

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

Zhixiang Chi¹, Yanan Wu², Li Gu³, Huan Liu⁴, Ziqiang Wang³, Yang Zhang⁵
Yang Wang³, Konstantinos N Plataniotis¹

¹University of Toronto ² China Agricultural University ³ Concordia University

⁴ McMaster University ⁵ Beijing Jiaotong University

zhixiang.chi@mail.utoronto.ca <https://github.com/chi-chi-zx/FSA>

Abstract

CLIP exhibits strong visual-textual alignment but struggle with open-vocabulary segmentation due to poor localization. Prior methods enhance spatial coherence by modifying intermediate attention. But, this coherence isn't consistently propagated to the final output due to subsequent operations such as projections. Additionally, intermediate attention lacks direct interaction with text representations, such semantic discrepancy limits the full potential of CLIP.

In this work, we propose a training-free, feedback-driven self-adaptive framework that adapts output-based patch-level correspondences back to the intermediate attention. The output predictions, being the culmination of the model's processing, encapsulate the most comprehensive visual and textual semantics about each patch. Our approach enhances semantic consistency between internal representations and final predictions by leveraging the model's outputs as a stronger spatial coherence prior. We design key modules, including attention isolation, confidence-based pruning for sparse adaptation, and adaptation ensemble, to effectively feedback the output coherence cues. Our method functions as a plug-in module, seamlessly integrating into four state-of-the-art approaches with three backbones (ViT-B, ViT-L, ViT-H). We further validate our framework across multiple attention types (Q-K, self-self, and Proxy augmented with MAE, SAM, and DINO). Our approach consistently improves their performance across eight benchmarks.

1. Introduction

Open-vocabulary semantic segmentation seeks to localize segments and align visual features with novel class descriptions [6, 51, 56]. CLIP models [9, 40], trained on web-scale datasets [41, 55], have demonstrated remarkable zero-shot performance in open-vocabulary tasks due to strong visual-text alignment [28, 40, 61]. However, adapting CLIP for

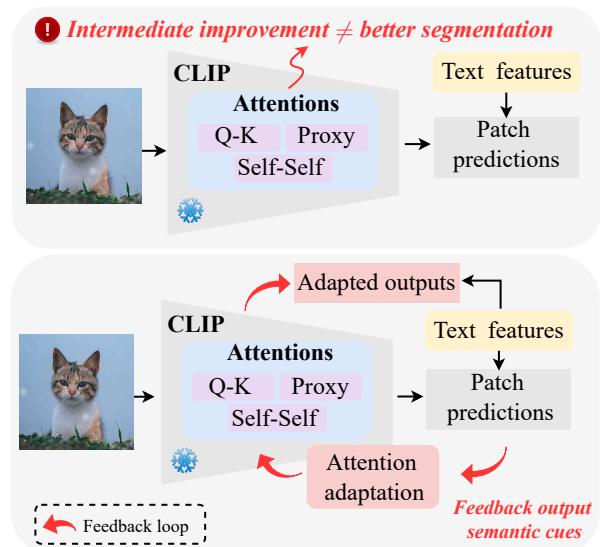


Figure 1. **Comparison with the existing training-free methods.**

Top: Prior works refine *intermediate* attention but face limitations: (1) improvements may not propagate to final segmentation; (2) attention lacks direct class information. **Bottom:** We introduce a feedback-driven self-adaptive mechanism that reintegrates semantic outputs with visual and textual cues into CLIP. Our approach is orthogonal to existing methods and serves as a plug-in.

semantic segmentation remains challenging [67] since its pretraining primarily contrasts global image-text representations, limiting fine-grained localization and causing spatial misalignment in patch-level features [27, 46].

Previous efforts to enhance visual localization have primarily focused on fine-tuning CLIP for segmentation tasks [7, 38, 47, 54, 62]. However, fine-tuning compromises CLIP's robust generalization [12, 13, 49, 67]. Consequently, recent studies have shifted toward a *training-free* paradigm, addressing the factors contributing to localization degradation in CLIP. ClearCLIP [26] identifies residual connections as a major source of noisy segmentation, while

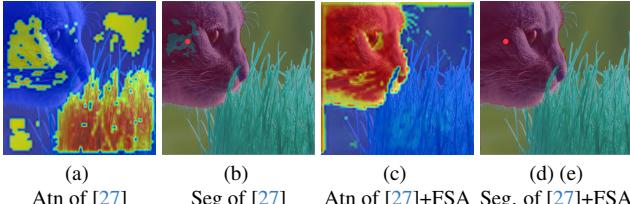


Figure 2. Effectiveness of our feedback self-adaptive mechanism on an image segmented into grass and cat. The attention maps correspond to a reference patch (red dot) on the cat face. In (a), the reference patch misaligns with grass in the attention map, causing incorrect segmentation in (b). Our proposed FSA adjusts (a) using output predictions to generate semantically aligned attention, focusing on cat in (c), and corrects segmentation in (d).

SCLIP [46] highlights the role of self-attention in disrupting spatial information arrangement. Since attention maps capture relationships among patches, patches with similar semantics should be strongly attended to reinforce spatial coherence. Building on this idea, recent works have modulated attention mechanisms, such as replacing Q-K attention with self-self [2, 26, 29, 46] or Proxy attentions [27], both demonstrating improved intermediate spatial consistency.

However, the final segmentation relies on the output predictions rather than the intermediate attention. Enhanced intermediate coherence may deteriorate due to subsequent operations and hence, might not translate into improved segmentation results. More importantly, intermediate attention does not interface directly with class information. Therefore, neglecting the crucial interplay between intermediate attention and output predictions incorporating text representation is inherently sub-optimal. The output predictions intuitively represent the model’s most refined patch-level semantic understanding, encapsulating high-level visual and textual features. We posit that patches with similar predictions likely belong to the same or related class. Leveraging the semantic coherence extracted from the output prediction can provide valuable feedback for further spatial information rearrangement refinement.

Building upon this idea, we propose a *training-free* Feedback-driven Self-adaptive Attention (FSA) framework that adapts output semantic cues back into CLIP through a feedback loop. Specifically, we compute pairwise patch similarities between their class predictions, which provides a stronger prior for guiding intermediate attention than using intermediate representations alone. Modulating the original attention maps with these output-based similarities improves consistency between internal representations and final predictions, enhancing information aggregation among semantically similar patches and improving segmentation accuracy. Fig. 1 shows a high-level comparison with the existing methods. The visual example in Fig. 2 demonstrates that our FSA successfully incorporates the semantic cues at the output to correct the wrongly seg-

TTA	KD	Our feedback mechanism
Minimizes entropy using output logits.	Treats output logits as soft labels.	Computes patch semantic similarity from output logits.

Table 1. Similarity of our FSA with other learning paradigms.

mented patches. Our FSA is an unsupervised adaptation method [5, 31] that leverages the model’s self-outputs, sharing similarities with learning paradigms such as test-time adaptation (TTA [34, 43, 44]) and knowledge distillation (KD [22, 57, 59, 64, 65]), as shown in Table 1.

To focus on the modulation only on the initial intermediate attention maps, we introduce attention isolation to ensure that the initial output predictions reflect the influence of the original attention rather than interference from subsequent operations. Additionally, we propose confidence-based pruning to filter out irrelevant patches while enhancing the impact of semantically relevant regions. Furthermore, we design three adaptation strategies and integrate them into an ensemble to achieve consistent improvements across diverse attention configurations and backbones.

To demonstrate the versatility of our FSA framework, we integrate it into four SoTA methods: MaskCLIP [67], SCLIP [46], ClearCLIP [26], and ProxyCLIP [27]. Additionally, we validate FSA across various attention types, including Q-K, self-self, and Proxy attention, augmented by MAE [21], SAM [25], and DINO [5, 39]. Evaluated on eight open-vocabulary segmentation benchmarks, our FSA consistently enhances performance across these configurations. Our contributions are summarized as:

- We introduce a feedback-driven self-adaptive mechanism that refines the attention process by integrating output-derived semantic coherence cues back into the model.
- We develop key modules, including attention isolation, sparse attention via confidence-based pruning, and adaptation ensemble, to effectively propagate semantic coherence cues from the output.
- Our FSA is training-free, and its adaptation process preserves the original model parameters.
- Our FSA functions as a plug-in module that enhances existing methods and attention configurations, demonstrating effectiveness across eight benchmarks.

