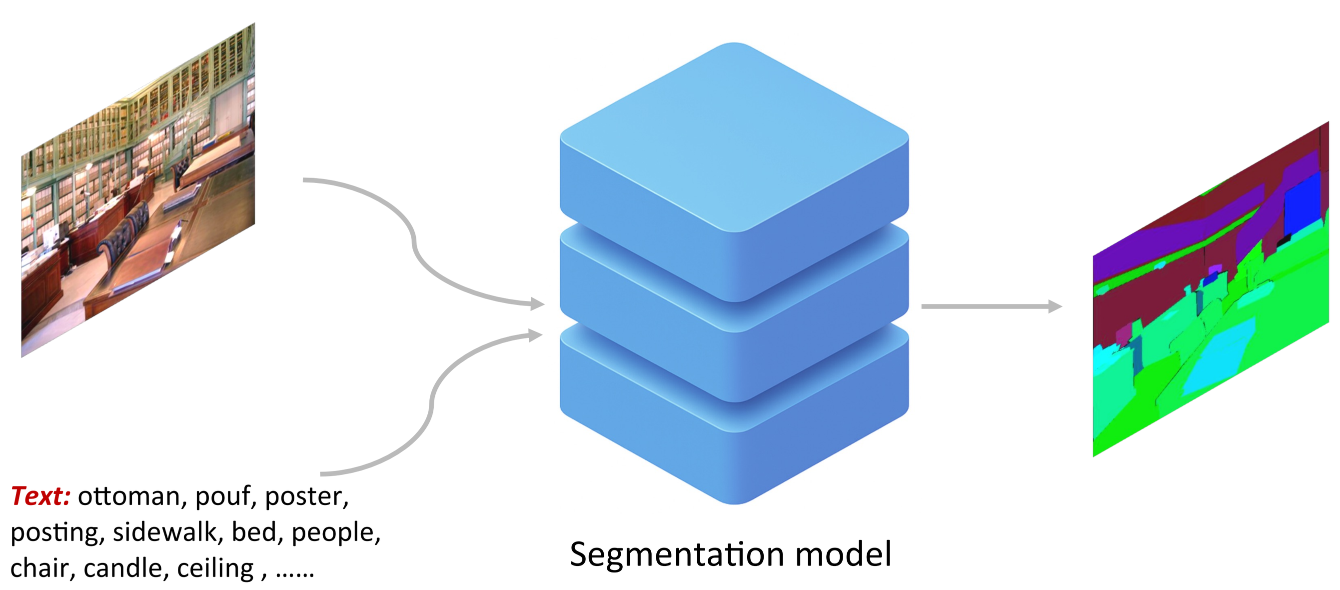
OPMapper: Enhancing Open-Vocabulary Semantic Segmentation with Multi-Guidance Information

Xuehui Wang[1*], Chongjie Si[1*], Xue Yang[2], Yuzhi Zhao[3], Wenhai Wang[4], Xiaokang Yang[1], Wei Shen[1]

[1]MoE Key Lab of Artificial Intelligence, AI Institute, School of Computer Science, SJTU
[2]School of Automation and Intelligent Sensing, SJTU
[3]City University of Hong Kong, [4]MMLab, Chinese University of Hong Kong

