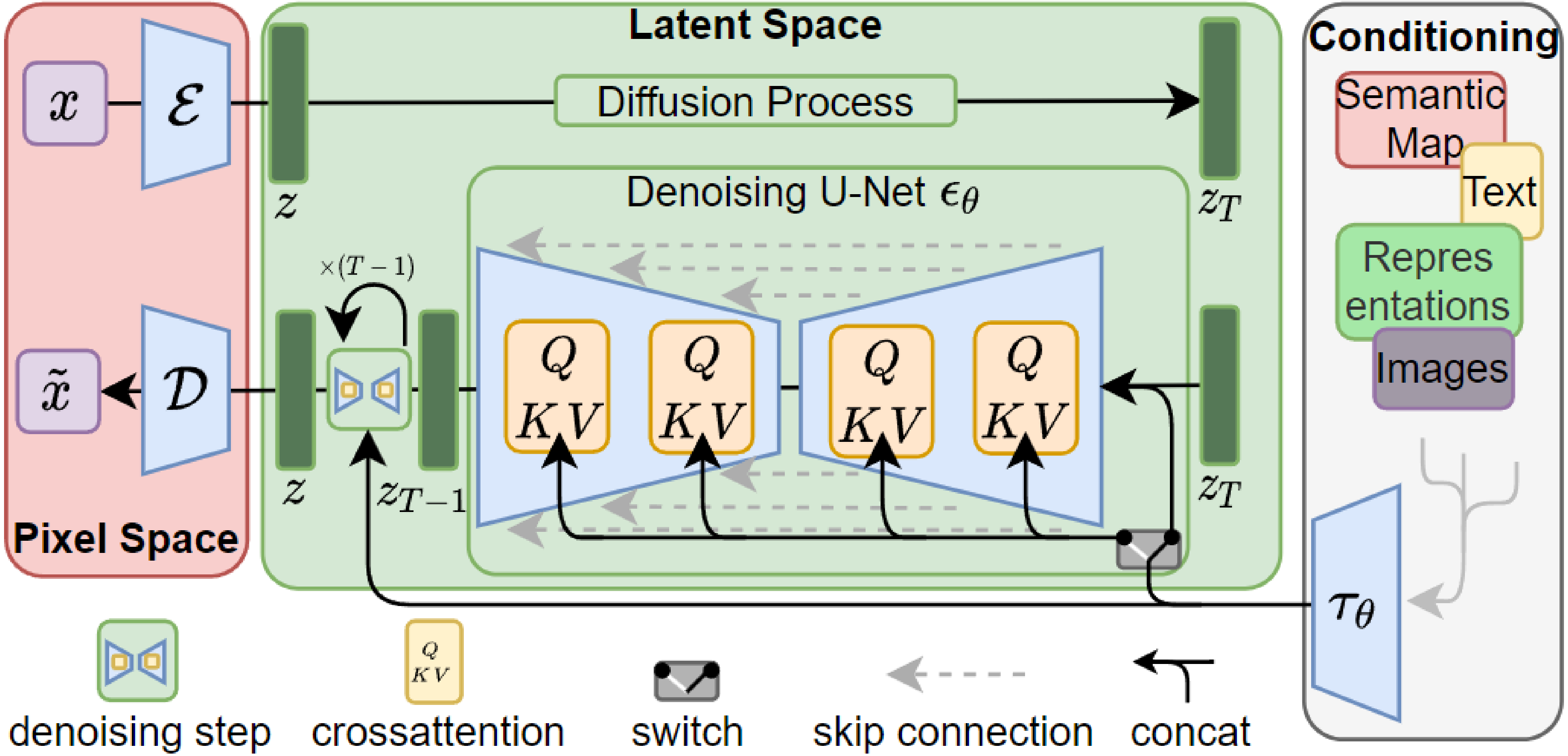
# Seminar

朱顺尧

2025.10.17

# Setting

Diffusion-based Training-free Segmentation



# Diffusion Model is Secretly a Training-free Open Vocabulary Semantic Segmenter

**Jinglong Wang**[1†], **Xiawei Li**[1†], **Jing Zhang**[1*], **Qingyuan Xu**[1]
**Qin Zhou**[1], **Qian Yu**[1], **Lu Sheng**[1], **Dong Xu**[2]
[1]Beihang University
[2]The University of Hong Kong
wjlzy@buaa.edu.cn, zy2121108@buaa.edu.cn, zhang_jing@buaa.edu.cn