2. Related work

Open-vocabulary segmentation. CLIP’s large-scale pre-training has enabled strong zero-shot transfer for open-vocabulary semantic segmentation [23, 24, 28, 61], but its global-level pretraining leads to patch-level localization issues [67]. Fine-tuning on segmentation datasets helps address this by enabling local-to-global alignment, often using self-distillation [7, 30, 50] or integrating spatial coherence from stronger VFM into CLIP [47, 54, 62]. However, fine-tuning can reduce CLIP’s robustness [49], prompting a shift

to training-free approaches that modulate problematic components of CLIP for improved performance. For example, ClearCLIP [26] improves segmentation by altering connections in the last attention block, while MaskCLIP [67] enhances localization with the value embedding. Methods like GEM [2], CLIPSurgery [29], and SCLIP [46] replace the original Q-K attention with self-self attentions. ProxyCLIP [27] combines CLIP and VFM, leveraging spatial feature correspondence for state-of-the-art performance. However, most approaches focus on modulating the forward pass, neglecting spatial consistency and class information encoded in the output predictions.

Output-based knowledge transfer. Output predictions have been effectively used across various machine learning areas, including self-distillation, model compression, domain adaptation, self-supervised learning, uncertainty estimation, and iterative refinement. Self-distillation leverages predictions for intra-model knowledge transfer [33], enhancing generalization and compression [20, 22, 63]. In domain and test-time adaptation, output entropy is minimized to adapt models dynamically without additional labeling [45, 48]. Self-supervised learning employs predictions to create pseudo-labels or enforce consistency across augmented views, refining representations without labeled data [8, 11, 18]. For uncertainty estimation and model calibration, analyzing prediction logits helps assess prediction confidence, with techniques like temperature scaling enhancing reliability [19, 58, 60]. Iterative refinement in vision tasks uses predictions to progressively improve predictions, boosting spatial consistency and segmentation accuracy [4, 10, 32]. These approaches highlight the versatility of output logits for adaptation, optimization, and feedback across diverse tasks. In this work, we introduce a new training-free self-adaptation method that utilizes patch similarity in the output to reorganize the spatial correspondence of intermediate attentions.

3. Preliminaries

Segmentation using CLIP ViT. ViT is composed of attention blocks [16]. Let $\mathbf{x} \in \mathbb{R}^{L \times v} = [x_1, \dots, x_L]^T$ denote the patch tokens, where L is the number of patches and v is the visual feature dimension. Note, we drop `CLS` token and layer norm [1] for simplicity and follow ProxyCLIP [27] and ClearCLIP [26] to modulate the last layer. The resulting visual dense patch features are then obtained by:

$$\mathbf{Q} = \mathbf{x} \cdot W_Q, \quad \mathbf{K} = \mathbf{x} \cdot W_K, \quad \mathbf{V} = \mathbf{x} \cdot W_V, \quad (1)$$

$$\mathbf{z} = \mathbf{y}' + \text{FFN}(\mathbf{y}'), \quad \mathbf{y}' = \mathbf{x} + \text{Proj}(\text{Attn}^{init} \cdot \mathbf{V}) \quad (2)$$

$$\mathbf{Z}^{dense} \in \mathbb{R}^{L \times d} = \mathcal{P}(\mathbf{z}), \quad (3)$$

where W_Q, W_K, W_V , Proj are the projection matrices, FFN is the feed-forward network. \mathcal{P} projects the visual features with dimension v to the visual-text shared space with

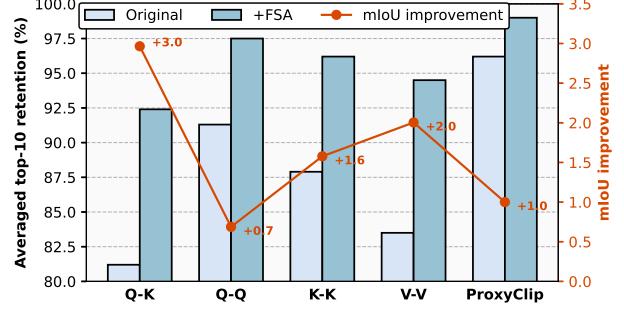


Figure 3. Semantic coherence retention between attention and final predictions (higher is better) and mIoU improvements averaged on 8 benchmarks. Our self-adaptive attention feeds the patch-wise semantics at output back to modulate the intermediate attention to both improve semantic retention and final mIoU.

dimension d , (\cdot) is the matrix multiplication, $\text{Attn}^{init} \in \mathbb{R}^{L \times L}$ represents the *initial* attention maps, which can be the default attention ($\text{SoftMax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\tau}\right)$, where τ is the scaling factor) or other attention configurations, such as self-self ($Q-Q$, $K-K$, $V-V$) [35, 46] and Proxy [27] attention. Note that, in certain methods, e.g., ProxyCLIP and ClearCLIP, the residual and FFN in Eq. 2 are omitted in the final layer.

To obtain the text representation $\mathbf{T} \in \mathbb{R}^{c \times d}$, we input a text prompt, "A photo of a [CLASS]", combined with c class names into the text encoder. Finally, the patch predictions are computed by taking the cosine similarity:

$$\mathbf{Y}^{dense} \in \mathbb{R}^{L \times c} = \cos(\mathbf{Z}^{dense}, \mathbf{T}). \quad (4)$$

The segmentation map $\mathcal{M} \in \mathbb{R}^{L \times 1}$ is obtained by $\arg \max$ along the c dimension of \mathbf{Y}^{dense} .

Motivation. CLIP demonstrates limited localization for segmentation [67]. Since attention governs the arrangement of spatial information [46], recent studies have empirically shown that self-self [35, 46] and Proxy attention [27] exhibit stronger local feature correspondence. However, segmentation ultimately depends on output predictions rather than intermediate attention. The spatial coherence established in Attn^{init} may be disrupted by subsequent operations, preventing it from being fully conveyed to the final output.

We first examine such spatial coherence retention by comparing the intermediate attention Attn^{init} (final layer as in ProxyCLIP) and the final segmentation map \mathcal{M} . Intuitively, if two patches have high mutual attention scores, they should be ideally predicted to the same class. For the i^{th} patch, we identify the k patches with Top- k score in the i^{th} row of Attn^{init} and check whether any of the k patches is predicted as the same class of i^{th} patch in \mathcal{M} :

$$\text{Retention} = \frac{1}{L} \sum_{i=1}^L \max_{j \in \text{Top-}k(\text{Attn}_i^{init})} \mathbb{I}(\mathcal{M}_i = \mathcal{M}_j), \quad (5)$$