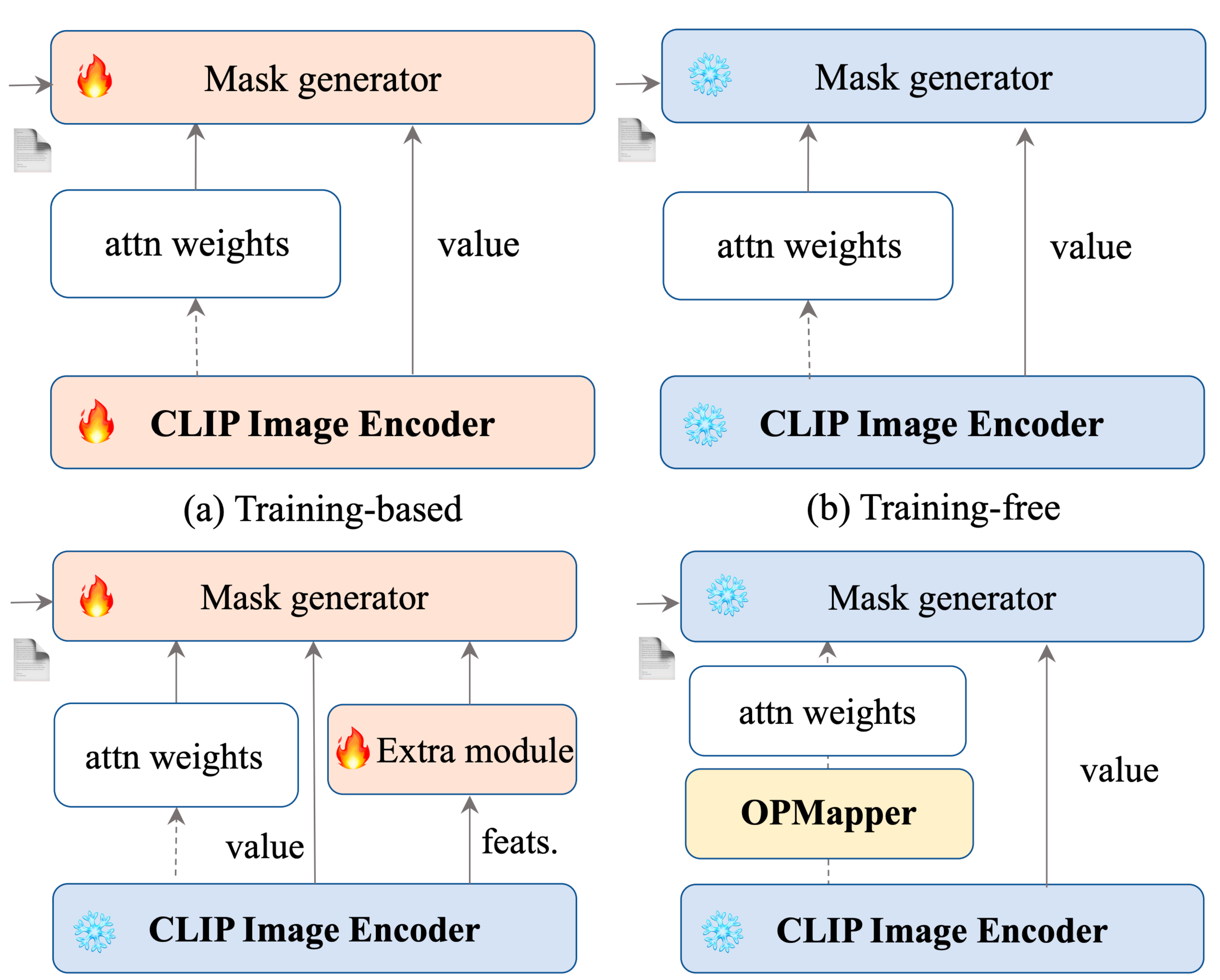
NEURAL INFORMATION PROCESSING SYSTEMS

Open-Vocabulary Semantic Segmentation



Open-vocabulary semantic segmentation aims to assign semantic labels to every pixel in an image, even for categories unseen during training. It leverages vision-language models (e.g., CLIP) to align visual and textual representations in a shared embedding space, enabling recognition beyond a fixed vocabulary.

Different adaptation styles for OVSS



(a) Training-based (b) Training-free (c) Extra trainable module (d) Our adapter plugin

OPMapper has great versatility. It is trained offline and can be applied across (a), (b), and (c), or their hybrid paradigms. Thus, **OPMapper** is a flexible plugin to boost other methods.