# Motivation

Pixel-level labels are expensive

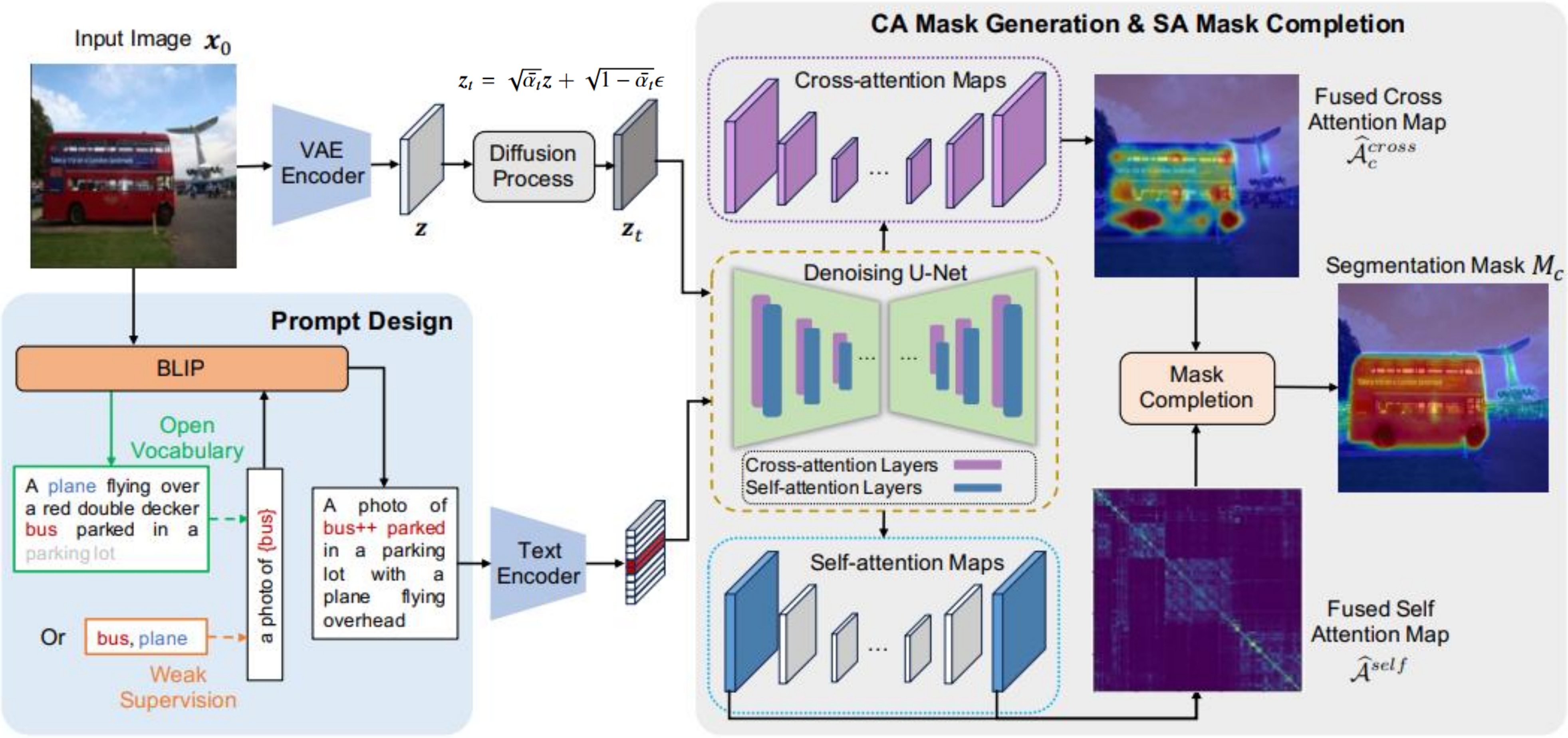
Models trained solely on fully annotated data are restricted to specific categories

Clip-based methods lack of crucial localization information and awareness of object shapes

A higher similarity leads to larger activation values in CAM,  indicating a closer relationship between the current pixel and the corresponding text

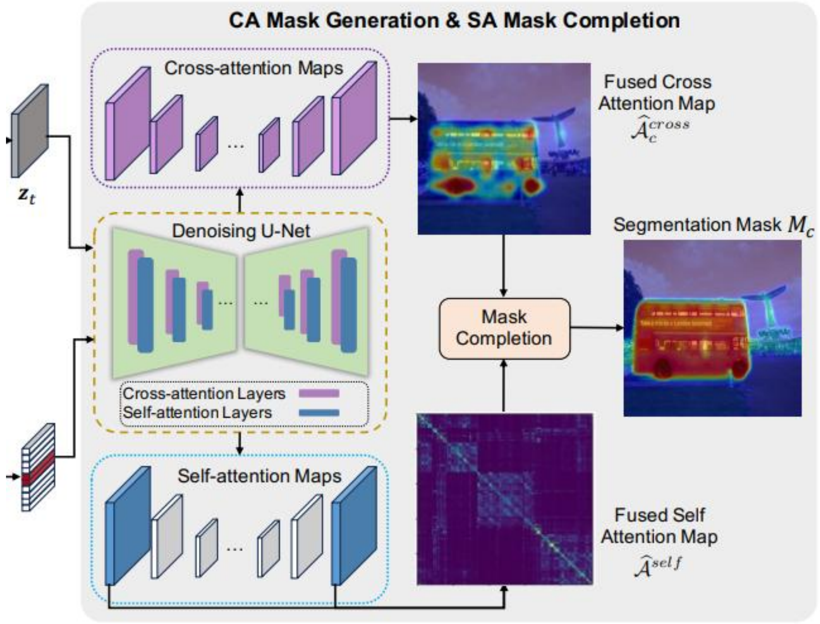
# Method

Overall architecture



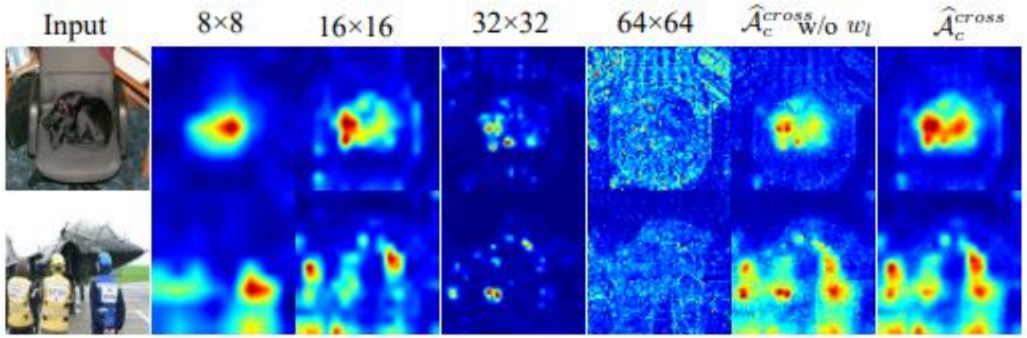
# Method

Cross-attention-based Score Map Generation
&Self-attention-based Score Map Completion




$$\widehat{\mathcal{A}}_c^{cross} = \sum_{l \in L} w_l \cdot \mathcal{A}_{c,l}^{cross} \in \mathbb{R}^{H \times W},$$

