SAM-CP: Marrying SAM with Composable Prompts for Versatile Segmentation

Pengfei Chen[1,2], Lingxi Xie[2], Xinyue Huo[2], Xuehui Yu[1], Xiaopeng Zhang[2], Yingfei Sun[1], Zhenjun Han[1 †], Qi Tian[2]

[1] University of Chinese Academy of Sciences [2] Huawei Inc.

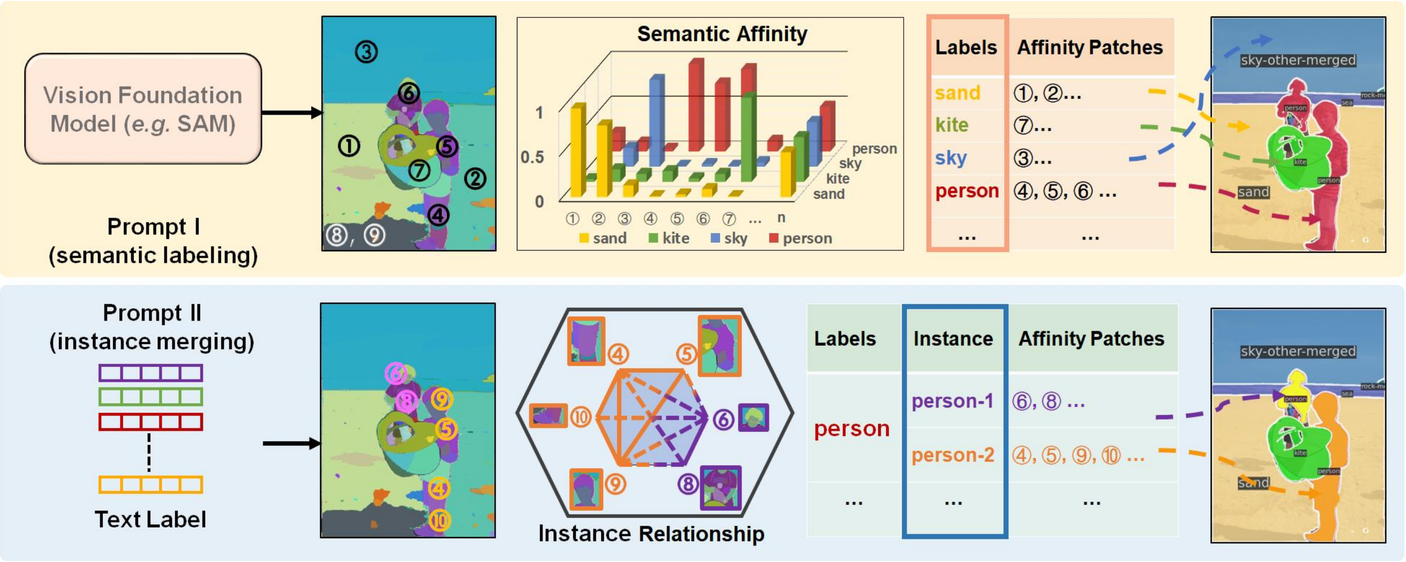
chenpengfei20@mails.ucas.ac.cn 198808xc@gmail.com hanzhj@ucas.ac.cn