Simple motivation

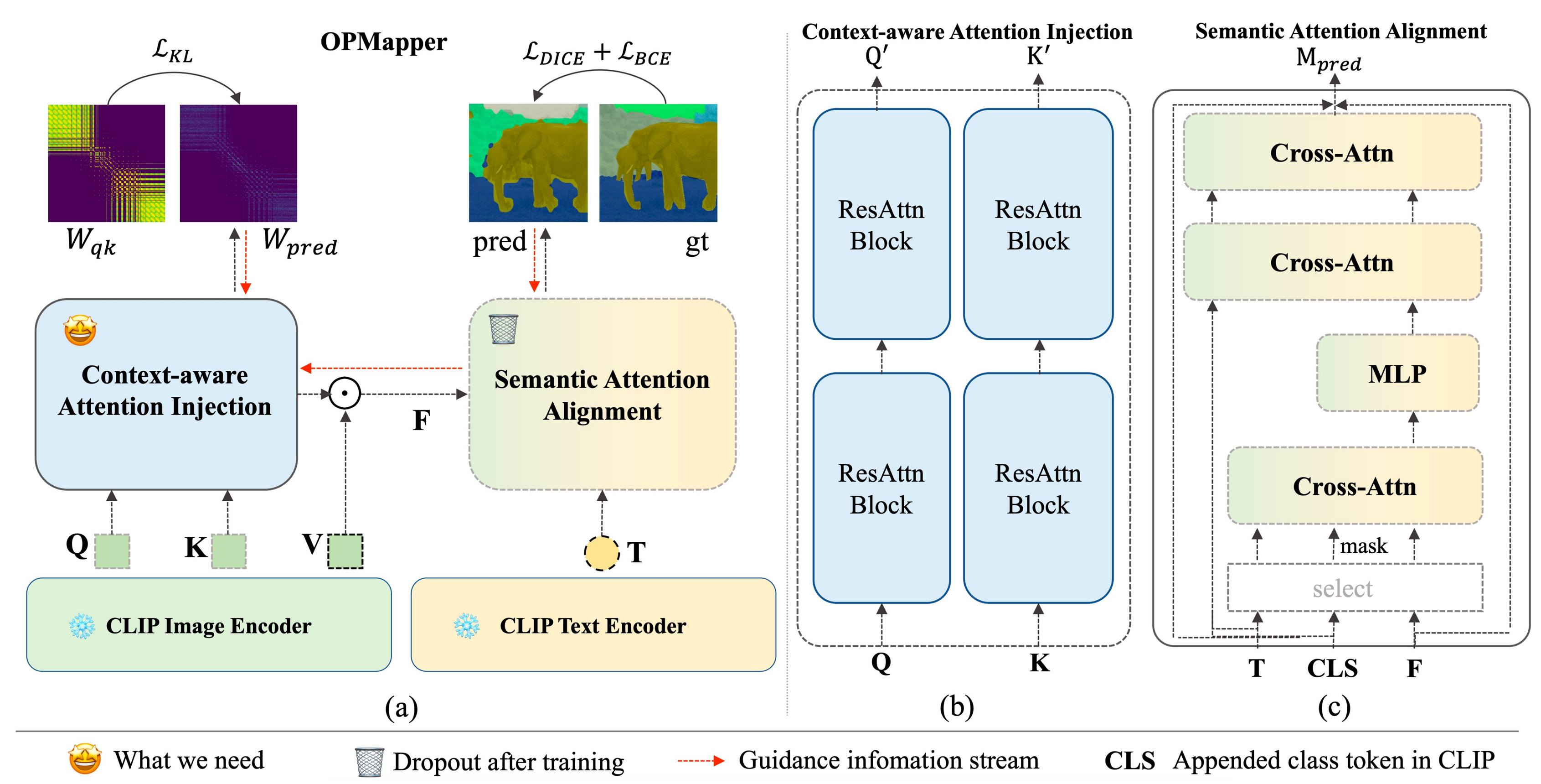
? *Can we avoid redesigning the entire attention pipeline or updating millions of parameters?*

- We directly adopt a mapper to map the object-level query/key into pixel-level query/key, thus obtaining pixel-level attention weights.
- The learning of the mapper should balance *local* compactness and *global connectivity*, while preserving CLIP's inherent vision-language alignment.