where \mathbb{I} is an indicator that equals 1 if the condition is

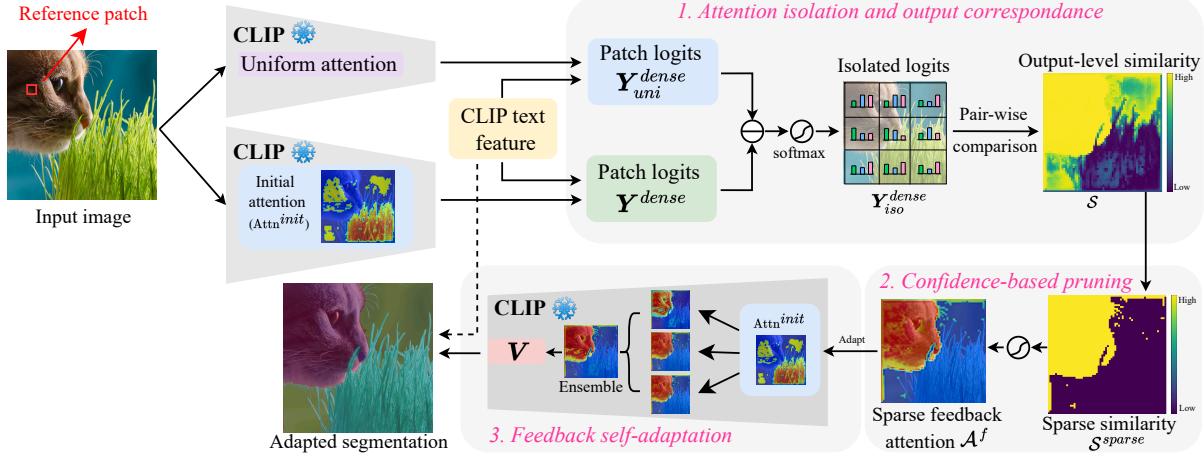


Figure 4. Overview of our feedback-driven self-adaptive attention framework. We isolate the contribution of initial attention using a dual branch with uniform attention. Pairwise patch similarity is then calculated by applying KL divergence to the output logits. A confidence-based pruning step follows to eliminate irrelevant patches and amplify the correlation between related patches. The resulting sparse feedback attention is adapted back to the initial attention maps using three operations, which are subsequently ensembled.

met. To avoid self-comparisons, we set the diagonal elements ($Attn_{ii}^{init}$) to zero. Fig. 3 shows the average Top-10 retention across 8 benchmarks for various attention configurations. Obviously, such spatial coherence is not fully retained in the final predictions. Furthermore, the attention map does not interface with class information directly. In contrast, output predictions capture comprehensive visual semantics and integrate class information from textual features. This motivates our feedback self-adaptive approach, which leverages these outputs as semantic cues to rearrange and modulate spatial information in the intermediate attention $Attn^{init}$. Our method enhances semantic consistency and segmentation performance, as demonstrated in Fig. 3. In Appendix C, we demonstrate another metric and show the retention degradation after each operation.

4. Method

In this section, we present the feedback self-adaptive attention (FSA) framework (Fig. 4). We first construct effective self-adaptive attention through attention isolation (Sec.4.1) and confidence-based pruning (Sec.4.2). Then, we adapt the self-adaptive attention to modulate the spatial information in the initial attention $Attn^{init}$ (Sec.4.3).

4.1. Stand-alone intermediate attention isolation

The initial attention map, $Attn^{init}$, serves as the foundation for spatial coherence, determining both the selection of correlated patches and the intensity of their attention. Thus, our objective is to modulate $Attn^{init}$ solely for spatial knowledge rearrangement. However, as shown in Eq. 2-4, the logits are derived through a sequential process: $Attn^{init} \cdot V \rightarrow$ sub-modules (projections, FFN, etc.) \rightarrow text \rightarrow logits. Consequently, the dense logits, \mathbf{Y}^{dense} , do not exclusively re-

flect the independent contribution of $Attn^{init}$, as it functions merely as an intermediate measure within this pipeline.

To address the interference, we devise an isolation procedure that preserves the stand-alone contribution of $Attn^{init}$ in \mathbf{Y}^{dense} , making it crucial that output relationships reflect only the initial attention maps. Specifically, we attach a parallel branch with a manually crafted *uniform* attention map $Attn^{uni} = \frac{1}{L}\mathbf{1}_{L \times L}$, ensuring every patch is equally attended. The resulting logits \mathbf{Y}_{uni}^{dense} are obtained following the same steps in Eqs. 2-4. Both $Attn^{uni}$ and $Attn^{init}$ traverse identical sub-modules and text-alignment stages, allowing us to isolate the learned attention’s net effect by subtracting the uniform-based logits from the original. Therefore, we can pinpoint how much of the final logits is genuinely driven by the learned selective attention patterns rather than a globally “spread out” context. The resulting subtraction is then transformed into a probability distribution via softmax:

$$\mathbf{Y}_{iso}^{dense} = \text{Softmax}(\mathbf{Y}^{dense} - \mathbf{Y}_{uni}^{dense}). \quad (6)$$

We then quantify the relationships between patches by calculating pair-wise KL divergence:

$$\mathcal{D}(i, j) = \text{KL}(\mathbf{Y}_{iso,i}^{dense} \parallel \mathbf{Y}_{iso,j}^{dense}), \quad (7)$$

where $\mathcal{D} \in \mathbb{R}^{L \times L}$ records the divergence for each pair of patches. To emphasize patch correspondence rather than difference, we convert the divergence to similarity as:

$$S \in \mathbb{R}^{L \times L} = \frac{1}{\mathcal{D} + 1}. \quad (8)$$

The KL divergence, bounded in $[0, -\infty]$, is mapped to $[0, 1]$, with a score of 1 indicating maximum similarity.

When attention isolation is applied, the output logits align better with the intermediate attention, leading to further segmentation gains.

4.2. Confidence-based sparse attention

The similarity map \mathcal{S} not only encodes visual semantic coherence among patches but more importantly, it integrates *class cues* from text representations. Capturing each patch’s response to each class offers a robust foundation for adapting attention with contextually relevant cues. However, in segmentation tasks, patches may belong to various classes, each with unique semantic information. Since \mathcal{S} contains similarities for every patch pair, low-correlated patches can dilute the relationships between strongly correlated patches. Thus, it is effective to suppress irrelevant patches while amplifying semantically coherent ones.

To this end, we propose a confidence-based sparse attention. Specifically, we first normalize each row of \mathcal{S} using softmax to obtain $\hat{\mathcal{S}}$, resulting in a confidence distribution across patches. Let $\hat{\mathcal{S}}_i$ denote the i^{th} row which represents the confidence of similarity between the i^{th} patch and all other patches. We sort $\hat{\mathcal{S}}_i$ in descending order ($\hat{\mathcal{S}}_i^{\text{sorted}}$) and record the sorting indices (idx). For each position, we compute the *cumulative* confidence probability:

$$\mathcal{C}_{i,j}^{\text{sorted}} = \sum_{k=1}^j \hat{\mathcal{S}}_{i,k}^{\text{sorted}}. \quad (9)$$

We then revert to their original order using idx : $\mathcal{C}_i = \mathcal{C}_i^{\text{sorted}}[idx]$. Repeating for each row yields the final cumulative confidence map \mathcal{C} . We then set a confidence level p to selectively amplify higher similarity values with adaptive exponential scaling and fully suppress the others:

$$\mathcal{S}_{i,j}^{\text{sparse}} = \begin{cases} \mathcal{S}_{i,j} \cdot e^{\lambda \cdot \mathcal{C}_{i,j}}, & \text{if } \mathcal{C}_{i,j} \leq p \text{ or } j = 1 \\ -\infty, & \text{otherwise,} \end{cases} \quad (10)$$

where λ and p are the only hyperparameters in our method, controlling the scaling sharpness and the cutoff confidence, set empirically to 2 and 0.45, respectively. Finally, we convert the sparse similarity map into our feedback attention: $\mathcal{A}^f = \text{softmax}(\mathcal{S}^{\text{sparse}})$.

In summary, our confidence-based sparse attention (Eq. 9 & 10) can be interpreted as: *selecting patches starting with the highest confidence until cumulative confidence of selected patches reaches p* . It emphasizes semantically coherent patches while ignoring low-confidence ones. A pseudo-code can be found in the supplementary.