https://github.com/ucas-vg/SAM-CP

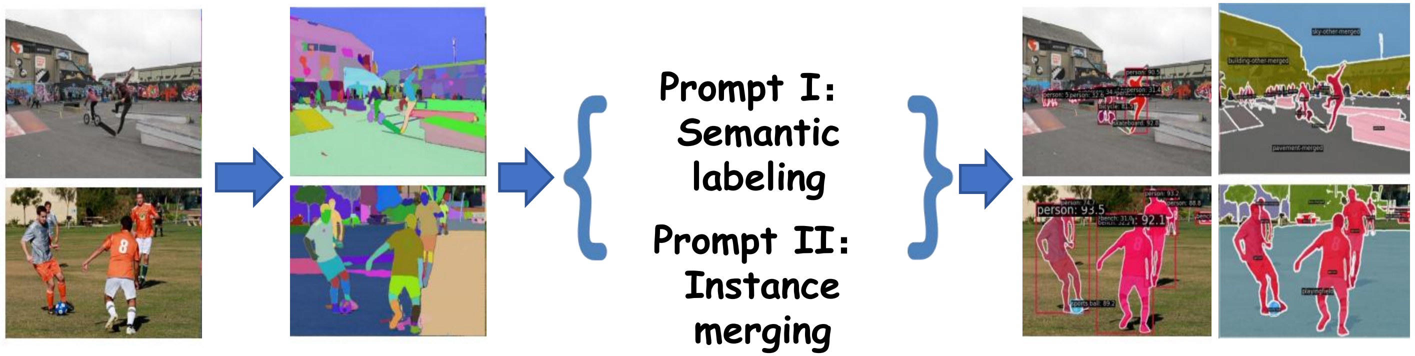
Introduction

Segment everything by grid points prompting, however:

- over-segmentation
- acking semantic labelling ability



Two composable prompts for versatile segmentation:
Prompt I – semantic labeling: whether a SAM patch aligns with the text label
Prompt II – instance merging: whether two patches belong to the same instance of the corresponding category



A new bottom-up visual sensing style

Methods

A unified affinity framework:

Segment patches P = $\{P_1, P_2, \ldots, P_N\}$ with SAM

Prompt I – semantic labeling. Given a text label T and one patch P, judge if P can be classified as T.

Prompt II – instance merging. Given a text label T and two patches P1 and P2 classified as T, judge if P1 and P2 belong to the same instance of T.

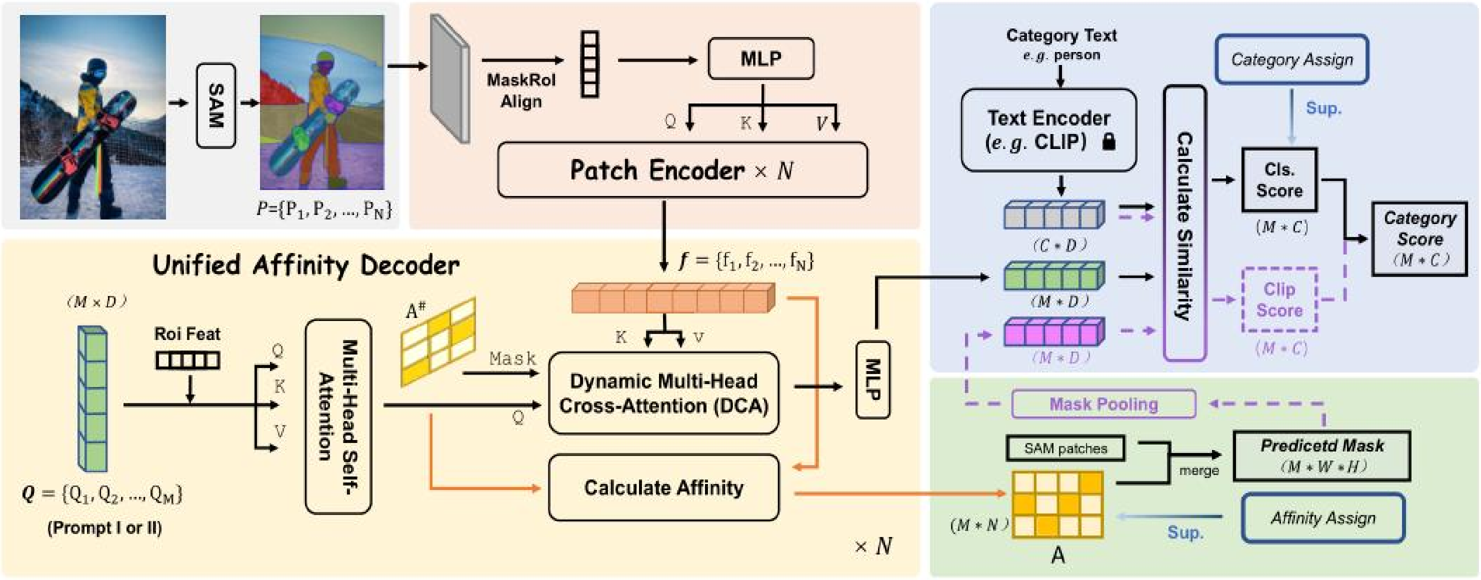


Figure 2: The unified affinity framework as an efficient implementation of SAM-CP. The input image with SAM patches is fed into a patch encoder. Type-I and Type-II prompts appear as two sets of queries. Affinity values are computed and the SAM patches are merged according to the affinity values. Semantic and instance level supervision are added to the merged patches. The purple arrows are present only in the inference stage of open-vocabulary segmentation. *Best viewed in color.*

