UNIVERSITY OF TORONTO

UNIVERSITÉ Concordia UNIVERSITY

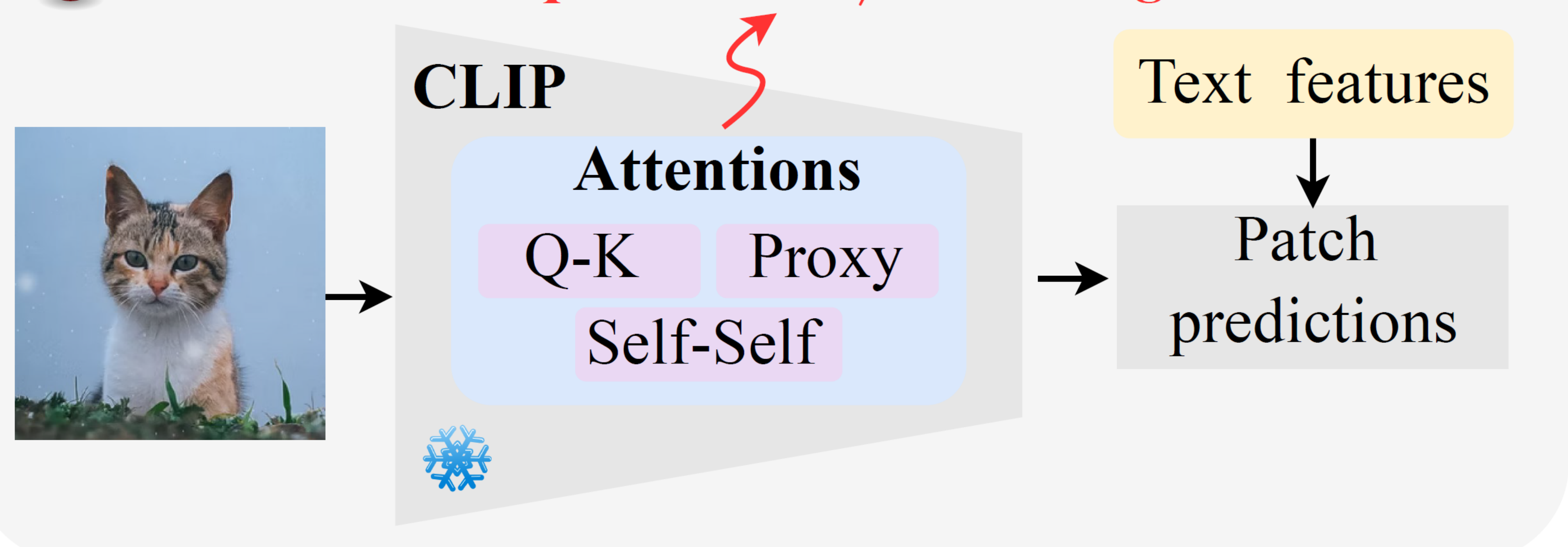
Plug-in Feedback Self-adaptive Attention in CLIP for Training-free Open-Vocabulary Segmentation

Zhixiang Chi, Yanan Wu, Li Gu, Huan Liu, Ziqiang Wang, Yang Zhang, Yang Wang, Konstantinos N. Plataniotis