4.3. Feedback self-adaptive attention

The resulting attention \mathcal{A}^f captures the ultimate semantic relationships between pairs of patches, derived from the initial attention. We reintegrate this feedback into the original pipeline (Eq. 2-4) to enhance the spatial arrangement

of information. Specifically, we propose three *training-free* adaptations to modify the interaction with \mathbf{V} in Eq. 2:

$$\text{Attn}^{\text{init}} \cdot \mathbf{V} \xrightarrow{\text{adaptation}} \left\{ \begin{array}{l} \mathcal{A}^f \cdot (\text{Attn}^{\text{init}} \cdot \mathbf{V}), \\ \text{Attn}^{\text{init}} \cdot (\mathcal{A}^f \cdot \mathbf{V}), \\ \mathcal{A}^f \cdot \mathbf{V}. \end{array} \right. \quad (11)$$

$$\text{Attn}^{\text{init}} \cdot (\mathcal{A}^f \cdot \mathbf{V}), \quad (12)$$

$$\mathcal{A}^f \cdot \mathbf{V}. \quad (13)$$

Eq. 11 captures the $\text{Attn}^{\text{init}}$ -driven relationships first, then selectively amplifies or suppresses them based on feedback, effectively introducing a two-step refinement. Eq. 12 immediately influences \mathbf{V} before initial attention is applied, giving higher priority to semantic coherence and confidence information from the output logits. Eq. 13 emphasizes only the feedback-driven information solely relying on the output semantic coherence and confidence. We observe that each of the three adaptations offers distinct benefits depending on the backbone, attention configuration, and benchmark. To leverage their combined strengths, we propose an ensemble approach by isolating \mathbf{V} in Eq. 11-13:

$$\left(\frac{\mathcal{A}^f \cdot \text{Attn}^{\text{init}} + \text{Attn}^{\text{init}} \cdot \mathcal{A}^f + \mathcal{A}^f}{3} \right) \cdot \mathbf{V}. \quad (14)$$

We denote the adapted logits from Eq. 11-14 as $\mathbf{Y}_1^{\text{dense}}$, $\mathbf{Y}_2^{\text{dense}}$, $\mathbf{Y}_3^{\text{dense}}$, and $\mathbf{Y}_{\text{ens}}^{\text{dense}}$ which are processed using the arg max to generate the adapted segmentation map.

5. Experiments

In this section, we first validate our FSA framework by integrating it with various SoTA methods and attention configurations. We then conduct comprehensive ablation studies to assess the effectiveness of each proposed component.

5.1. Implementation details and datasets

Architectures. To show the versatility of FSA, we test on 3 backbones: ViT-B/16 and ViT-L/14 from CLIP [40], and ViT-H/14 from OpenCLIP [9]. We adopt DINO ViT-B/8 [5] for ProxyCLIP as default unless specified otherwise.

Implementations. Our FSA functions as a plug-in module that can be integrated into SoTA methods, modifying only the model component. Note that, we preserve their original data processing pipeline. For example patch size and stride size are kept the same. All experiments report mean Intersection over Union (mIoU) on validation sets, without training or post-processing. Implementation is based on the MMSegmentation codebase [14].

Datasets. We follow TCL [6], ProxyCLIP [27], SCLIP [46] and CLIP-DINOiser [53] to evaluate on eight standard benchmarks: (i) with background class: PASCAL VOC [17] (VOC), PASCAL Context [37] (Context), and COCO Object [3] (Object); and (ii) without background class: PASCAL VOC20 [17] (VOC20), PASCAL Context59 [37] (Context59), COCO Stuff [3] (Stuff), Cityscapes [15] (City), and ADE20K [66] (ADE).

Models	Methods	VOC	Context	Object	VOC20	Context59	Stuff	City	ADE	Avg.
CLIP ViT-B/16	MaskCLIP _{ECCV'22} [67]	38.8	23.6	20.6	74.9	26.4	16.4	12.6	9.8	27.9
	+ FSA	47.7	31.0	29.2	78.3	34.2	21.4	28.4	16.0	35.8 (+7.9)
	SCLIP _{ECCV'24} [46]	59.1	30.4	30.5	80.4	34.2	22.4	32.2	16.1	38.2
	+ FSA	61.5	33.3	33.9	82.8	36.8	24.4	34.7	17.5	40.6 (+2.4)
	ClearCLIP _{ECCV'24} [26]	51.8	32.6	33.0	80.9	35.9	23.9	30.0	16.7	38.1
	+ FSA	53.0	36.6	33.2	81.3	33.8	24.3	30.8	17.4	38.8 (+0.7)
CLIP ViT-L/14	ProxyCLIP _{ECCV'24} [27]	61.3	35.3	37.5	80.3	39.1	26.5	38.1	20.2	42.3
	+ FSA (Ours)	63.7	36.1	38.0	82.3	39.9	27.0	38.8	20.5	43.3 (+1.0)
	MaskCLIP _{ECCV'22} [67]	23.3	11.7	7.2	29.4	12.4	8.8	11.5	7.2	13.9
	+ FSA	44.8	26.8	27.8	73.9	29.4	19.0	23.1	16.2	32.6 (+18.7)
	SCLIP _{ECCV'24} [46]	43.5	22.3	25.0	69.1	25.2	17.6	18.6	10.9	29.0
	+ FSA	48.1	27.8	30.8	79.9	30.3	20.4	27.1	15.9	35.0 (+6.0)
CLIP ViT-H/14	ClearCLIP _{ECCV'24} [26]	46.1	29.6	26.7	80.0	30.1	19.9	27.9	15.0	34.4
	+ FSA	47.5	30.8	27.9	80.4	30.2	20.4	27.2	16.8	35.2 (+0.8)
	ProxyCLIP _{ECCV'24} [27]	60.6	34.5	39.2	83.2	37.7	25.6	40.1	22.6	42.9
	+ FSA (Ours)	61.8	34.9	40.2	84.1	38.1	25.9	41.2	22.9	43.6 (+0.7)
	MaskCLIP _{ECCV'22} [67]	31.4	13.3	16.2	41.7	15.8	8.4	17.7	10.4	19.3
	+ FSA	44.8	27.7	27.6	71.3	30.5	20.7	26.1	19.0	33.4 (+14.1)

Table 2. **Integration of FSA into the state-of-the-art methods for open-vocabulary semantic segmentation on eight standard benchmarks.** Our feedback self-adaptive framework consistently improves previous approaches across all datasets and three different backbones.

CLIP	VFM	Methods	V21	C60	Obj	V20	C59	Stf	City	ADE	Avg.
ViTB/16	SAM ViTB/16	Proxy	59.3	33.6	35.4	80.4	37.0	25.0	37.0	19.1	40.8
	+ FSA	60.7	34.0	35.8	81.8	37.4	25.2	37.9	19.3	41.5 (+0.7)	
	MAE ViTB/16	Proxy	52.2	30.4	30.8	76.3	33.5	23.1	30.1	17.1	36.7
	+ FSA	54.3	30.9	31.3	78.1	33.9	23.4	33.6	17.5	37.9 (+1.2)	
	DINOv2 ViTB/14	Proxy	58.6	33.8	37.0	83.0	37.2	25.4	33.9	19.7	41.1
	+ FSA	59.2	33.9	37.4	84.0	37.5	25.5	34.4	19.7	41.4 (+0.3)	
ViTL/14	SAM ViTB/16	Proxy	57.2	32.6	36.5	82.3	35.6	24.2	39.1	20.7	41.0
	+ FSA	58.5	33.0	38.0	83.1	36.1	24.5	40.5	21.0	41.8 (+0.8)	
	MAE ViTB/16	Proxy	49.0	27.8	31.6	78.3	30.2	20.8	31.8	17.2	35.8
	+ FSA	52.8	29.9	34.9	80.2	32.5	22.6	34.7	19.0	38.3 (+2.5)	
	DINOv2 ViTB/14	Proxy	56.6	33.0	36.7	85.2	36.2	24.6	35.2	21.6	41.1
	+ FSA	57.4	33.3	38.0	85.8	36.5	24.7	36.1	21.9	41.7 (+0.6)	
ViTH/14	SAM ViTB/16	Proxy	63.5	34.1	36.7	84.0	37.9	25.0	41.1	22.0	43.1
	+ FSA	64.9	34.4	37.6	85.5	38.0	25.1	42.6	21.9	43.7 (+0.6)	
	MAE ViTB/16	Proxy	54.7	29.8	32.2	80.6	32.9	21.8	34.9	19.4	38.3
	+ FSA	58.7	32.0	35.2	82.2	35.3	24.0	37.9	21.3	40.8 (+2.5)	
	DINOv2 ViTB/14	Proxy	61.5	34.0	37.3	86.1	37.8	26.2	37.8	23.4	43.0
	+ FSA	63.0	34.4	38.5	87.4	38.4	26.4	39.7	23.7	43.9 (+0.9)	

Table 3. **Integration of our FSA into ProxyCLIP with various VFMs.** Consistent improvements across VFMs, benchmarks, and backbones highlight the versatility and effectiveness of our FSA.