Methods

A unified affinity framework:

Patches encoder:
extract features with MaskRoI Align， and then embed the feature with patch encoder (multi-head attention layers)

Unified affinity decoder:
Dynamic multi-head cross-attention to distinguish which patches belong to (calculate the affinity score) the semantic query for semantic segmentation (or instance query for instance segmentation)

Classifier:
The learnable classifier for close-vocabualry, and the CLIP classifier for open-vocabulary

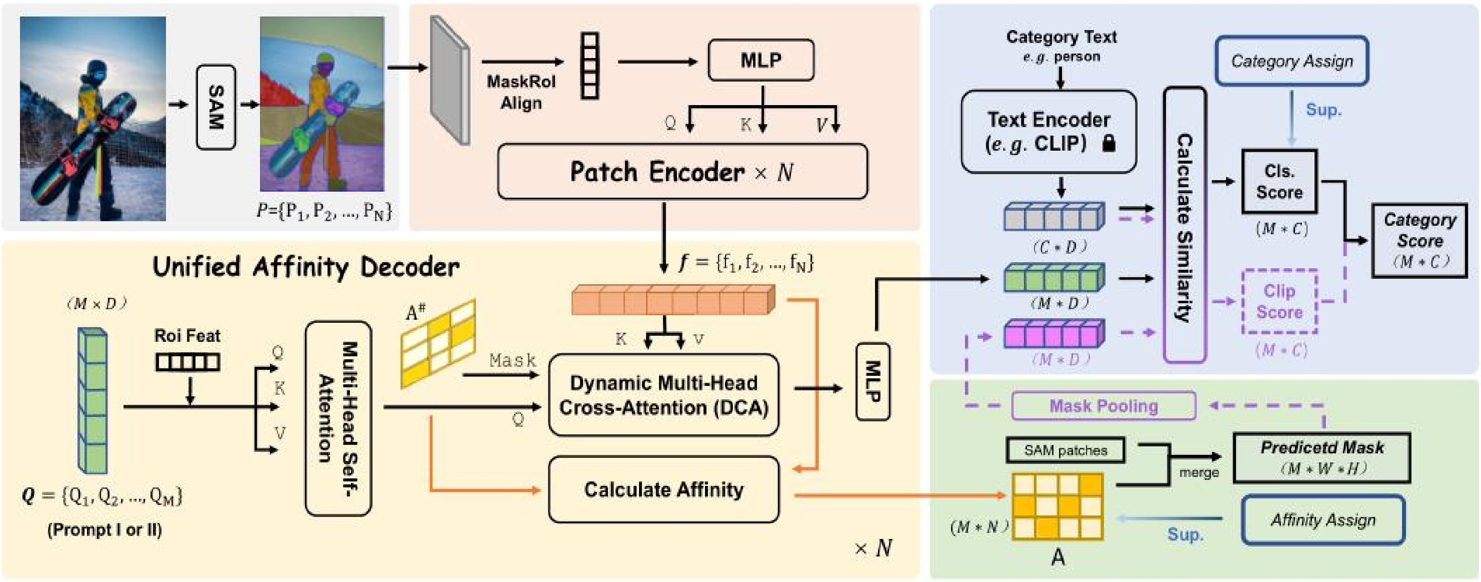


Figure 2: The unified affinity framework as an efficient implementation of SAM-CP. The input image with SAM patches is fed into a patch encoder. Type-I and Type-II prompts appear as two sets of queries. Affinity values are computed and the SAM patches are merged according to the affinity values. Semantic and instance level supervision are added to the merged patches. The purple arrows are present only in the inference stage of open-vocabulary segmentation. *Best viewed in color.*