$$[0.3,\ 0.5,\ 0.1,\ 0.1]$$



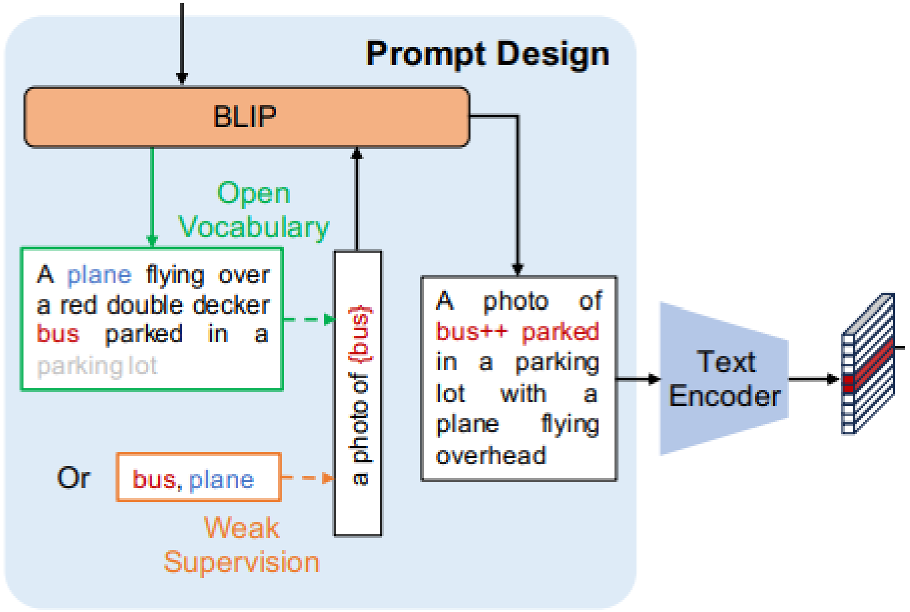

lack clear object boundaries and may exhibit internal holes:
use SAM to perform region completion

$$\widehat{\mathcal{A}}^{self} = \frac{1}{L} \sum_{l \in L} \mathcal{A}_l^{self} \in \mathbb{R}^{HW \times HW},$$

$$M_c = \mathrm{norm}(\widehat{\mathcal{A}}^{self} \cdot vec(\widehat{\mathcal{A}}_c^{cross})),$$

# Method

Prompt Design for Semantic Enhancement

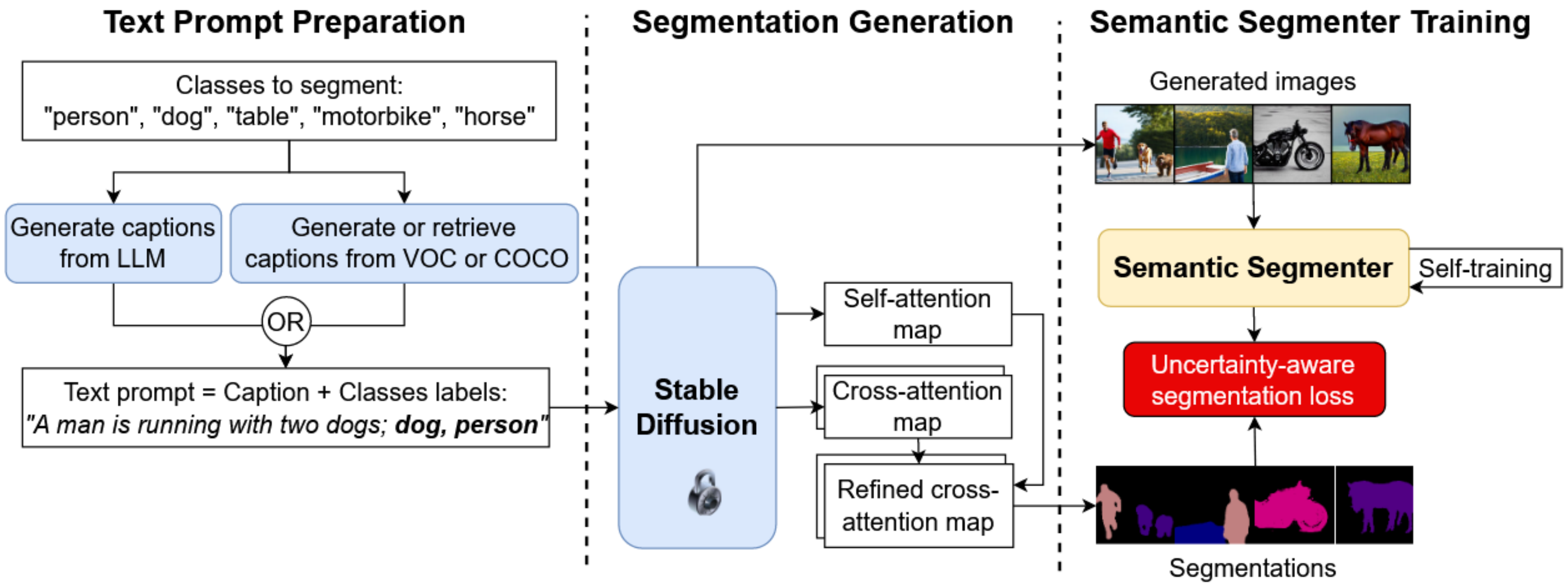



The cross-attention maps of the class names and the adverbs or adjectives are fused to obtain the segmentation score maps.

Class Token Re-weighting

# Method

## Dataset Diffusion



# Results

RESULTS OF ZERO-SHOT OPEN-VOCABULARY SEMANTIC SEGMENTATION ON THREE BENCHMARK DATASETS

| Method | VOC | Context | Object |
|---|---|---|---|
| *Training-involved* | | | |
| ReCo [45] | 25.1 | 19.9 | 15.7 |
| ViL-Seg [46] | 37.3 | 18.9 | - |
| MaskCLIP [23] | 38.8 | 23.6 | 20.6 |
| TCL [47] | 51.2 | 24.3 | 30.4 |
| CLIPpy [48] | 52.2 | - | 32.0 |
| GroupViT [49] | 52.3 | 22.4 | - |
| ViewCo [50] | 52.4 | 23.0 | 23.5 |
| SegCLIP [51] | 52.6 | 24.7 | 26.5 |
| OVSegmentor [25] | 53.8 | 20.4 | 25.1 |
| *Training-free* | | | |
| DiffSeg [17] | 39.4 | 16.7 | 19.1 |
| OVDiff(+CutLER+DINO&CLIP) [15] | **67.1** | **30.1** | 34.8 |
| OVDiff(+DINO&CLIP) [15] | 62.8 | 28.6 | 34.9 |
| OVDiff [15] | 60.4 | 27.6 | - |
| DiffSegmenter (Ours) | 60.1 | 27.5 | **37.9** |