5.2. Main results

Integration into the SoTA methods. Table 2 reports the integration of our FSA with training-free SoTA methods including: MaskCLIP [67], SCLIP [46], ClearCLIP [26] and ProxyCLIP [27]. Our method consistently improves all of them across all eight benchmarks. Notably, we surpass MaskCLIP by a significant margin of **+7.9 mIoU**, **+18.7 mIoU** and **+14.1 mIoU** using ViT-B/16, L/14, and H/14, respectively. Obvious improvement is also observed on SCLIP. ProxyCLIP leverages external knowledge from VFMs, providing strong spatial coherence. However, our FSA system further enhances performance across diverse

CLIP	Attention	V21	C60	Obj	V20	C59	Stf	City	ADE	Avg.
ViTB/16	Q-K + FSA	34.8	19.0	21.4	76.1	21.5	14.2	15.9	10.7	26.7
	+ FSA	40.7	22.2	23.5	76.9	25.2	16.4	20.1	12.3	29.7 (+3.0)
	Q-Q + FSA	55.6	31.7	32.9	81.4	35.5	23.9	30.3	17.9	38.7
	+ FSA	56.9	32.2	34.2	81.9	35.9	24.2	31.2	18.2	39.3 (+0.6)
	K-K + FSA	54.5	31.5	30.0	77.5	35.4	23.4	33.0	18.1	37.9
	+ FSA	57.5	32.6	33.3	80.4	36.2	24.1	33.5	18.4	39.5 (+1.6)
ViTL/14	V-V + FSA	51.4	29.7	30.1	74.1	32.8	21.5	30.9	16.2	35.8
	+ FSA	54.4	31.4	31.9	78.4	34.6	22.6	32.2	17.2	37.8 (+2.0)
	Q-K + FSA	34.7	19.2	24.7	75.7	21.3	14.6	20.6	11.0	27.7
	+ FSA	40.8	22.0	27.7	77.3	24.1	16.5	24.3	12.2	30.6 (+2.9)
	Q-Q + FSA	49.2	26.7	31.1	80.1	29.5	19.8	30.6	16.6	35.5
	+ FSA	50.4	26.9	31.4	80.4	29.8	19.9	30.7	16.8	35.8 (+0.3)
ViTH/14	K-K + FSA	49.3	26.6	28.7	76.6	30.7	20.5	31.8	17.0	35.2
	+ FSA	51.1	27.3	30.3	78.5	30.8	20.4	31.7	17.2	35.9 (+0.7)
	V-V + FSA	47.4	27.0	31.0	78.7	29.8	20.2	31.1	18.0	35.4
	+ FSA	49.4	27.7	31.4	79.7	30.6	20.6	31.8	18.3	36.2 (+0.8)
	Q-K + FSA	36.6	20.9	23.7	81.5	23.0	15.5	25.3	13.4	30.0
	+ FSA	46.3	24.7	30.8	83.5	26.9	18.1	30.8	15.5	34.6 (+4.6)
ViTH/14	Q-Q + FSA	51.1	26.8	29.7	79.1	29.9	19.6	32.7	17.7	35.8
	+ FSA	52.8	27.7	30.5	79.2	30.8	20.4	34.0	18.6	36.7 (+0.9)
	K-K + FSA	51.6	27.2	30.1	73.5	30.3	20.0	34.0	18.8	35.7
	+ FSA	53.4	28.1	30.8	75.0	31.2	20.8	34.2	19.5	36.6 (+0.9)
	V-V + FSA	51.2	28.1	30.0	80.2	31.0	21.3	33.7	19.7	36.9
	+ FSA	52.7	28.6	30.7	80.5	31.7	21.6	34.4	19.9	37.5 (+0.6)

Table 4. **Integration of our FSA into Q-K and self-self attentions.** Our method consistently shows improvements across all attention configurations. Specifically, we improve the original Q-K by a large margin, which is less effective in localization.

CLIP architectures, pushing their boundaries.

Integration with various attention configurations. The proposed feedback self-adaptive mechanism operates independently of the specific intermediate attention, demonstrating its versatility when integrated with various attention configurations. Table 3 shows results of applying our method to ProxyCLIP with different VFMs (MAE [21],

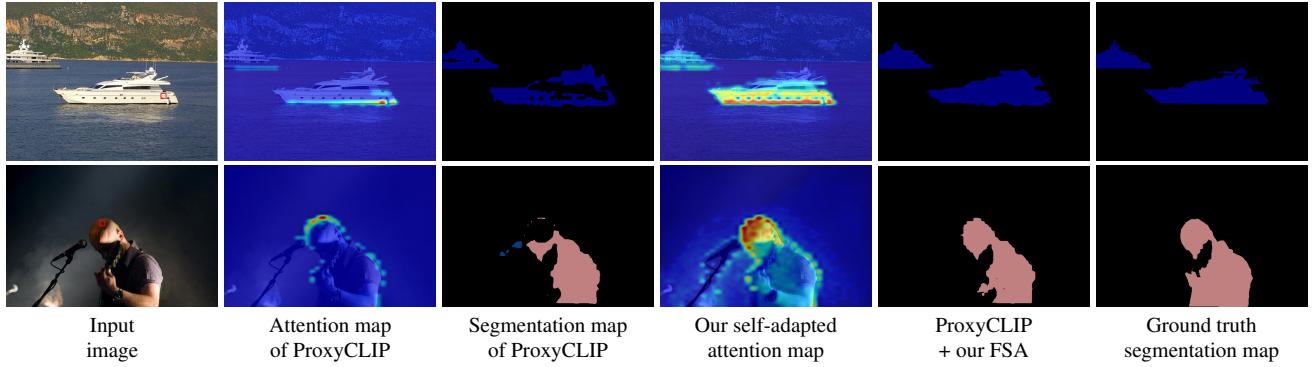


Figure 5. **Visualization of segmentation results of ProxyCLIP integrated with our FSA.** The attention maps (2^{nd} and 4^{th} columns) correspond to the reference patch shown in the 1^{st} column. ProxyCLIP produces with holes within the same object due to weak attention across regions of the object. In contrast, our FSA effectively aggregates similar patches, enabling the correction of missegmented regions.

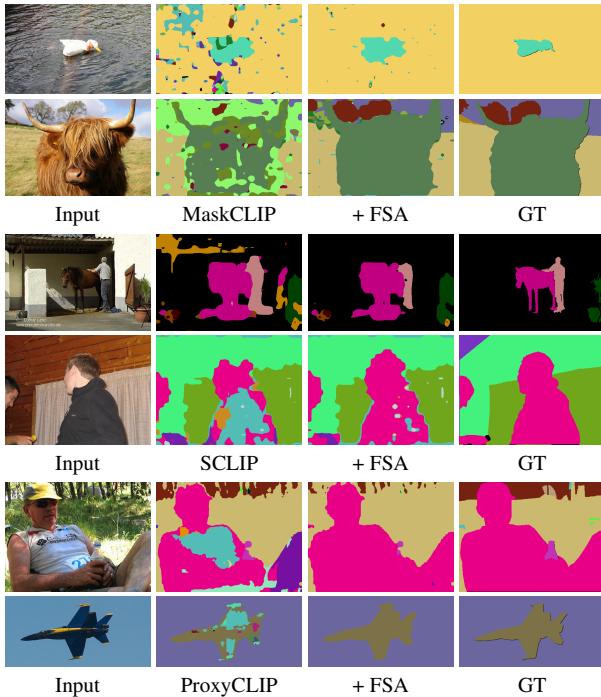


Figure 6. **Qualitative results.** By integrating our feedback self-adaptive mechanism, we correct missegmented patches, ensuring consistent segmentation within each object.