Methods

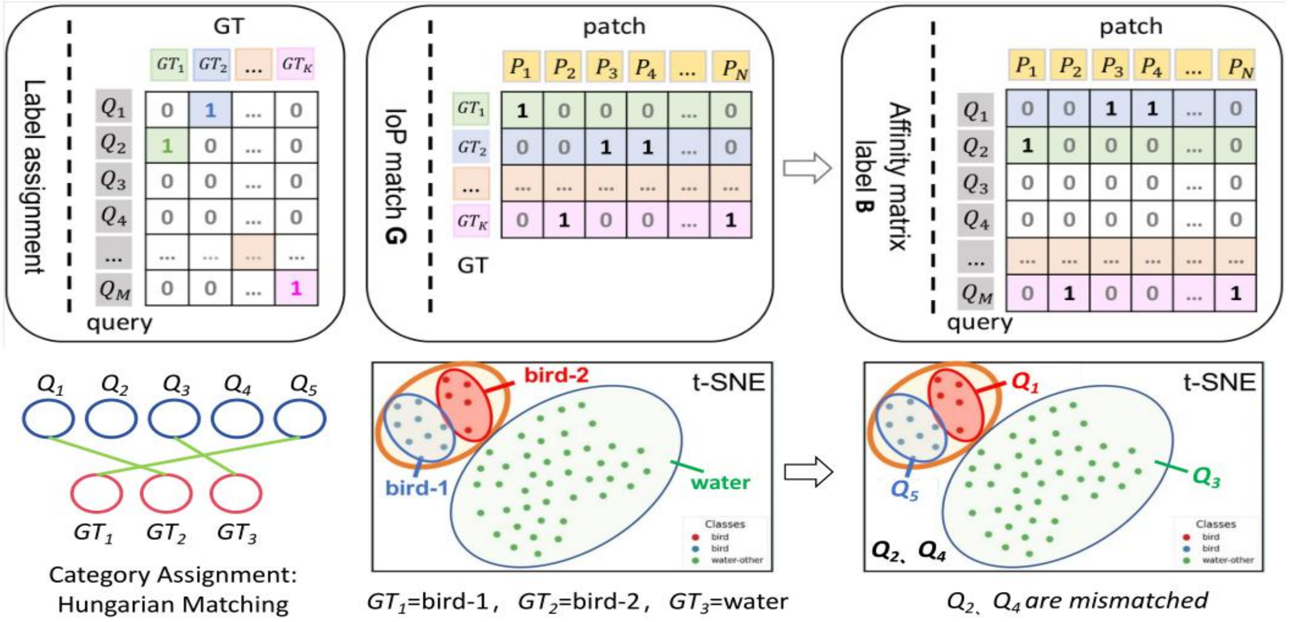


Figure 9: The illustration of how to get the ground-truth affinity matrix $\mathbf{B}$. The left is GTs&Q, the middle is GTs&P and the right is Q&P. Line 2 is the t-SNE visualization of category assignment.

How to determine the GT affinity supervision and the lable assignment

Algorithm 1 Affinity Similarity Calculation

Input: Query vectors $\mathtt{Q}$, Patch features $\mathtt{K}$, Head number η, Stage number ω.

Output: Affinity similarity $\hat{\mathbf{A}}$.

Note: $\mathtt{Q} \in \mathbb{R}^{M\times D}$, $\mathtt{K} \in \mathbb{R}^{N\times D}$, where M and N is the number of $\mathtt{Q}$ and $\mathtt{K}$. D is the feature dimension, which is a multiple of η. $s \in \mathbb{R}^1$, $\mathbf{b}_0 \in \mathbb{R}^D$ and $\mathbf{b}_1 \in \mathbb{R}^D$ are the learnable scaling factor and bias parameters to initialize the score to 0.01 for the focal loss.

$\mathtt{Q} \leftarrow fc^{\mathtt{Q}}(\mathtt{Q})$;
$\mathtt{K} \leftarrow fc^{\mathtt{K}}(\mathtt{K})$;
Reshape $\mathtt{Q}$ to $\mathbb{R}^{M\times\eta\times(D/\eta)}$ and transpose $\mathtt{Q}$ to $\mathbb{R}^{\eta\times M\times(D/\eta)}$;
Reshape $\mathtt{K}$ to $\mathbb{R}^{N\times\eta\times(D/\eta)}$ and transpose $\mathtt{K}$ to $\mathbb{R}^{\eta\times(D/\eta)\times N}$;
$\hat{\mathbf{A}} \leftarrow \frac{\mathtt{Q}\mathtt{K}^{\top}}{\sqrt{D/\eta}} \in \mathrm{R}^{\eta\times M\times N}$;
$\hat{\mathbf{A}} \leftarrow \mathrm{MLP}(\hat{\mathbf{A}}) \in \mathbb{R}^{1\times M\times N}$
Reshape $\mathbf{A}$ to $\mathbb{R}^{M\times N}$;
$\hat{\mathbf{A}} \leftarrow \frac{1}{s}\cdot\hat{\mathbf{A}} + \mathbf{b}$, where $\mathbf{b} = \mathbf{b}_1\mathtt{K} + \mathbf{b}_0$;
$\hat{\mathbf{A}}_\omega \leftarrow \hat{\mathbf{A}}$
if $\omega > 0$ **then**
    $\hat{\mathbf{A}} \leftarrow \hat{\mathbf{A}} + \hat{\mathbf{A}}_{\omega-1}$
end if