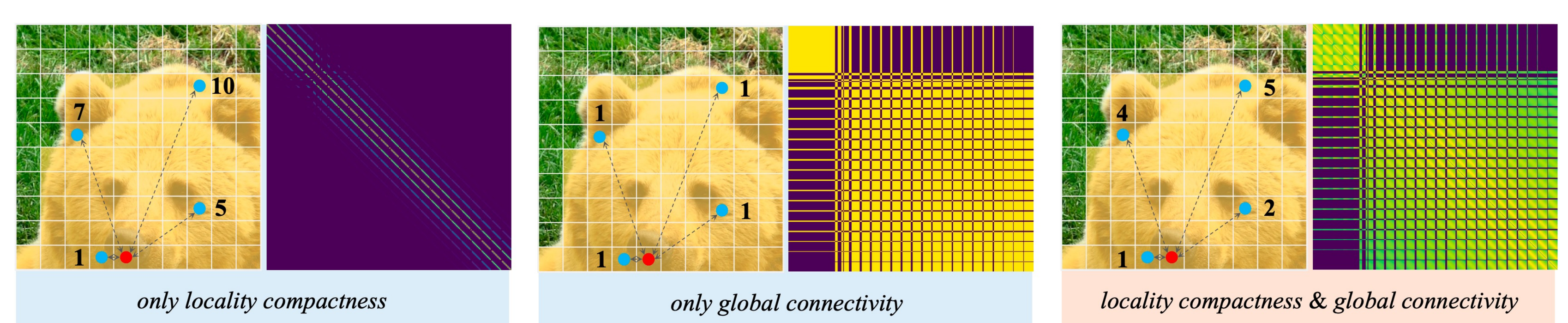
Methodology

We design:

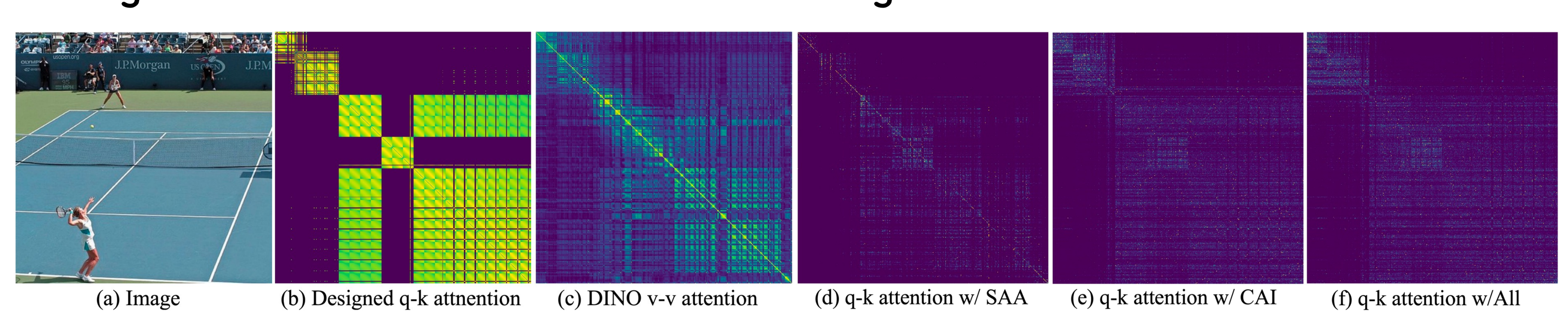
- A lightweight *Mapper* which only contains two block for the input embedding.
- A *prior attention map* that considers *locality compactness* and *global connectivity*.
- A *disposable-after-training module* that further keeps the modality alignment.



How to build the prior attention map?