SAM [25], and DINOv2 [39]), each displaying varying levels of spatial correspondence. Despite these differences, our method proves flexible, achieving consistent improvements across all VFM, benchmarks, and backbones. Notably, while MAE shows weaker baseline performance, our feedback correction enhances it by **+1.2 mIoU**, **+2.5 mIoU**, and **+2.5 mIoU** across the three backbones.

We also integrate our method with inherent Q-K and self-self attention mechanisms using the codebase from ProxyCLIP. As summarized in Table 4, improvements are observed across all attention types on the three backbones.

Scaling	Pruning	ViT-B/16	ViT-L/14	ViT-H/14
✗	✗	22.6	23.5	25.5
✗	Fixed ratio	36.6	37.8	41.0
✗	Fixed threshold	37.2	38.6	40.1
✗	Confidence-based	38.2	39.4	42.5
✓	Confidence-based	43.3	43.6	45.8

Table 5. **Ablation on confidence-based pruning with ProxyCLIP over 8 benchmarks.** Our adaptive confidence-based pruning selects relevant patches while suppressing irrelevant ones. Additionally, adaptive exponential scaling emphasizes the importance of these selected patches, further enhancing their contribution.

The most significant gains are seen with Q-K attention (**+3.0 mIoU**, **+2.9 mIoU**, and **+4.6 mIoU**), as it is less effective for patch-level alignment. Overall, our self-adaptive framework is compatible with various attention configurations, yielding performance gains ranging from **+0.3 mIoU** to **+4.6 mIoU** as shown in Tables 3 and 4.

Qualitative comparison with SoTA. Fig. 5 illustrates that ProxyCLIP often focuses on object edges, resulting in unstable segmentation. By integrating our adaptive FSA, we refine the intermediate attention to focus solely on the main regions of the same object, thereby improving the segmentation. Fig. 6 presents additional qualitative results with our FSA incorporated into MaskCLIP, SCLIP, and ProxyCLIP. Leveraging the feedback mechanism, we correct misclassified regions, leading to more consistent segmentation. Additional results are provided in Appendix B.

6. Ablation studies

Confidence-based pruning. We generate sparse feedback attention through confidence-based pruning, highlighting semantically related patches likely belonging to the same class. As shown in Table 5, both fixed-ratio and threshold-based pruning improve segmentation by filtering irrelevant patches, while confidence-based pruning achieves greater

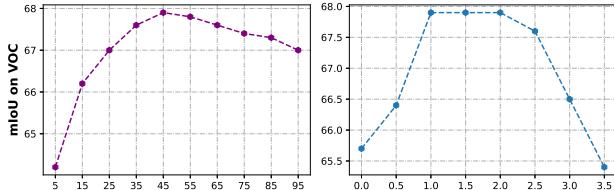


Figure 7. **Sensitivity on p (left) and λ (right) with ViT-L/14.** As p increases, mIoU rises then falls with added irrelevant patches. λ adjusts amplification and remains stable between 1 and 2.5.

Attention isolation	MaskCLIP	SCLIP	ProxyCLIP
w/o FSA (baseline)	13.9	29.0	42.9
\times	29.8 (+15.9)	33.2 (+4.2)	43.4 (+0.5)
\checkmark	32.6 (+18.7)	35.0 (+6.0)	43.6 (+0.7)

Table 6. **Validation on attention isolation with ViT-L/14.** Such isolation ensures that the output logits directly reflect the intermediate attention, enabling more consistent feedback adaptation.

gains by adaptively selecting correlated patches. Additionally, adaptive exponential scaling amplifies relative differences in patch similarity, further enhancing segmentation.

Fig. 7 illustrates the sensitivity of parameters p and λ using ViT-L/14 on the VOC dataset. As p increases, more relevant patches are included; however, beyond a certain point, irrelevant patches are also incorporated, which affects attention on spatially correlated patches. Increasing λ improves relative distances between patch similarities, allowing relevant patches to dominate. Performance remains stable between 1.0 and 2.5 but decreases when too few patches dominate, potentially excluding important patches. Fig. 9 shows the distributions of row-wise pruning ratios, underscoring the necessity of an adaptive pruning mechanism.

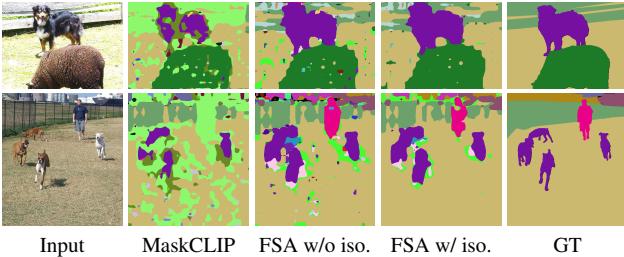


Figure 8. **Qualitative results on attention isolation (iso.).**

Attention isolation. Our feedback attention modulates the intermediate attention, making it crucial that output relationships reflect only the initial attention maps. As shown in Table 6, we observe improvements over 3 methods even without isolation. When attention isolation is applied, the output logits align better with the intermediate attention, leading to further segmentation gains. The qualitative examples in Fig. 8 also demonstrate that attention isolation is

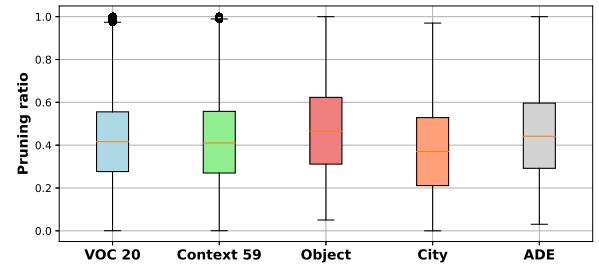


Figure 9. **Distribution of pruning ratio.** The variety of distribution demonstrates the necessity of our adaptive mechanism.

\mathbf{Y}_1^{dense}	\mathbf{Y}_2^{dense}	\mathbf{Y}_3^{dense}	MaskCLIP	SCLIP	ProxyCLIP
w/o FSA (baseline)			13.9	29.0	42.9
\checkmark			32.1 (+18.2)	34.2 (+5.2)	43.5 (+0.6)
	\checkmark		31.8 (+17.9)	33.9 (+4.9)	43.4 (+0.5)
		\checkmark	30.9 (+17.0)	34.5 (+5.5)	43.5 (+0.6)
		Ensenble	32.6 (+18.7)	35.0 (+6.0)	43.6 (+0.7)

Table 7. **Validation on ensembling with ViT-L/14.** Our ensembling strategy combines the strengths of each adaptation, resulting in consistent and enhanced segmentation across methods.

Methods	ViT-B/16	ViT-L/14	ViT-H/14
Proxy	12.9ms	20.1ms	28.1ms
+ FSA	13.6ms	21.1ms	29.2ms

Table 8. **Speed on V100 GPU with patch resolution of 224x224.**

effective in removing noise in segmentation maps.

Ensembling on adapted attentions. Table 7 presents the individual segmentation results for each adaptation strategy in Eqs. 11-13. While each method shows improvement, the best performance varies across SoTA methods. Our ensembling strategy combines their strengths, resulting in consistent and enhanced segmentation performance.

Cost analysis. Table 8 reports speed averaged over 3 trials of 100 forward passes. As we only modify the last attention as in ProxyCLIP, FSA only adds 3-5% overhead.

Iterative adaptation. Intuitively, our FSA supports iterative adaptation, but, further improvement is not observed.