The affinity similarity calculation algorithm

Experiments and analysis

Method	Backbone	COCO→ADE20K					ADE20K→COCO					COCO→Cityscapes		
		PQ	SQ	RQ	AP	mIoU	PQ	SQ	RQ	AP	mIoU	PQ	AP	mIoU
MaskCLIP [15]	VIT-L	15.1	70.5	19.2	6.0	23.7	–	–	–	–	–	–	–	–
FreeSeg [39]	VIT-B	16.3	71.8	21.6	6.5	24.6	21.7	72.0	21.6	6.6	21.7	–	–	–
ODISE [52]	VIT-H	23.3	74.4	27.9	13.0	29.2	25.0	79.4	30.4	–	–	23.9	–	–
OPSNet [10]	VIT-L	19.0	52.4	23.0	–	–	–	–	–	–	–	41.5	–	–
MaskQCLIP [53]	VIT-L	23.3	–	–	–	30.4	–	–	–	–	–	–	–	–
X-Decoder [63]	Focal-L	21.8	–	–	13.1	29.6	–	–	–	–	–	38.1	24.9	52.0
FCCLIP [56]	CN-L	26.8	71.5	32.3	16.8	34.1	27.0	78.0	32.9	–	–	44.0	26.8	56.2
SAM-CP	CN-L	27.2	77.7	32.9	17.0	31.8	28.6	78.4	34.5	21.9	34.3	41.0	29.3	47.9

Table 1: Accuracy (%) of Open-vocabulary panoptic segmentation (in PQ, SQ and RQ), instance segmentation (in AP) and semantic segmentation (in mIoU). CN-L means ConvNext-L.

The performance of SAM-CP on COCO→ADE20K、ADE20K→COCO and COCO→Cityscapes about instance & semantic & panoptic segmentation

The visualization of SAM-CP on COCO dataset about panoptic segmentation

Experiments and analysis

The main ablation studies of SAM-CP:

- different loss & label assignment
- different modules

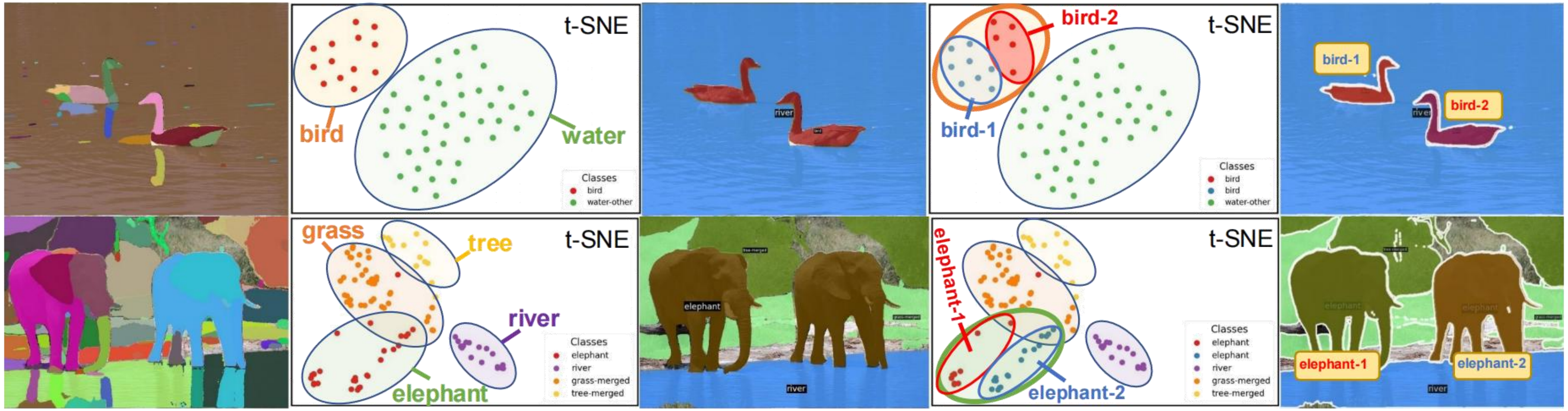
Loss	Label Assignment	Closed-domain (COCO) PQ	AP^{det}	AP	mIoU	Open-domain (COCO→ADE20K) PQ	SQ	RQ	AP	mIoU
all	all	47.0	45.8	41.4	54.2	27.2	77.7	32.9	17.0	31.8
w/o $\mathcal{L}_{\mathrm{mfl}}$	w/o mfl	0.0	3.5	0.0	0.0	0.6	22.0	0.9	0.0	3.4
w/o $\mathcal{L}_{\mathrm{dice}}$	w/o dice	41.3	35.1	34.3	48.3	23.8	73.4	29.1	15.8	28.6
all	w/o mfl	42.8	44.0	39.8	51.4	26.5	78.2	32.3	17.2	31.6
all	w/o dice	45.3	44.8	40.6	53.7	26.6	76.6	32.4	16.7	31.5
all	w/o box & giou	45.5	44.0	40.7	53.9	25.9	76.1	31.6	16.4	30.5