Designed & DINO & Predicted attention weights

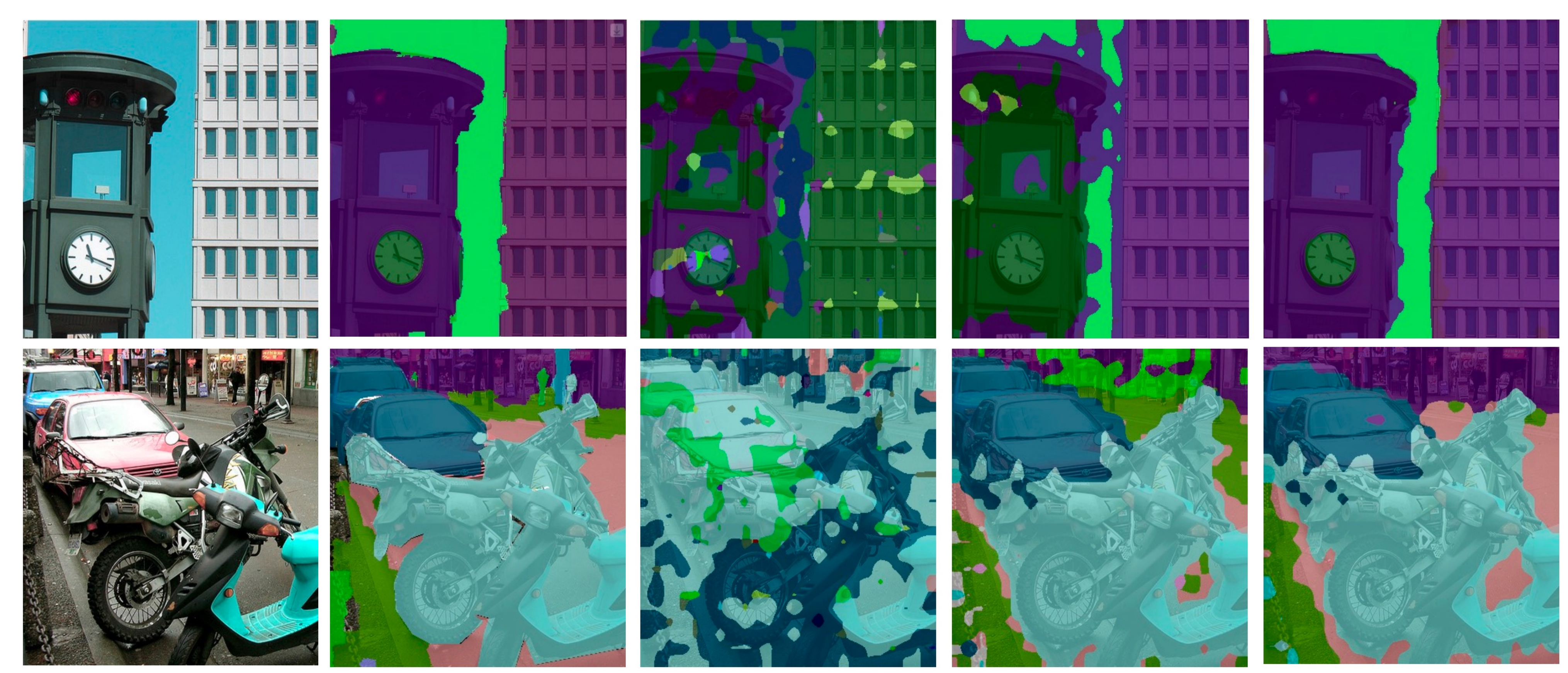


Quantitative comparison

Extensive experiments demonstrate **OPMapper's** effectiveness, yielding significant improvements across 8 open-vocabulary segmentation benchmarks for 9 different methdos.