ICCV OCT 19-23, 2025 HONOLULU HAWAII

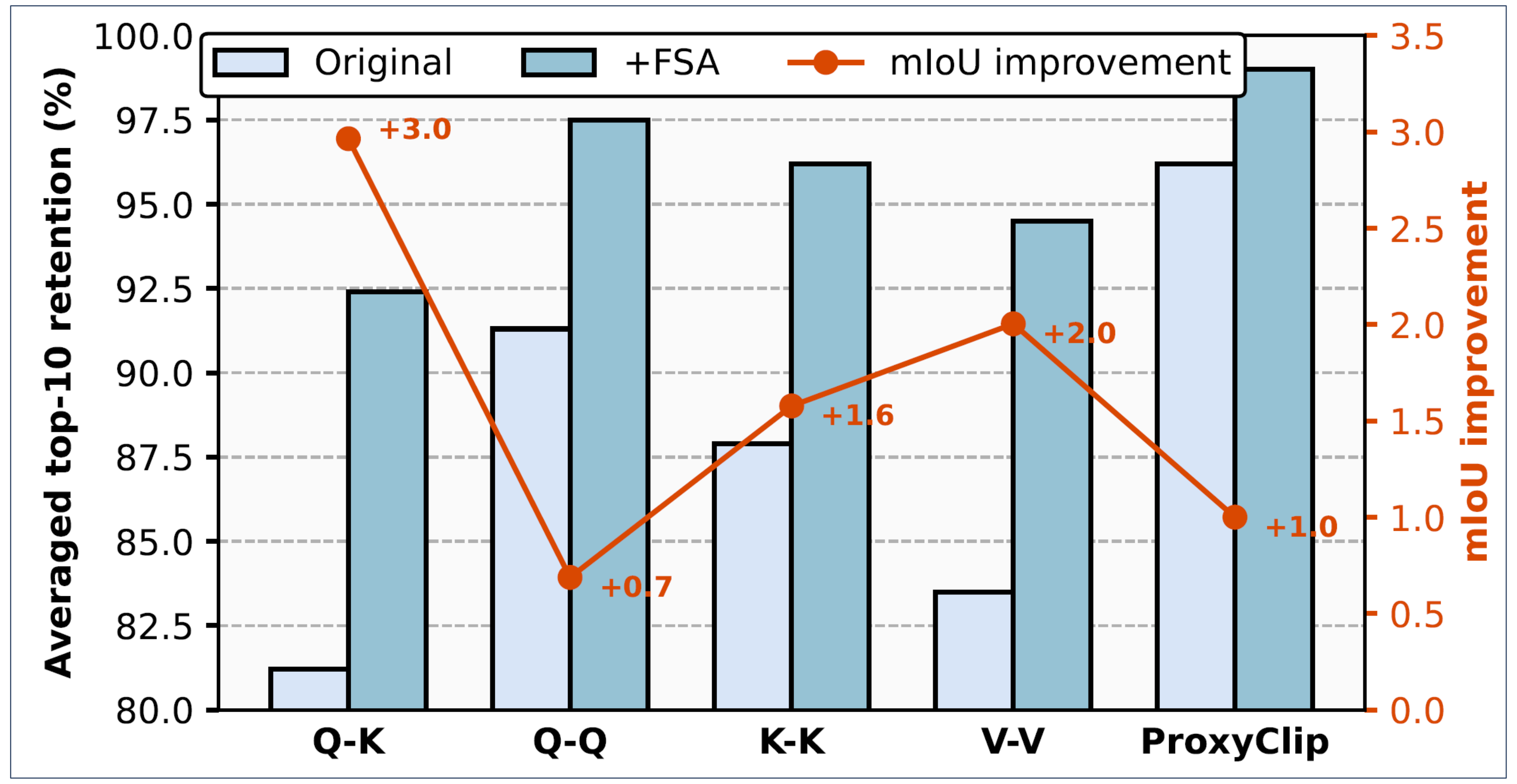
Motivation

! *Intermediate improvement ≠ better segmentation*



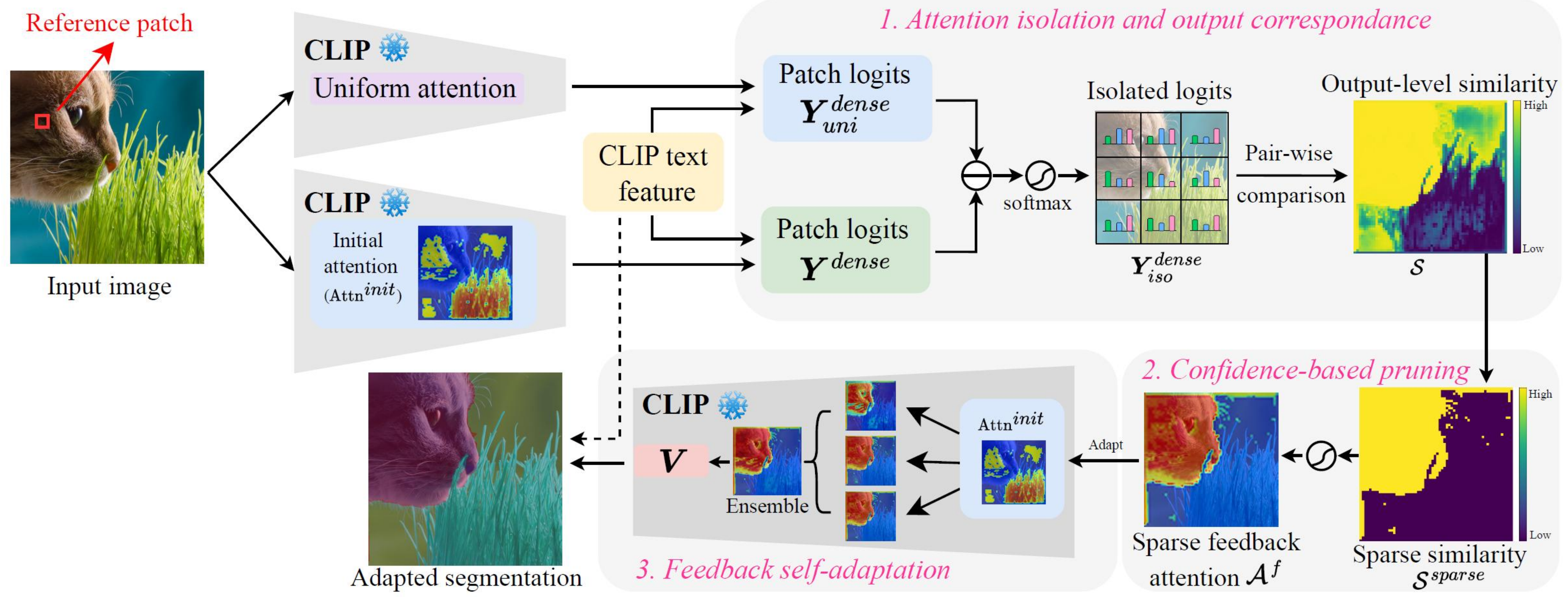
❖ Intermediate attention refinement does not yield improvements in the final segmentation results.

Observation



❖ Intermediate semantic coherence is not preserved in the output, leading to suboptimal segmentation.

Method



- ➢ **(Step 1) Stand-alone intermediate attention isolation:**
  - ➢ **Use an uniform attention to catch the interference of downstream operations.**
  - ➢ **Purify the contribution of intermediate attention.**
- ➢ **(Step 2) Confidence-based sparse attention**
  - ➢ **Suppress irrelevant patches while amplifying semantically coherent ones for the output similarity map.**
- ➢ **(Step 3) Feedback self-adaptive attention**
  - ➢ **Feed the sematic coherence at the output back to intermediate attention maps.**
  - ➢ **Better sematic coherence preservation.**

Our method improves four baselines across three backbone architectures and is validated on 33 intermediate attention configurations.