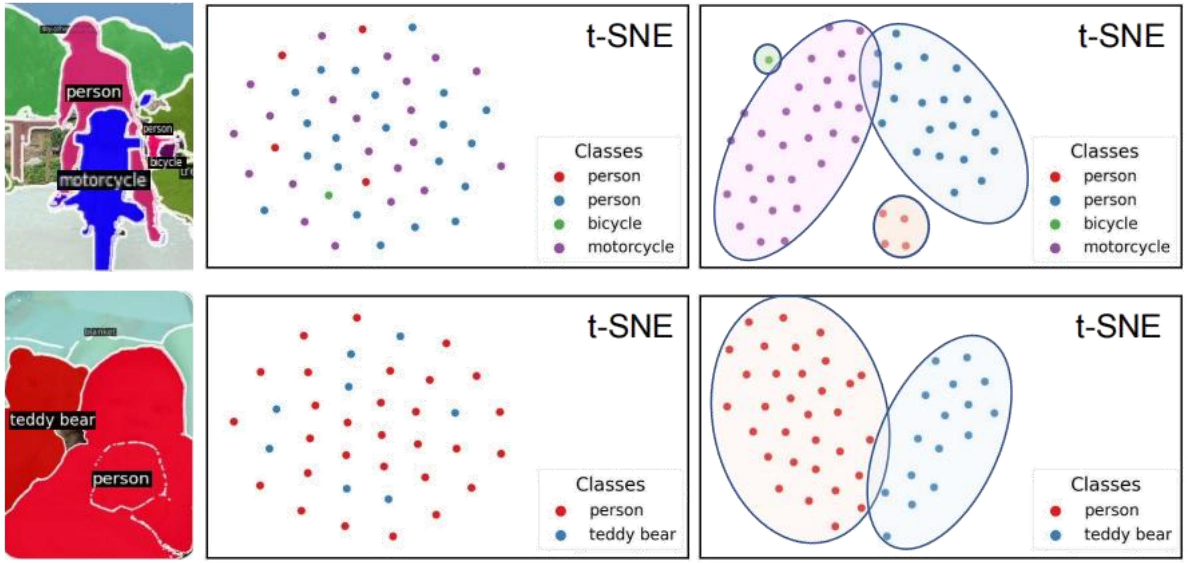
Table 3: Accuracy (%) in open and closed domains with different loss terms and matching strategies.

DCA	AR	MaskRoI	QE	BG	Closed-domain (COCO) PQ	AP^{det}	AP	mIoU	Open-domain (COCO→ADE20K) PQ	SQ	RQ	AP	mIoU
	✓	✓	✓	✓	45.4	45.6	41.1	51.8	26.6	76.9	32.5	16.6	31.7
✓		✓	✓	✓	43.5	44.0	39.9	51.1	25.8	76.8	31.3	16.3	30.5
✓	✓		✓	✓	44.1	45.3	40.6	51.1	25.6	74.4	31.1	16.5	30.3
✓	✓	✓		✓	44.8	44.5	40.5	51.6	26.5	75.7	32.1	16.5	31.4
✓	✓	✓	✓		45.2	45.4	41.3	52.6	25.5	75.7	31.2	16.1	30.3
✓	✓	✓	✓	✓	47.0	45.8	41.4	54.2	27.2	77.7	32.9	17.0	31.8

Table 4: Accuracy (%) in open and closed domains with different modules in the SAM-CP framework.