OVDiff necessitates a complex image synthesis process and involve additional pre-trained segmenters and feature extractors for prototype generation

| Method | VOC train |
|---|---|
| *Image-level Supervsion* | |
| IRN [Ahn *et al.*, 2019] | 48.8 |
| SC-CAM [Chang *et al.*, 2020] | 50.9 |
| SEAM [Wang *et al.*, 2020] | 55.4 |
| AdvCAM [Lee *et al.*, 2021b] | 55.6 |
| RIB [Lee *et al.*, 2021a] | 56.5 |
| OoD [Lee *et al.*, 2022] | 59.1 |
| MCTfomer [Xu *et al.*, 2022b] | 61.7 |
| DiffSegmenter (Ours) | **70.5** |
| *Image-level Supervision+Language Supervision* | |
| CLIMS [Xie *et al.*, 2022] | 56.6 |
| CLIP-ES [Lin *et al.*, 2023] | 70.8 |

Table 2: Segmentation results of on PASCAL VOC 2012 train sets with image-level object labels.

| Method | Backbone | Val | Test |
|---|---|---|---|
| *Image-level Supervsion* | | | |
| AdvCAM [Lee *et al.*, 2021b] | R101 | 68.1 | 68.0 |
| RIB [Lee *et al.*, 2021a] | R101 | 68.3 | **69.1** |
| ReCAM [Chen *et al.*, 2022] | R101 | 68.5 | 68.4 |
| DiffSegmenter (Ours) | R101 | **69.1** | **68.6** |
| *Image-level Supervision+Language Supervision* | | | |
| CLIMS [Xie *et al.*, 2022] | R101 | 69.3 | 68.7 |
| CLIP-ES [Lin *et al.*, 2023] | R101 | 71.1 | 71.4 |

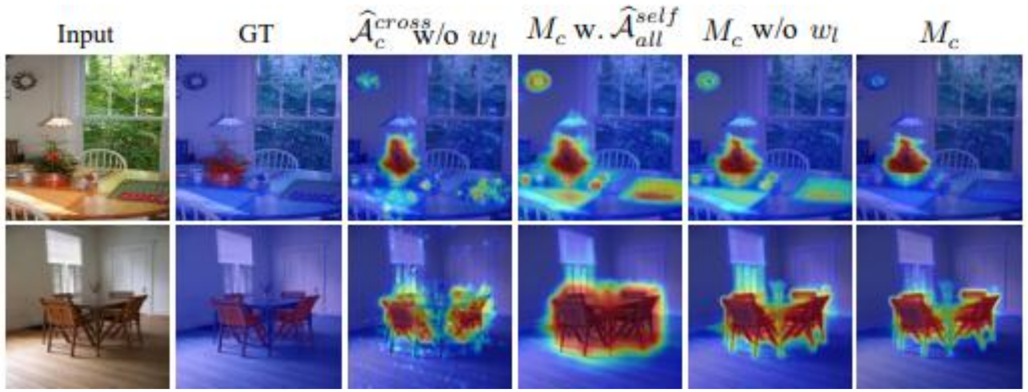
Table 3: Weakly-supervised semantic segmentation results on PASCAL VOC 2012 validation and test sets.

# Ablation

| Method | | | | | VOC train |
|---|---|---|---|---|---|
| $\widehat{\mathcal{A}}_c^{cross}$ | $\widehat{\mathcal{A}}_{all}^{self}$ | $\widehat{\mathcal{A}}^{self}$ | BLIP | "++" | **mIoU** |
| w/o $w_l$ | | | ✓ | ✓ | 61.25 |
| w/o $w_l$ | ✓ | | ✓ | ✓ | 65.01 |
| w/o $w_l$ | | ✓ | ✓ | ✓ | 67.89 |
| ✓ | | ✓ | | | 65.32 |
| ✓ | | ✓ | ✓ | | 67.99 |
| ✓ | | ✓ | | ✓ | 69.46 |
| ✓ | | ✓ | ✓ | ✓ | **70.49** |

| **Method** | t=1 | t=50 | t=100 | t=150 | **Avg.** |
|---|---|---|---|---|---|
| **mIoU** | 69.10 | 69.94 | 70.30 | 69.69 | 70.49 |

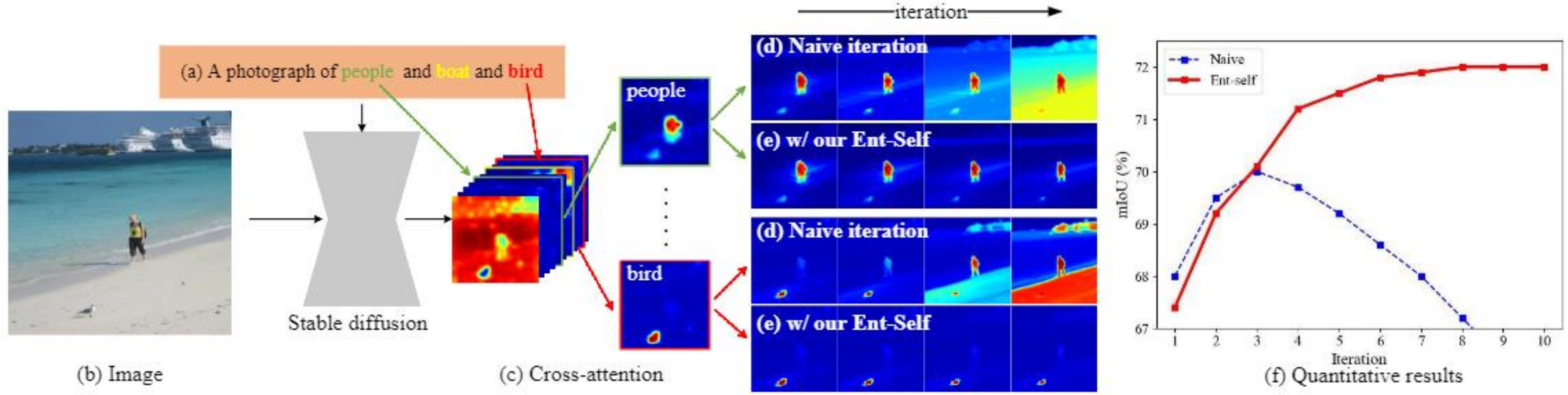
Table 5: Results of different timesteps. **Avg.** is calculated by averaging the results of t=1,t=50,t=100 and t=150.

