Supplementary material for Plug-in Feedback Self-adaptive Attention in CLIP for Training-free Open-Vocabulary Segmentation

A. Summary

In this supplementary material, we present the following additional content to complement the main paper:

- Additional qualitative comparisons on various datasets.
- We present more details motivation and additional observations.
- Sensitivity on similarity metric.
- Impact of different configuration of CLIP.
- Additional speed analysis.

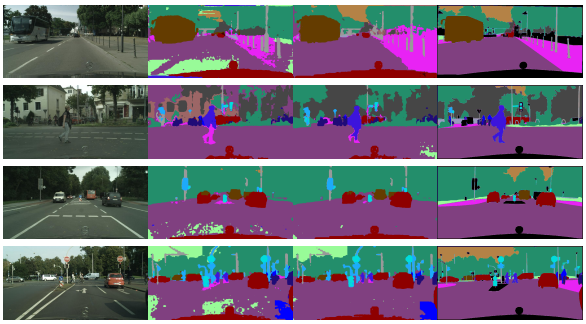
B. Additional qualitative results

In Fig. S1, we provide additional qualitative comparisons with ProxyCLIP on the Cityscapes dataset. By incorporating our self-adaptive framework, we successfully correct missegmented regions. Notably, for the same object, certain regions are initially misclassified; however, our feedback-adaptive method aggregates information from similar patches in the output, enabling further refinement and correction. In Fig. S2, we show expanded comparison with MaskCLIP and SCLIP with the exmaples in Row 5-6 of Fig. 6.

In Fig. S5, we present additional results from the VOC21 dataset, along with attention maps corresponding to the reference patch (indicated by the red box in the first column). The segmentation results of ProxyCLIP (third column) exhibit flaws, as certain regions within the main object are incorrectly segmented. This issue arises because those patches fail to attend correctly to the same object, as illustrated in their attention maps (second column). In contrast, our feedback self-adaptive method successfully corrects the segmentation (fifth column) across the entire object by attending to more regions belonging to the same object.

C. Additional motivation and observation

Our proposed FSA aims to improve the spatial coherence among similar patches using the feedback loop. The feedback loop is derived using self-predicted logits for each patch. The concept is similar to knowledge distillation [2, 12], where the output logits of a stronger model is used as a soft label to guide the current model to learn extended knowledge, instead of the sparse labels from ground truth.



(a) Input (b) ProxyCLIP (c) +FSA (d) GT

Figure S1. **Qualitative results on Cityscapes.** By integrating our feedback self-adaptive mechanism, we correct missegmented patches by ProxyCLIP, ensuring consistent segmentation within each object. We can clearly observe that our segmentation is more consistent across whole objects.

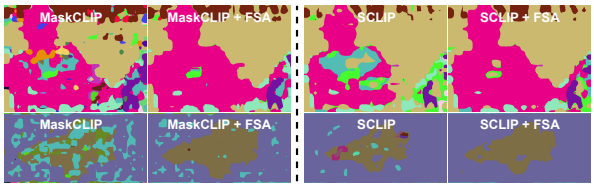


Figure S2. **Examples of Row 5-6 in Fig.6 for MaskCLIP and SCLIP.**

More specifically, it is close to self-distillation [10, 11] where both the teacher and student are the model itself. On the other hand, our methodology is also closely related to test-time adaptation, which normally adapt the model towards one specific test data instance [7, 8] or specific domain [1, 9]. The process is normally self-supervised without any additional manual labeling [5, 6].

In the main paper, we have illustrated the semantic coherence retention. To quantify subsequent degradation, we introduce a new metric: using $Attn^{init}$ as reference, for each patch i, we get its most attended patch j. After each operation in Eq.2 (residuals, FFNs), we compute pairwise token similarities and check whether j remains among the top-10 similar patches to i. Fig. S3 illustrates this process and the metric drops (ave of 8 datasets) sharply after residuals in MaskCLIP, indicating noise injection [21]. In con-

Similarity metric	ViT-B/16	ViT-L/14	ViT-H/14
Cosine	43.2	43.3	45.4
KL divergence	43.3	43.6	45.8

Table S1. **Sensitivity on similarity metric.** KL divergence evaluates entire distributions and emphasizes differences in probabilistic outputs, making it ideal for capturing detailed semantic coherence and supporting effective feedback adaptation.

trast, our FSA better preserves spatial coherence. Fig. S4 compares attention maps (Fig.2) after the proj: although both methods reduce focus on the cat's face, our improved intermediate attention provides greater *resistance to degradation.*