Experiments and analysis





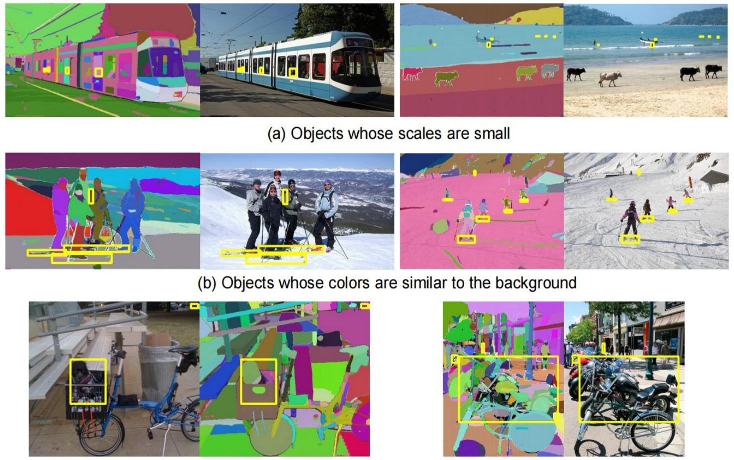
Through t-SNE visualization, we can see that SAM fragments belonging to the same category converge together in the feature space, while SAM fragments belonging to different instances within the same category converge together

Through t-SNE visualization, each SAM fragment is almost independent in the feature space of SAM, while in our SAM-CP feature space, they can be clustered into corresponding clusters based on semantics/instances

Experiments and analysis

Figure 8: The visualization of part and general instance segmentation. The result is obtained with one model and text labels of different granularities. On the left are the GTs, and on the right are the results.

Other Versatile Segmentation: Part segmentation



(a) Objects whose scales are small

(b) Objects whose colors are similar to the background

(c) Object which is occluded

(d) Object which has cluttered components

Limitation of SAM's mask quality

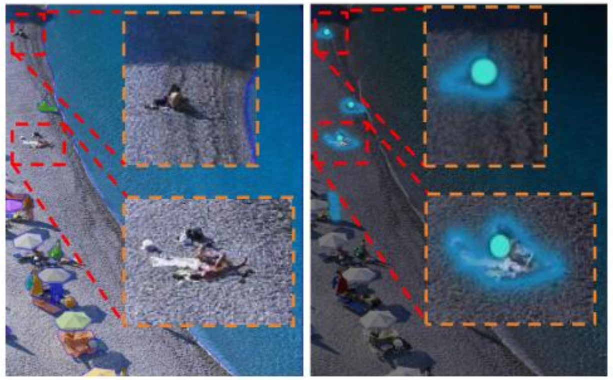


Figure 6: Dynamic prompts for looking for small objects.

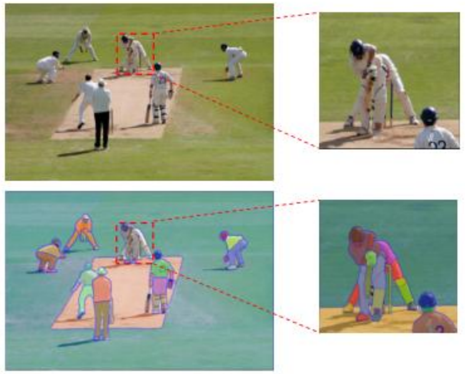


Figure 7: Interactive SAM calling yields finer results.

Swin-L	300	100queries	seg.	52.7
Swin-L	100	200queries	seg.	57.8
Swin-L	50	300quries	reg.+seg.	58.3
R50	24	patch+text	SAM*	48.4
R50	24	patch+text	SAM*+MD	50.7
R50	50	patch+text	SAM*+MD	51.5
Swin-L	36	patch+text	SAM*	52.9
Swin-L	36	patch+text	SAM*+MD	54.7
Swin-L	50	patch+text	SAM*+MD	54.5

The closed domain is not SOTA yet, how can we make up for it?

Future work

Thank you!