7. Conclusion

In this work, we introduce FSA, a novel training-free self-adaptive framework that enhances spatial coherence in attention maps by leveraging the model’s own patch predictions as feedback. Our method effectively bridges the gap between intermediate attention and final outputs by integrating visual and textual cues, resulting in more accurate patch-level correspondence and improved segmentation. Integrated into various attention configurations, FSA consistently improves on 8 standard benchmarks.

References

- [1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 3
- [2] Walid Bousselham, Felix Petersen, Vittorio Ferrari, and Hilde Kuehne. Grounding everything: Emerging localization properties in vision-language transformers. *arXiv preprint arXiv:2312.00878*, 2023. 2, 3
- [3] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 5
- [4] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6154–6162, 2018. 3
- [5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2, 5
- [6] Junbum Cha, Jonghwan Mun, and Byungseok Roh. Learning to generate text-grounded mask for open-world semantic segmentation from only image-text pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11165–11174, 2023. 1, 5
- [7] Jun Chen, Deyao Zhu, Guocheng Qian, Bernard Ghanem, Zhicheng Yan, Chenchen Zhu, Fanyi Xiao, Sean Chang Culatana, and Mohamed Elhoseiny. Exploring open-vocabulary semantic segmentation from clip vision encoder distillation only. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 699–710, 2023. 1, 2
- [8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, pages 1597–1607. PMLR, 2020. 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023. 1, 5
- [10] Zhixiang Chi, Rasoul Mohammadi Nasiri, Zheng Liu, Juwei Lu, Jin Tang, and Konstantinos N Plataniotis. All at once: Temporally adaptive multi-frame interpolation with advanced motion modeling. In *European conference on computer vision*, pages 107–123. Springer, 2020. 3
- [11] Zhixiang Chi, Yang Wang, Yuanhao Yu, and Jin Tang. Test-time fast adaptation for dynamic scene deblurring via meta-auxiliary learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9137–9146, 2021. 3
- [12] Zhixiang Chi, Li Gu, Tao Zhong, Huan Liu, YUANHAO YU, Konstantinos N Plataniotis, and Yang Wang. Adapting to distribution shift by visual domain prompt generation. In *The Twelfth International Conference on Learning Representations*, 2024. 1
- [13] 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, 12
- [14] MMSegmentation Contributors. Mmsegmentation: Open-mmlab semantic segmentation toolbox and benchmark, 2020. 5
- [15] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 5
- [16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3
- [17] Mark Everingham and John Winn. The pascal visual object classes challenge 2012 (voc2012) development kit. *Pattern Anal. Stat. Model. Comput. Learn., Tech. Rep.*, 2007(1-45):5, 2012. 5
- [18] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:21271–21284, 2020. 3
- [19] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1321–1330. PMLR, 2017. 3
- [20] Shayan Mohajer Hamidi, Xizhen Deng, Renhao Tan, Linfeng Ye, and Ahmed Hussein Salamat. How to train the teacher model for effective knowledge distillation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024. 3
- [21] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2, 6
- [22] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2, 3, 12
- [23] 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 *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2
- [24] Aisha Urooj Khan, Hilde Kuehne, Chuang Gan, Niels Da Victoria Lobo, and Mubarak Shah. Weakly supervised grounding for vqa in vision-language transformers. In *European*

- Conference on Computer Vision*, pages 652–670. Springer, 2022. 2
- [25] 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. *arXiv preprint arXiv:2304.02643*, 2023. 2, 7
- [26] 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. 1, 2, 3, 6, 12
- [27] 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. 1, 2, 3, 5, 6, 12
- [28] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 1, 2
- [29] Yi Li, Hualiang Wang, Yiqun Duan, and Xiaomeng Li. Clip surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv preprint arXiv:2304.05653*, 2023. 2, 3
- [30] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023. 2
- [31] Hanwen Liang, Niamul Quader, Zhixiang Chi, Lizhe Chen, Peng Dai, Juwei Lu, and Yang Wang. Self-supervised spatiotemporal representation learning by exploiting video continuity. In *Proceedings of the AAAI conference on artificial intelligence*, 2022. 2
- [32] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1925–1934, 2017. 3
- [33] Huan Liu, Li Gu, Zhixiang Chi, Yang Wang, Yuanhao Yu, Jun Chen, and Jin Tang. Few-shot class-incremental learning via entropy-regularized data-free replay. In *European Conference on Computer Vision*, pages 146–162. Springer, 2022. 3
- [34] Huan Liu, Zhixiang Chi, Yuanhao Yu, Yang Wang, Jun Chen, and Jin Tang. Meta-auxiliary learning for future depth prediction in videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023. 2
- [35] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 3
- [36] 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. 12
- [37] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 891–898, 2014. 5
- [38] Jishnu Mukhoti, Tsung-Yu Lin, Omid Poursaeed, Rui Wang, Ashish Shah, Philip HS Torr, and Ser-Nam Lim. Open vocabulary semantic segmentation with patch aligned contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19413–19423, 2023. 1
- [39] Maxime Oquab, Timothée Darcret, 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. 2, 7
- [40] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 5
- [41] 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. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 1
- [42] 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. 12
- [43] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*. PMLR, 2020. 2
- [44] 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. 2, 12
- [45] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations (ICLR)*, 2021. 3
- [46] Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. *arXiv preprint arXiv:2312.01597*, 2023. 1, 2, 3, 5, 6
- [47] 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 *UniReps*:

- the First Workshop on Unifying Representations in Neural Models*, 2023. 1, 2
- [48] 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. 3, 12
- [49] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7959–7971, 2022. 1, 2
- [50] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. Clipself: Vision transformer distills itself for open-vocabulary dense prediction. *arXiv preprint arXiv:2310.01403*, 2023. 2
- [51] Yanan Wu, Zhixiang Chi, Yang Wang, and Songhe Feng. Metagcd: Learning to continually learn in generalized category discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1655–1665, 2023. 1
- [52] 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. 12
- [53] Monika Wysoczańska, Oriane Siméoni, Michaël Ramamonjisoa, Andrei Bursuc, Tomasz Trzcinski, and Patrick Pérez. Clip-dinoiser: Teaching clip a few dino tricks. *arXiv preprint arXiv:2312.12359*, 2023. 5
- [54] Monika Wysoczańska, Michaël Ramamonjisoa, Tomasz Trzcinski, and Oriane Siméoni. Clip-diy: Clip dense inference yields open-vocabulary semantic segmentation for-free. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1403–1413, 2024. 1, 2
- [55] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. *arXiv preprint arXiv:2309.16671*, 2023. 1
- [56] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022. 1
- [57] En-hui Yang and Linfeng Ye. Markov knowledge distillation: Make nasty teachers trained by self-undermining knowledge distillation fully distillable. In *European Conference on Computer Vision*, pages 154–171. Springer, 2024. 2
- [58] En-Hui Yang, Shayan Mohajer Hamidi, Linfeng Ye, Renhao Tan, and Beverly Yang. Conditional mutual information constrained deep learning: Framework and preliminary results. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pages 569–574. IEEE, 2024. 3
- [59] Linfeng Ye, Shayan Mohajer Hamidi, Renhao Tan, and EN-HUI YANG. Bayes conditional distribution estimation for knowledge distillation based on conditional mutual information. In *The Twelfth International Conference on Learning Representations*, 2020. 2
- [60] Linfeng Ye, Shayan Mohajer Hamidi, and EN-HUI YANG. Towards undistillable models by minimizing conditional mutual information. *Transactions on Machine Learning Research*, 2025. 3
- [61] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 1, 2
- [62] Haobo Yuan, Xiangtai Li, Chong Zhou, Yining Li, Kai Chen, and Chen Change Loy. Open-vocabulary sam: Segment and recognize twenty-thousand classes interactively. *arXiv preprint arXiv:2401.02955*, 2024. 1, 2
- [63] 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. 3, 12
- [64] 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. 2, 12
- [65] 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. 2, 12
- [66] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019. 5
- [67] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022. 1, 2, 3, 6

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. S10, 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. S11, we show expanded comparison with MaskCLIP and SCLIP with the examples in Row 5-6 of Fig. 6.