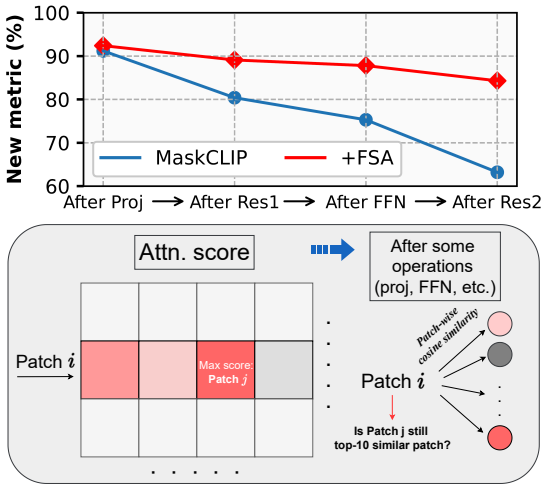


Figure S3. Illustration and analysis for new metric.

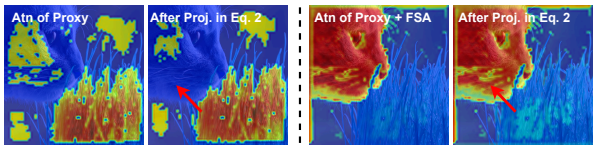


Figure S4. Attention visualization of Fig.2 after Proj. in Eq.2.

D. Similarity metric

Table S1 compares cosine similarity and KL divergence for computing logit similarity. KL divergence proves more effective due to its ability to assess full distributions and highlight differences in probabilistic outputs, making it better suited for capturing detailed semantic coherence and enabling effective feedback adaptation.

E. Impact of different configuration of CLIP

ProxyCLIP and ClearCLIP omit residual and FFN modules, identified as sources of noisy segmentation [3, 4], thus better preserving spatial consistency than MaskCLIP or SCLIP, which retain them. As our method primarily enhances semantic consistency, it yields larger improvements on baselines with weaker spatial coherence. As shown in Tab. S2, FSA improves MaskCLIP under different configurations of CLIP, though the margin is smaller in the latter.

Methods	ViT-B/16	ViT-L/14	ViT-H/14
MaskCLIP	27.9	13.9	19.3
+FSA	35.8 (+7.9)	32.6 (+18.7)	33.4 (+14.1)
MaskCLIP (w/o FFN, Res)	29.5	29.7	29.8
+FSA	36.8 (+7.3)	34.1 (+4.4)	34.7 (+4.9)

Table S2. **Improvement over MaskCLIP with different configurations.** Average mIoU reported on 8 datasets.

F. Additional speed analysis

Following Clear/Mask/SCLIP, we modify only the *last layer*, incurring a 4.3–11.7% overhead depending on layer count (Tab. S3).

Methods	B/16(12-layers)	L/14(24-layers)	H/14(32-layers)
Clear/+FSA	4.9/5.4	13.1/13.9	21.1/22.2
Mask/+FSA	5.1/5.7	13.4/14.1	21.9/22.9
SCLIP/+FSA	5.2/5.7	13.4/14.2	22.0/23.0

Table S3. Speed (ms) on V100 GPU with 224x224 input.

References

[1] Zhixiang Chi, Li Gu, Huan Liu, Ziqiang Wang, Yanan Wu, Yang Wang, and Konstantinos N Plataniotis. Learning to adapt frozen clip for few-shot test-time domain adaptation. *arXiv preprint arXiv:2506.17307*, 2025. 1

[2] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1

[3] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Clearclip: Decomposing clip representations for dense vision-language inference. *arXiv preprint arXiv:2407.12442*, 2024. 2

[4] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Proxyclip: Proxy attention improves clip for open-vocabulary segmentation. *arXiv preprint arXiv:2408.04883*, 2024. 2

[5] Yuejiang Liu, Parth Kothari, Bastien Van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. Ttt++: When does self-supervised test-time training fail or thrive? *Advances in Neural Information Processing Systems*, 34: 21808–21820, 2021. 1