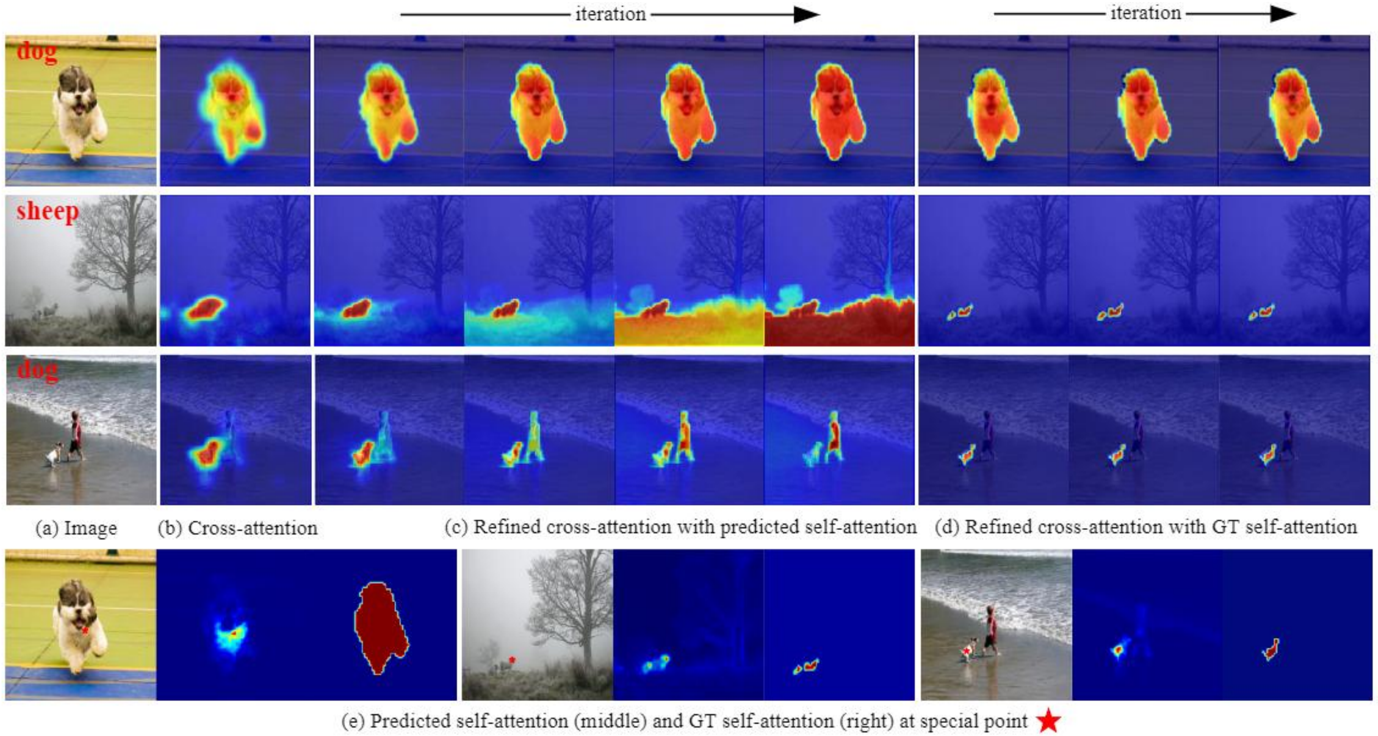


# iSeg: An Iterative Refinement-based Framework for Training-free Segmentation

Lin Sun, Jiale Cao, Jin Xie, Fahad Shahbaz Khan, *Senior Member, IEEE*,
and Yanwei Pang, *Senior Member, IEEE*