Model	Image Encoder	VOC-21	VOC-20	Context-60	Context-59	Object-171	Stuff-171	ADE-150	CityScapes	Avg.
▼ *Training based*										
GroupViT [47]*	VIT-B/16	52.3	74.1	22.4	23.4	24.3	-	10.6	-	-
TCL [5]*	VIT-B/16	51.2	77.5	30.3	24.3	30.4	-	14.9	-	-
ZegFormer [14]*	VIT-B/16	65.5	89.5	-	45.5	-	-	18.0	-	-
OVSeg [34]*	VIT-B/16	-	92.6	-	53.3	-	-	24.8	-	-
SAN [49]‡	VIT-B/16	-	92.7	-	51.8	-	40.2	29.9	-	-
+OPMapper‡	VIT-B/16	-	93.2 (+0.5)	-	52.7 (+0.9)	-	41.3 (+1.1)	30.1 (+0.2)	-	-
CAT-Seg [10]‡	VIT-B/16	75.2	93.2	-	55.1	-	43.8	30.7	-	-
+ OPMapper‡	VIT-B/16	74.8(-0.4)	94.0 (+0.8)	-	56.2 (+1.1)	-	43.9 (+0.1)	31.3 (+0.6)	-	-
SCAN [17]‡	VIT-B/16	-	94.9	-	56.8	-	44.1	30.5	-	-
+OPMapper‡	VIT-B/16	-	96.0 (+1.1)	-	58.3 (+1.5)	-	44.8 (+0.7)	31.0 (+0.5)	-	-
▼ *Training free*										
Vanilla CLIP [40]	VIT-B/16	16.4	41.9	8.4	9.2	5.6	4.4	2.9	5.0	11.7
MaskCLIP [15]	VIT-B/16	38.8	74.9	23.6	26.4	20.6	16.4	9.8	12.6	27.89
+OPMapper	VIT-B/16	50.7 (+11.9)	79.1 (+4.2)	32.5 (+8.9)	35.8 (+9.4)	31.8 (+11.2)	23.2 (+6.8)	16.8 (+7.0)	31.5 (+18.9)	37.68 (+9.79)
SCLIP [45]	VIT-B/16	52.5	78.2	30.4	35.5	33.2	23.6	16.8	31.0	37.65
+OPMapper	VIT-B/16	56.2 (+3.7)	84.1 (+5.9)	35.2 (+4.8)	38.8 (+5.6)	36.8 (+3.6)	25.9 (+2.3)	18.3 (+1.5)	33.6 (+2.6)	41.11 (+3.46)
ClearCLIP [27]	VIT-B/16	51.9	80.9	32.4	35.9	33.2	23.9	16.7	30.0	38.11
+OPMapper	VIT-B/16	55.6 (+3.7)	84.7 (+3.8)	34.8 (+2.4)	38.6 (+2.7)	36.5 (+3.3)	25.9 (+2.0)	18.1 (+1.4)	32.9 (+2.9)	40.89 (+2.78)
ProxyCLIP [28]	VIT-B/16	61.3	80.3	35.3	39.1	37.5	26.5	20.2	38.1	42.29
+OPMapper	VIT-B/16	62.8 (+1.5)	84.3 (+4.0)	36.0 (+0.7)	40.1 (+1.0)	38.7 (+1.2)	27.1 (+0.6)	20.3 (+0.1)	37.2 (-0.9)	43.33 (+1.04)
LPOSS [42]	VIT-B/16	60.2	80.2	35.0	36.9	34.7	25.3	21.2	37.6	41.39
+OPMapper	VIT-B/16	63.7 (+3.5)	84.9 (+4.7)	35.7 (+0.7)	37.9 (+1.0)	35.2 (+0.5)	26.4 (+1.1)	22.1 (+0.9)	40.0 (+2.4)	43.24 (+1.85)
CASS [26]	VIT-B/16	64.3	88.3	36.9	39.6	38.1	26.2	20.1	39.8	44.16
+OPMapper	VIT-B/16	67.0 (+2.7)	90.0 (+1.7)	38.2 (+1.3)	40.1 (+0.5)	38.8 (+0.7)	27.1 (+0.9)	20.8 (+0.7)	41.0 (+1.2)	45.37 (+1.21)
Vanilla CLIP [40]	VIT-L/14	8.2	15.6	4.1	4.4	2.7	2.4	1.7	2.5	5.2
CaR [43]†	VIT-L/14&ViT-B/16	67.6	91.4	30.5	39.5	36.6	-	17.7	-	-
MaskCLIP [15]	VIT-L/14	41	65.1	24.5	26.5	26.4	17.6	15.1	21.2	29.68
+OPMapper	VIT-L/14	50.9 (+9.9)	81.9 (+16.8)	30.6 (+6.1)	33.7 (+7.2)	33.5 (+7.1)	22.3 (+4.7)	18.3 (+3.2)	29.7 (+8.5)	37.61 (+7.94)
SCLIP [45]	VIT-L/14	47.4	79.3	27.8	30.6	30.1	20.5	15.6	27.8	34.89
+OPMapper	VIT-L/14	50.4 (+3.0)	81.6 (+2.3)	29.6 (+1.8)	32.6 (+2.0)	33.5 (+3.4)	21.8 (+1.3)	16.5 (+0.9)	30.3 (+2.5)	37.03 (+2.14)
ClearCLIP [27]	VIT-L/14	46.1	80	26.8	29.6	30.1	19.9	15	27.9	34.43
+OPMapper	VIT-L/14	49.2 (+3.1)	81.5 (+1.5)	28.7 (+1.9)	31.7 (+2.1)	33.1 (+3.0)	21.3 (+1.4)	16.1 (+1.1)	29.8 (+1.9)	36.43 (+2.00)
ProxyCLIP [28]	VIT-L/14	59.3	82.6	33.1	35.7	38.2	24.2	20.8	36.3	41.28
+OPMapper	VIT-L/14	59.2(-0.1)	84.4 (+1.8)	33.7(+0.6)	36.9 (+1.2)	39.1 (+0.9)	25.1 (+0.9)	21.7 (+0.9)	37.8 (+1.5)	42.24(+0.96)

Visualization results



(a) Image (b) GroundTruth (c) CLIP (d) ClearCLIP (e) CLIP+OPMapper

Prof. Wei Shen: https://shenwei1231.github.io | Xuehui Wang: https://huiserwang.site | GitHub: https://github.com/SJTU-DeepVisionLab/OPMapper | Email: wangxuehui@sjtu.edu.cn

NeurIPS 2025, San Diego