[6] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022. 1

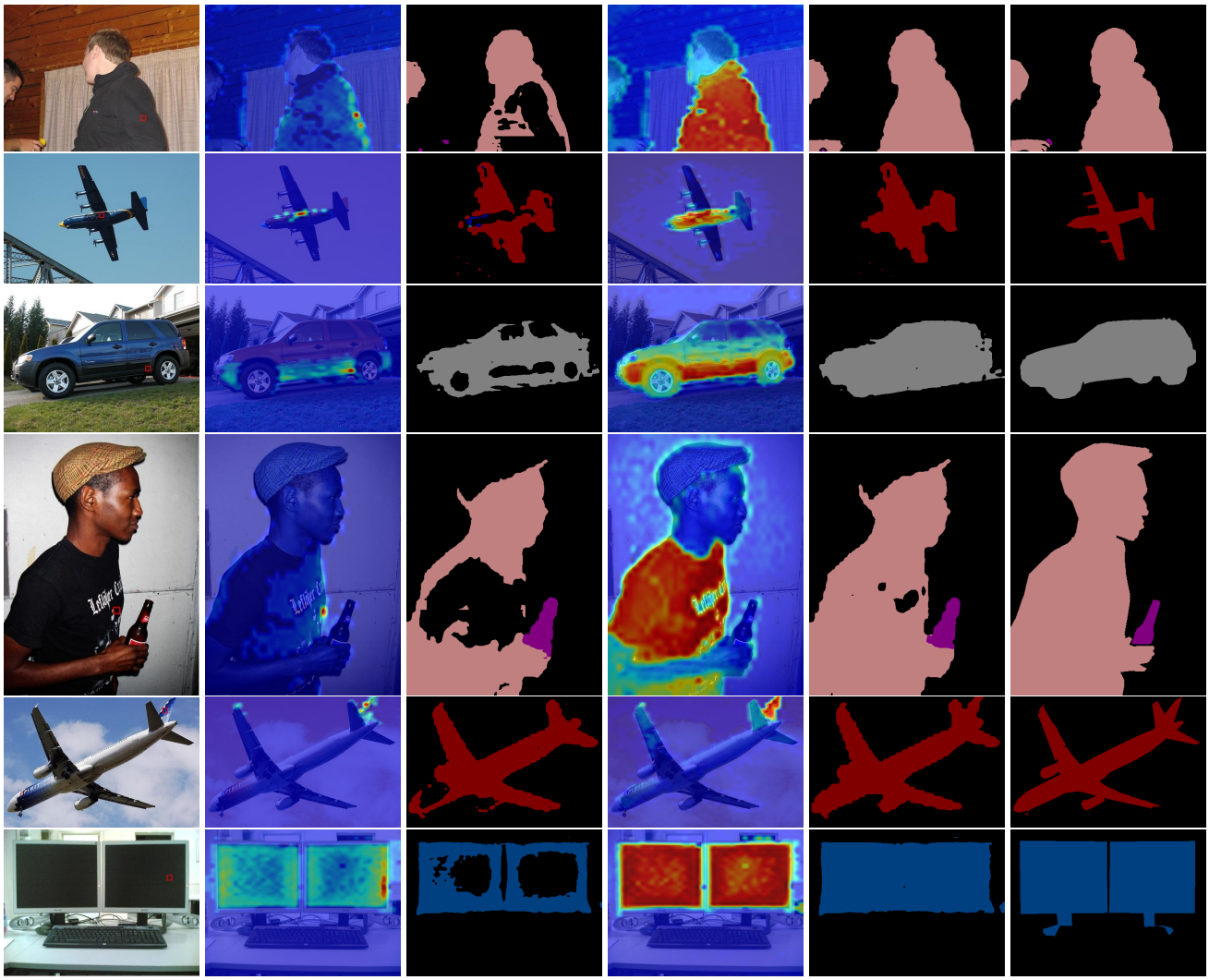


Figure S5. **Comparative visualization of segmentation results: ProxyCLIP vs. integration with our FSA.** The attention maps (second and fourth columns) correspond to the reference patch shown in the first column. ProxyCLIP produces segmentation maps with holes within the same object due to weak attention across regions of the object. In contrast, integrating our FSA effectively aggregates similar patches, enabling the correction of missegmented regions.

[7] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020. 1

[8] Ziqiang Wang, Zhixiang Chi, Yanan Wu, Li Gu, Zhi Liu, Konstantinos Plataniotis, and Yang Wang. Distribution alignment for fully test-time adaptation with dynamic online data streams. In *European Conference on Computer Vision*, pages 332–349. Springer, 2024. 1

[9] Yanan Wu, Zhixiang Chi, Yang Wang, Konstantinos N Plataniotis, and Songhe Feng. Test-time domain adaptation by learning domain-aware batch normalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 15961–15969, 2024. 1

[10] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3713–3722, 2019. 1

[11] Linfeng Zhang, Chenglong Bao, and Kaisheng Ma. Self-distillation: Towards efficient and compact neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4388–4403, 2021. 1

[12] Tao Zhong, Zhixiang Chi, Li Gu, Yang Wang, Yuanhao Yu, and Jin Tang. Meta-dmoe: Adapting to domain shift by meta-distillation from mixture-of-experts. *Advances in Neural Information Processing Systems*, 2022. 1