# Motivation

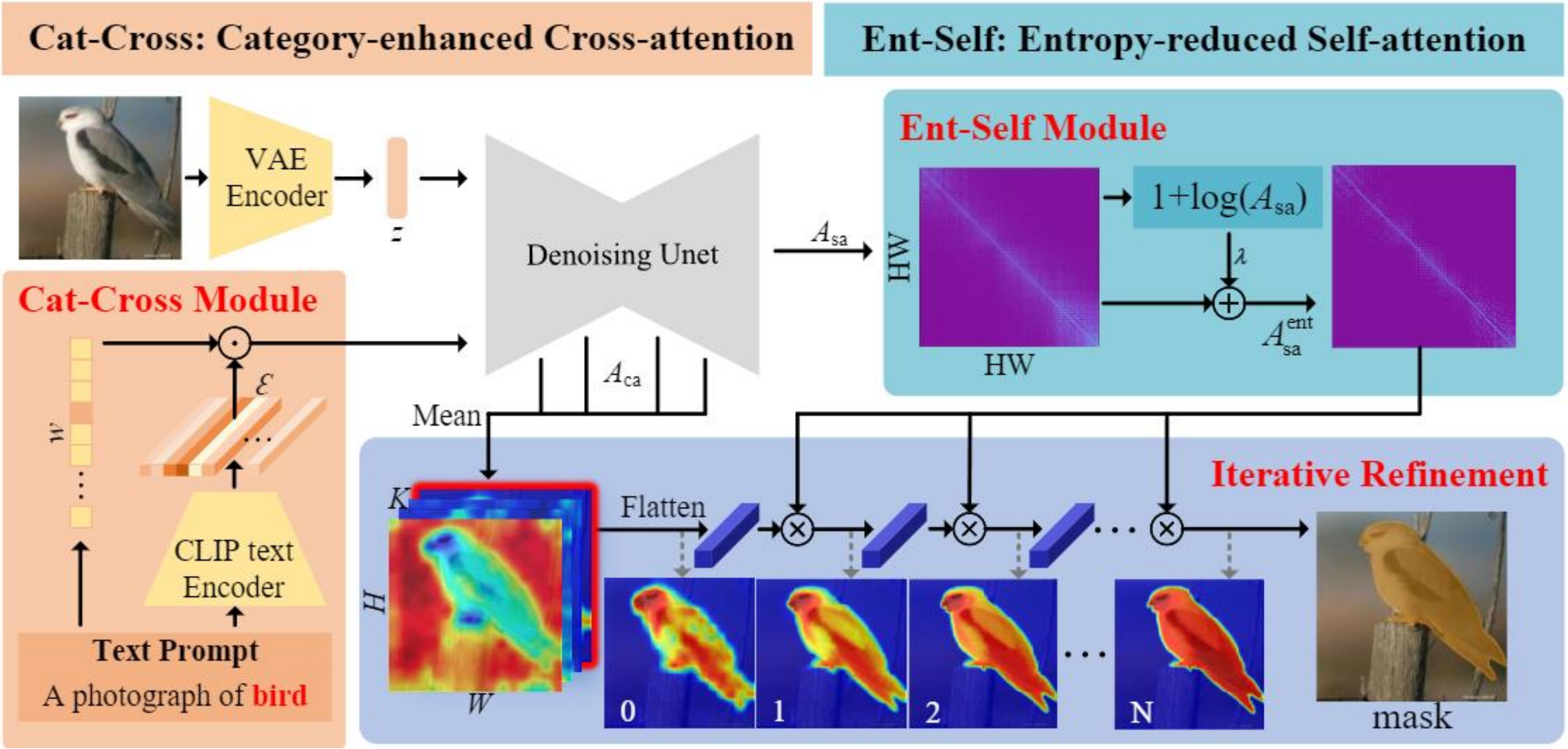



Naive use of self-attn map to iteratively refine CAM may aggregates global information from irrelevant regions