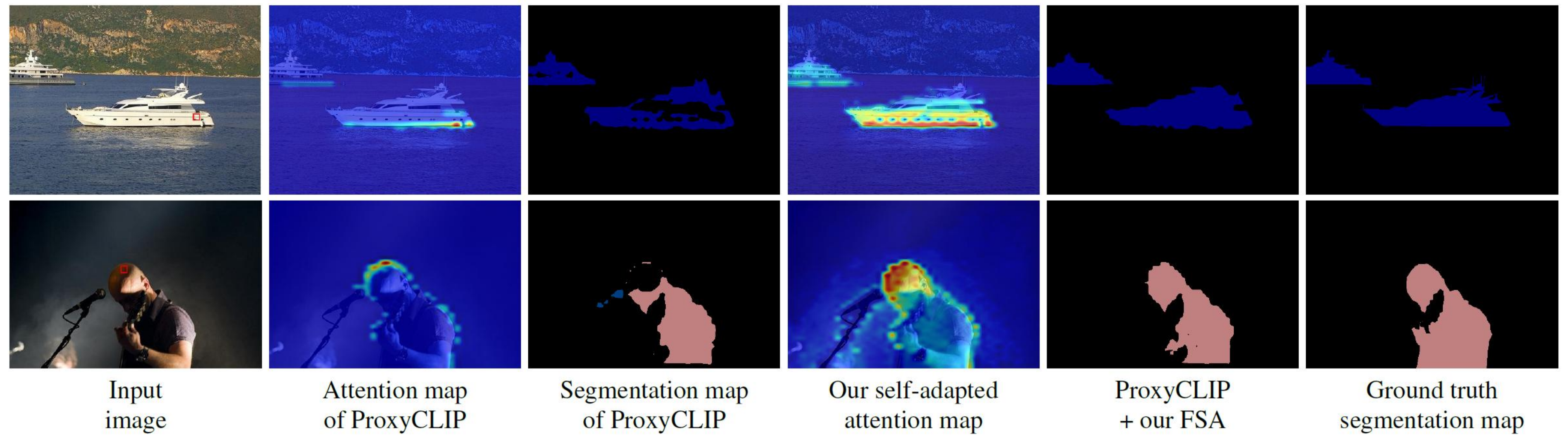
Implementation

Our code is available

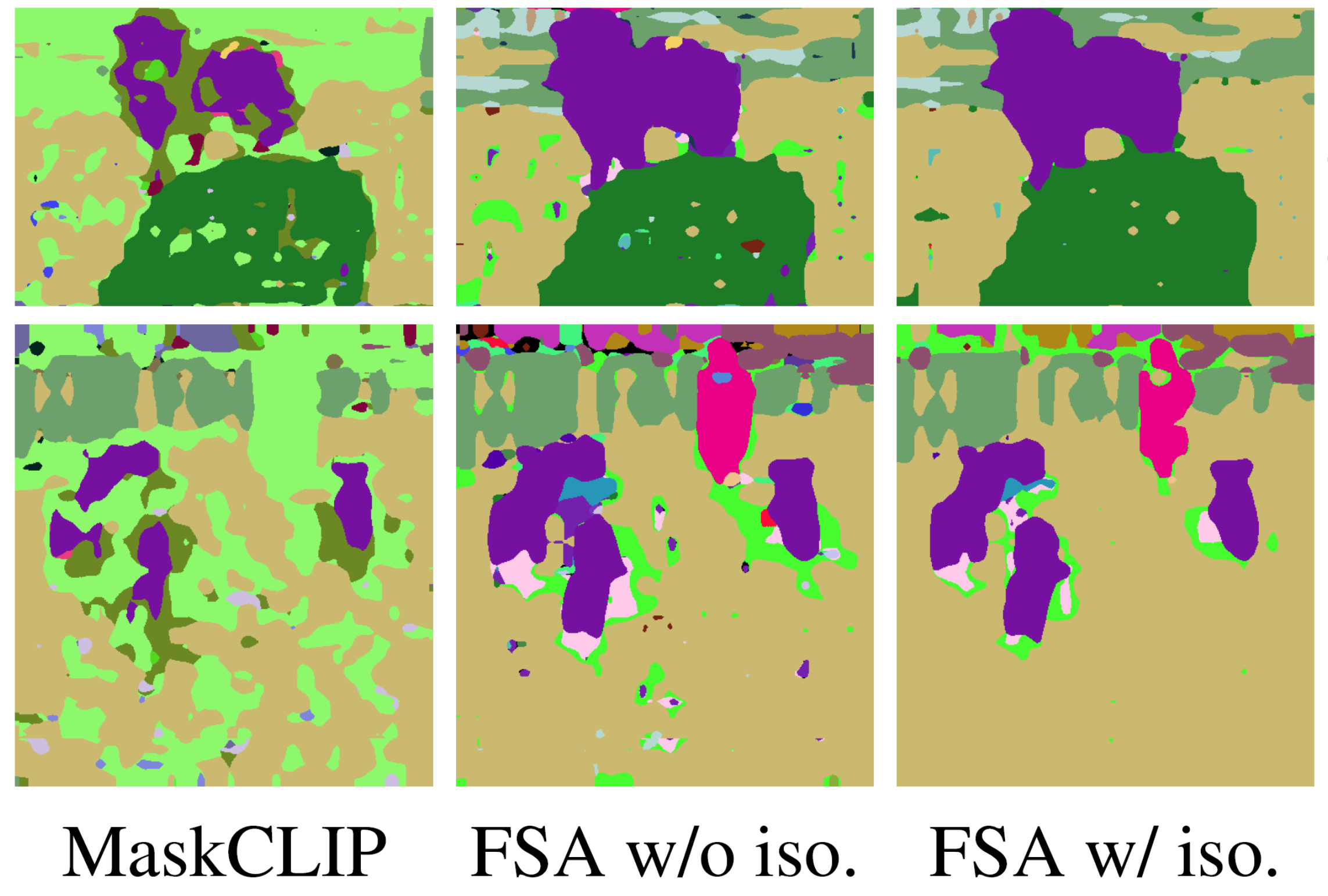
Main results

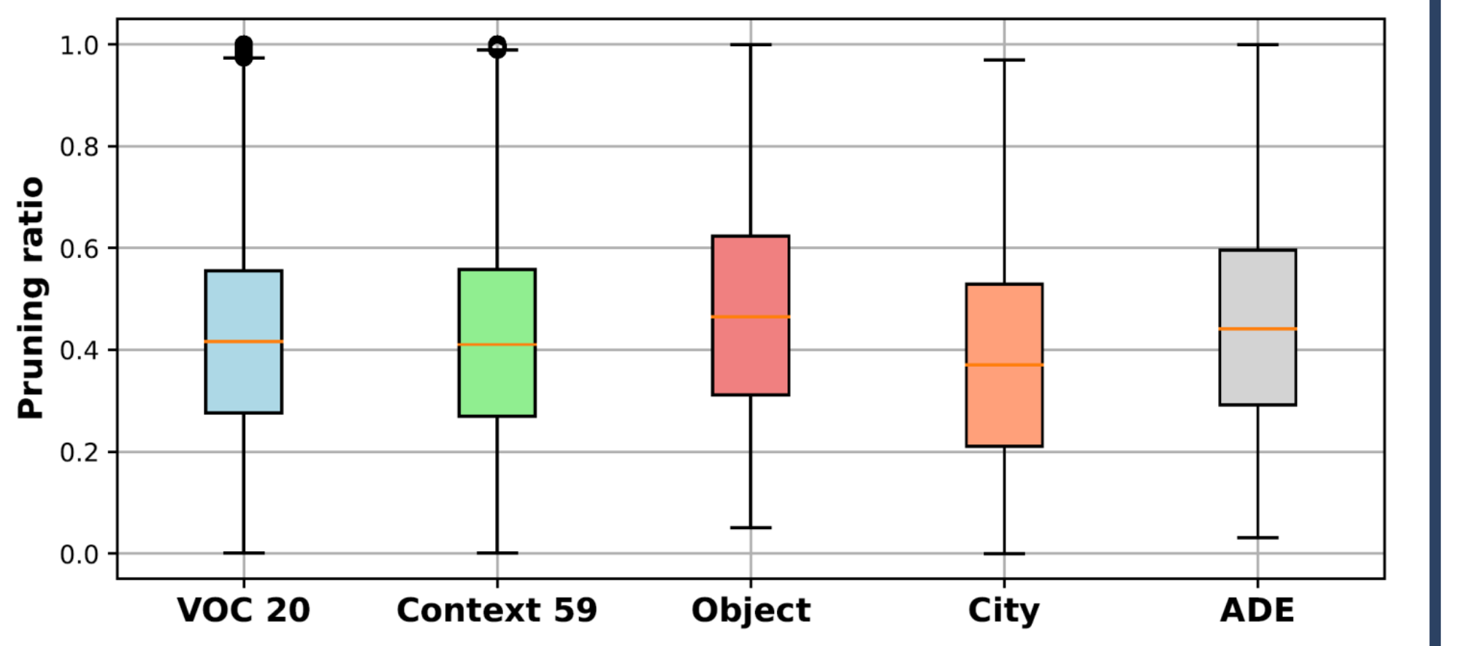
Models	Methods	VOC	Context	Object	VOC20	Context59	Stuff	City	ADE	Avg.
CLIP ViT-B/16	MaskCLIP$_{ECCV'22}$ [67]	38.8	23.6	20.6	74.9	26.4	16.4	12.6	9.8	27.9
	+ FSA	47.7	31.0	29.2	78.3	34.2	21.4	28.4	16.0	35.8 (+7.9)
	SCLIP$_{ECCV'24}$ [46]	59.1	30.4	30.5	80.4	34.2	22.4	32.2	16.1	38.2
	+ FSA	61.5	33.3	33.9	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)
	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)
CLIP ViT-L/14	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)
	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)

Visualization


Input image
Attention map of ProxyCLIP
Segmentation map of ProxyCLIP
Our self-adapted attention map
ProxyCLIP + our FSA
Ground truth segmentation map

⬆ Intermediate attention and final segmentation when integrated into ProxyCLIP.





⬆ Distribution of pruning ratios across datasets, illustrating adaptive pruning.

⬅ Integration with MaskCLIP with and without stand-alone attention isolation.