In Fig. S14, 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.

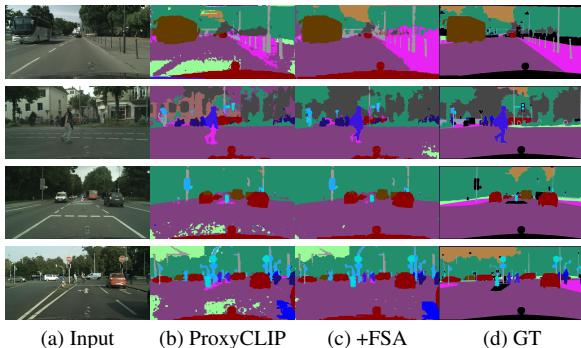


Figure S10. Qualitative results on Cityscapes. By integrating our feedback self-adaptive mechanism, we correct missegmented patches by ProxyCLIP, ensuring consistent segmentation within each object. We can clearly observe that our segmentation is more consistent across whole objects.

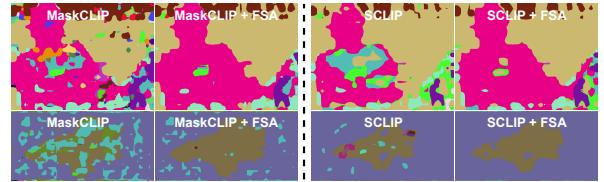


Figure S11. Examples of Row 5-6 in Fig.6 for MaskCLIP and SCLIP.

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 [22, 65], 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. More specifically, it is close to self-distillation [63, 64] 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 [44, 48] or specific domain [13, 52]. The process is normally self-supervised without any additional manual labeling [36, 42].

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. S12 illustrates this process and the metric drops (ave of 8 datasets) sharply after residuals in MaskCLIP, indicating noise injection [21]. In contrast, our FSA better preserves spatial coherence. Fig. S13 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*.

D. Similarity metric

Table S9 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 [26, 27], thus better preserving spatial consistency than MaskCLIP

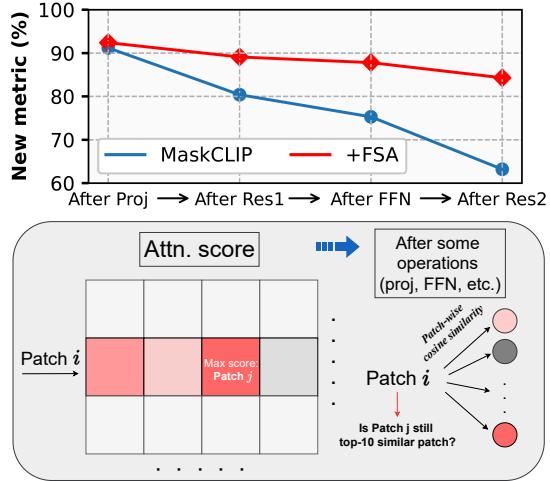


Figure S12. Illustration and analysis for new metric.

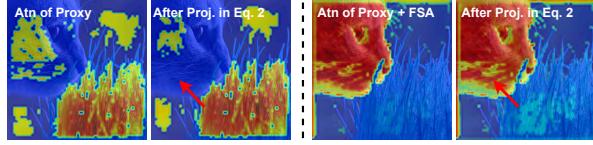


Figure S13. Attention visualization of Fig.2 after Proj. in Eq.2.

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 S9. **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.

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. S10, 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 S10. **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. S11).

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 S11. Speed (ms) on V100 GPU with 224x224 input.

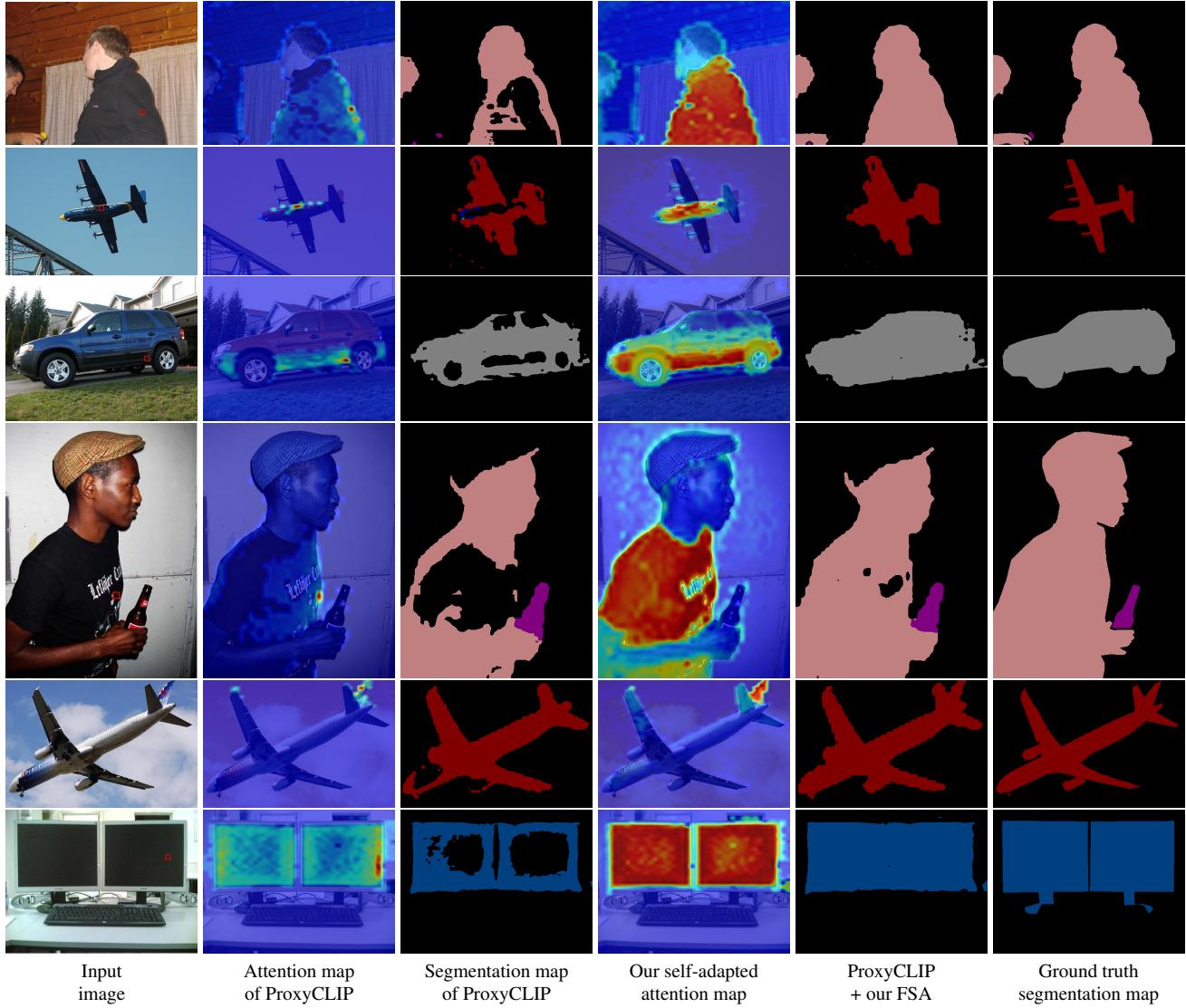


Figure S14. **Comparative visualization of segmentation results: ProxyCLIP vs. integration with our FSA.** The attention maps (second and fourth columns) correspond to the reference patch shown in the first column. ProxyCLIP produces segmentation maps with holes within the same object due to weak attention across regions of the object. In contrast, integrating our FSA effectively aggregates similar patches, enabling the correction of missegmented regions.