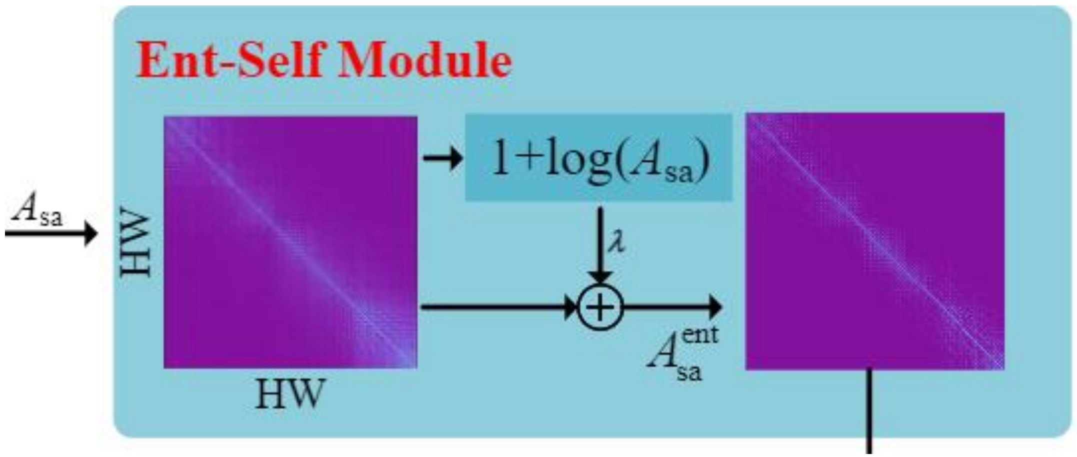
# Method

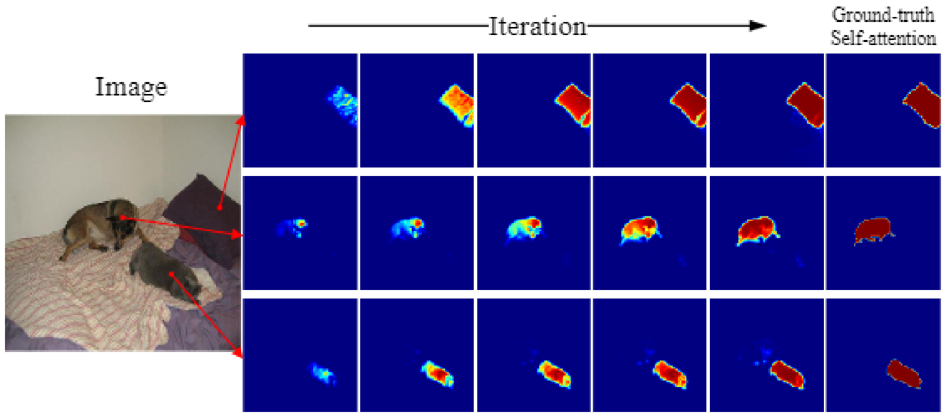
Overall architecture



# Method

Entropy-reduced self-attention



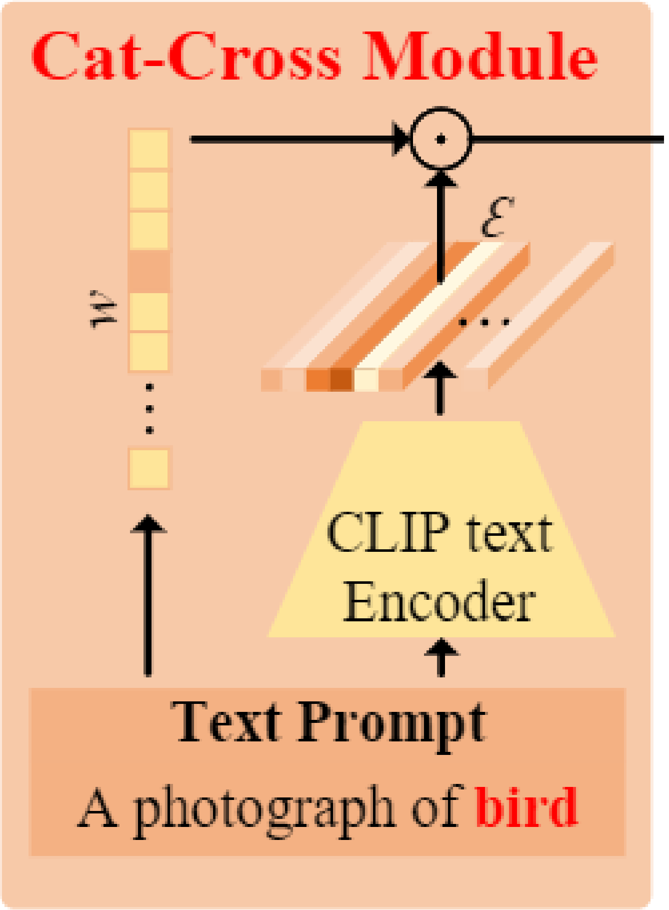

$$E = -\sum_{i=1}^{HW}\sum_{j=1}^{HW} A_{\mathrm{sa}}[i,j]\log(A_{\mathrm{sa}}[i,j]).$$

$$\frac{\mathrm{d}E}{\mathrm{d}A_{\mathrm{sa}}^{ij}} = -(1+\log(A_{\mathrm{sa}}^{ij})).$$

$$A_{\mathrm{sa}}^{ij} = A_{\mathrm{sa}}^{ij} + \lambda(1+\log(A_{\mathrm{sa}}^{ij})),$$

# Method

Category-enhanced cross-attention




$$W[j] = \begin{cases} \gamma, & \text{if } j \in \mathcal{C}, \\ 1, & \text{if } j \notin \mathcal{C}, \end{cases}$$

$$A_{\mathrm{ca}} = \mathrm{Softmax}\left(\frac{Q(W \cdot K)^{\mathrm{T}}}{\sqrt{d}}\right),$$

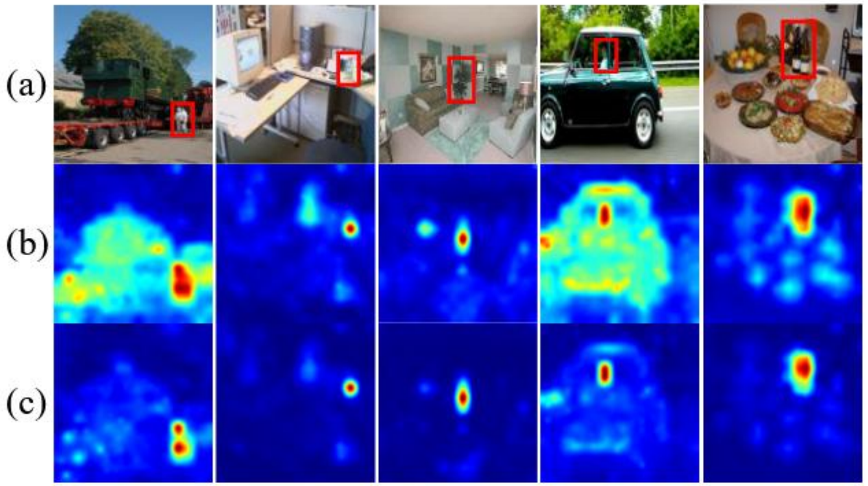



Fig. 5. **Comparison of cross-attention maps** before and after Cat-Cross module. Compared to the original cross-attention map (b), the refined cross-attention map (c) is more clean, and has strong response around corresponding objects in red bounding-box.

# Results

TABLE 1

**Comparison of pseudo mask generation with weakly-supervised semantic segmentation approaches.** We report the mIoU results on PASCAL VOC 2012 and MS COCO training sets. Our proposed method outperforms various training-based and training-free approaches.

| Type | Method | Publication | Training | VOC | COCO |
|---|---|---|---|---|---|
| CNN-based | IRN [1] | CVPR2019 | ✓ | 66.5 | 42.4 |
| | AdvCAM [30] | CVPR2021 | ✓ | 55.6 | 35.8 |
| | BAS [77] | IJCV2023 | ✓ | 57.7 | 36.9 |
| | HSC [67] | IJCAI2023 | ✓ | 71.8 | - |
| Transformer-based | MCTformer [74] | CVPR2022 | ✓ | 61.7 | - |
| | MCTformer+ [73] | arXiv2023 | ✓ | 68.8 | - |
| | ToCo [55] | CVPR2023 | ✓ | 72.2 | - |
| | WeakTr [81] | arXiv2023 | ✓ | 66.2 | - |
| CLIP-based | CLIMS [70] | CVPR2022 | ✓ | 56.6 | - |
| | CLIP-ES [37] | CVPR2023 | ✗ | 70.8 | 39.7 |
| Diffusion-based | DiffSegmenter [64] | arXiv2023 | ✗ | 70.5 | - |
| | T2M [68] | arXiv2023 | ✗ | 72.7 | 43.7 |
| | **iSeg (Ours)** | - | ✗ | **75.2** | **45.5** |

