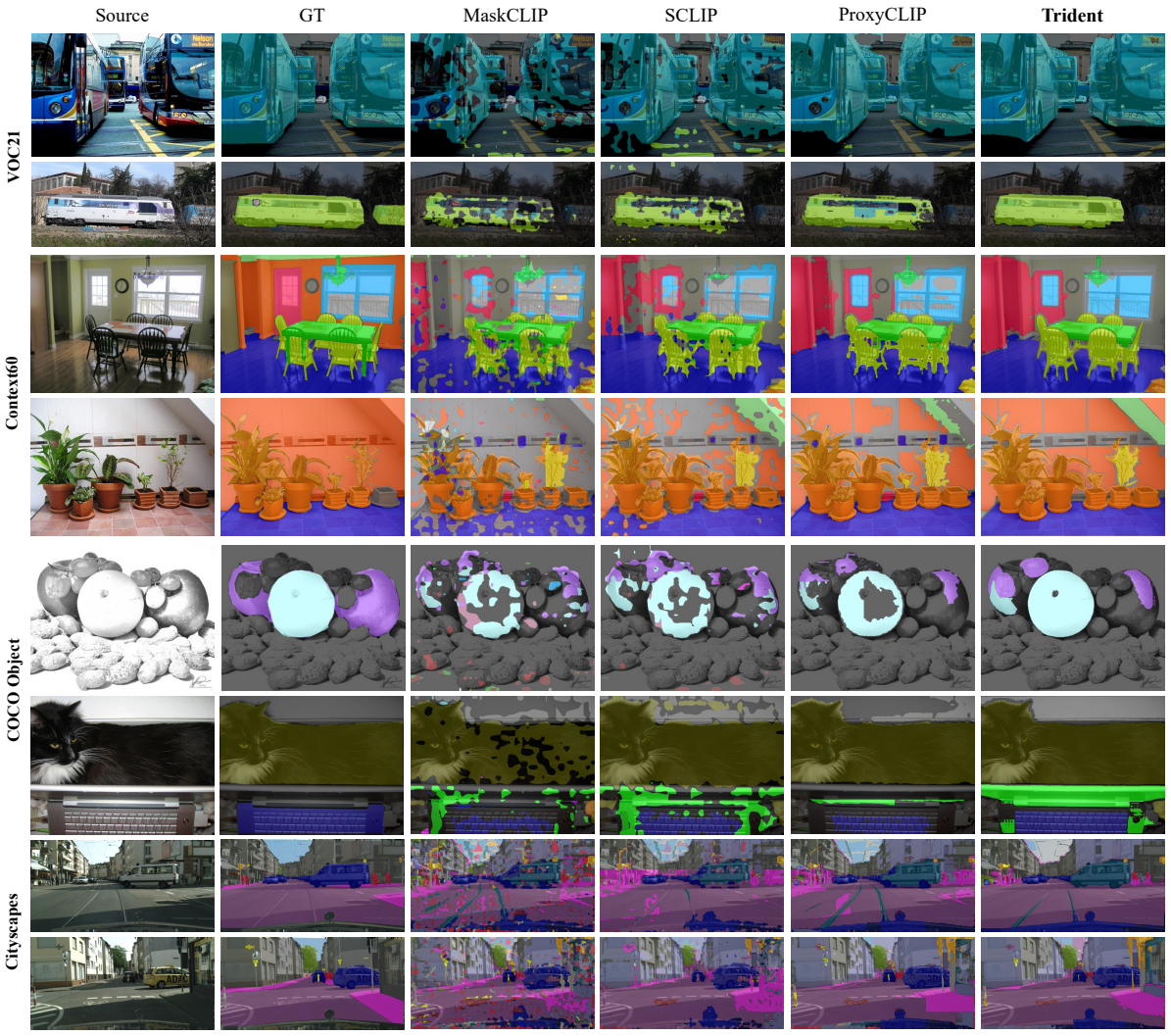


Figure 4. **Additional visualization comparison with previous SOTA training-free methods.**

References

[1] 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*, 2015. 1

[2] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 3

[3] 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*, 2023. 1, 2

[4] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Proxyclip: Proxy attention improves clip for open-vocabulary segmentation. In *ECCV*, 2024. 1

[5] M. Lan et al. Text4seg: Reimagining image segmentation as text generation. In *ICLR*, 2025. 2

[6] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 3

[7] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 3

[8] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2023. 3

[9] Cong Wei, Haoxian Tan, Yujie Zhong, Yujiu Yang, and Lin Ma. Lasagna: Language-based segmentation assistant for complex queries. *arXiv preprint arXiv:2404.08506*, 2024. 2

[10] Monika Wysoczańska, Oriane Siméoni, Michaël Ramamonjisoa, Andrei Bursuc, Tomasz Trzciński, and Patrick Pérez. Clip-dinoiser: Teaching clip a few dino tricks. *arXiv preprint arXiv:2312.12359*, 2023. 3

[11] Runpeng Yu, Weihao Yu, and Xinchao Wang. Api: Attention prompting on image for large vision-language models. In *ECCV*, 2024. 3

[12] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *ICML*, 2024. 3

[13] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024. 3