TABLE 2

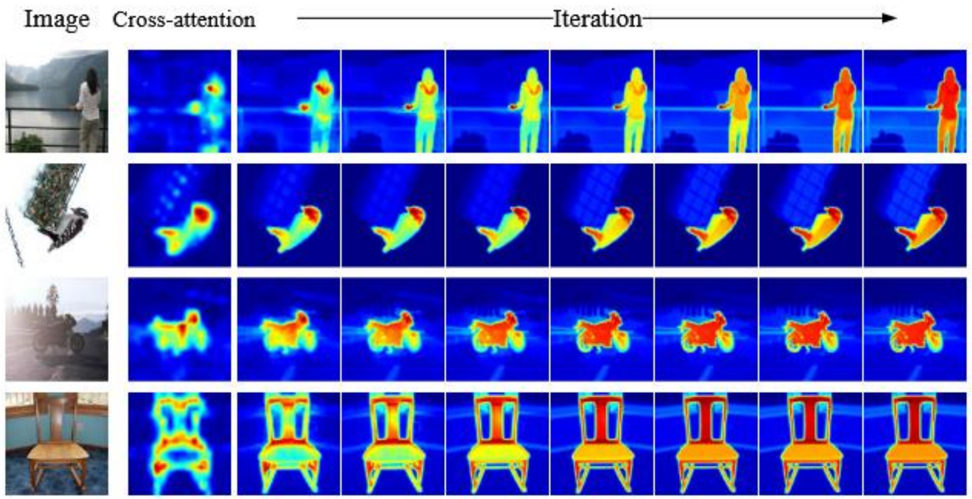
**Comparison with open-vocabulary segmentation approaches.** We reports the mIoU results on PASCAL VOC 2012 validation set, PASCAL-VOC Context validation set, and MS COCO-Object validation set. Our proposed method achieves the promising performance.

| Type | Method | Publication | Training | VOC | Context | Object |
|---|---|---|---|---|---|---|
| CLIP-based | ReCo [58] | NeurIPS2022 | ✓ | 25.1 | 19.9 | 15.7 |
| | MaskCLIP [80] | ECCV2022 | ✓ | 38.8 | 23.6 | 20.6 |
| | SegCLIP [41] | ICML2023 | ✓ | 52.6 | 24.7 | 26.5 |
| | CLIPpy [51] | ICCV2023 | ✓ | 52.2 | - | 32.0 |
| | ViewCo [52] | ICLR2023 | ✓ | 52.4 | 23.0 | 23.5 |
| | OVSegmenter [72] | CVPR2023 | ✓ | 53.8 | 20.4 | 25.1 |
| | TCL [8] | CVPR2023 | ✓ | 51.2 | 24.3 | 30.4 |
| | TagCLIP [38] | AAAI2024 | ✗ | 64.8 | - | - |
| | CaR [59] | CVPR2024 | ✗ | 67.6 | 30.5 | 36.6 |
| SAM-based | SAM-CLIP [62] | CVPRW2024 | ✓ | 60.6 | 29.2 | 31.5 |
| Diffusion-based | OVDiff [26] | ECCV2024 | ✗ | 67.1 | 30.1 | 34.8 |
| | DiffSegmenter [64] | arXiv2023 | ✗ | 60.1 | 27.5 | 37.9 |
| | iSeg (Ours) | - | ✗ | **68.2** | **30.9** | **38.4** |

TABLE 3

**Comparison with some unsupervised semantic segmentation approaches.** We report the results on Cityscapes and COCO-Stuff-27 validation sets. Our iSeg stably outperforms DiffSeg and other approaches on these two datasets in terms of mIoU and ACC.

| Method | Publication | Training | Cityscapes | | COCO-Stuff-27 | |
|---|---|---|---|---|---|---|
| | | | ACC | mIoU | ACC | mIoU |
| MDC [7] | ECCV2018 | ✓ | 40.7 | 7.1 | 32.3 | 9.8 |
| IIC [23] | ICCV2019 | ✓ | 47.9 | 6.4 | 21.8 | 6.7 |
| PICLE [14] | CVPR2021 | ✓ | 65.5 | 12.3 | 48.1 | 13.8 |
| STEGO [42] | ICLR2022 | ✓ | 73.2 | 21.0 | 56.9 | 28.2 |
| MaskCLIP [80] | ECCV2022 | ✓ | 35.9 | 10.0 | 32.2 | 19.6 |
| RoCo [58] | NeurIPS2022 | ✓ | 74.6 | 19.3 | 46.1 | 26.3 |
| ACSeg [33] | CVPR2023 | ✓ | - | - | - | 28.1 |
| DiffSeg [60] | CVPR2024 | ✗ | 76.0 | 21.2 | 72.5 | 43.6 |
| **iSeg (Ours)** | - | ✗ | **78.7** | **25.0** | **74.5** | **45.2** |



# Ablation

| Ent-Self | Cat-Cross | Weakly-supervised VOC | Weakly-supervised COCO | Open-vocabulary VOC | Open-vocabulary Context | Open-vocabulary Object | Unsupervised Cityscapes | Unsupervised COCO-Stuff |
|---|---|---|---|---|---|---|---|---|
| ✗ | ✗ | 68.2 | 40.1 | 63.7 | 26.4 | 36.6 | 22.8 | 44.4 |
| ✓ | ✗ | 72.0 | 42.5 | 67.1 | 28.2 | 37.5 | 25.0 | 45.2 |
| ✓ | ✓ | **75.2** | **45.5** | **68.2** | **30.9** | **38.4** | N/A | N/A |

(a) Iteration

| $N$ | 1 | 2 | 4 | 6 | 8 | 10 | 12 |
|---|---|---|---|---|---|---|---|
| mIoU | 71.0 | 72.9 | 74.5 | 75.0 | 75.1 | **75.2** | 74.9 |

(b) Updating factor

| $\lambda$ | 0 | 0.001 | 0.005 | 0.01 | 0.05 | 0.1 |
|---|---|---|---|---|---|---|
| mIoU | 69.1 | 74.3 | 75.0 | **75.2** | 74.6 | 74.1 |

(c) Weighting factor

| $\gamma$ | 1 | 1.2 | 1.4 | 1.6 | 1.8 | 2 |
|---|---|---|---|---|---|---|
| mIoU | 72.0 | 73.6 | 74.7 | **75.2** | 75.2 | 74.9 |

(a) Cross-attention map

| Level | 16×16 | 32×32 | Both |
|---|---|---|---|
| mIoU | 74.8 | 56.9 | **75.2** |

(b) Self-attention map

| Layer | #-3 | #-2 | #-1 |
|---|---|---|---|
| mIoU | 68.5 | 71.1 | **75.2** |

(c) Time-step

| Number | 1 | 50 | 100 | 150 | 200 |
|---|---|---|---|---|---|
| mIoU | 73.2 | 74.6 | **75.2** | 74.5 | 74.3 |