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

Yu Jiening
Yu Jiening@umac.mo

um 澳大

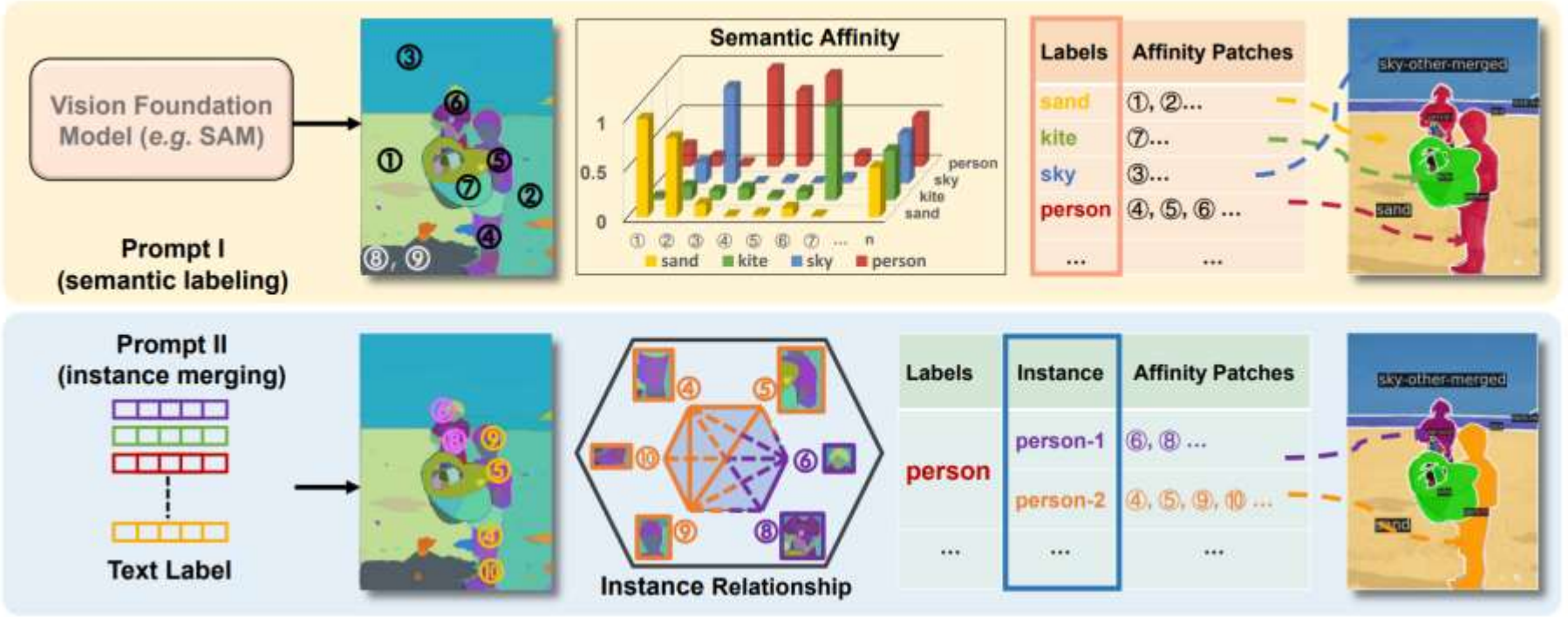
Catalogue

- Introduction
- Method
- Experiments
- Conclusions

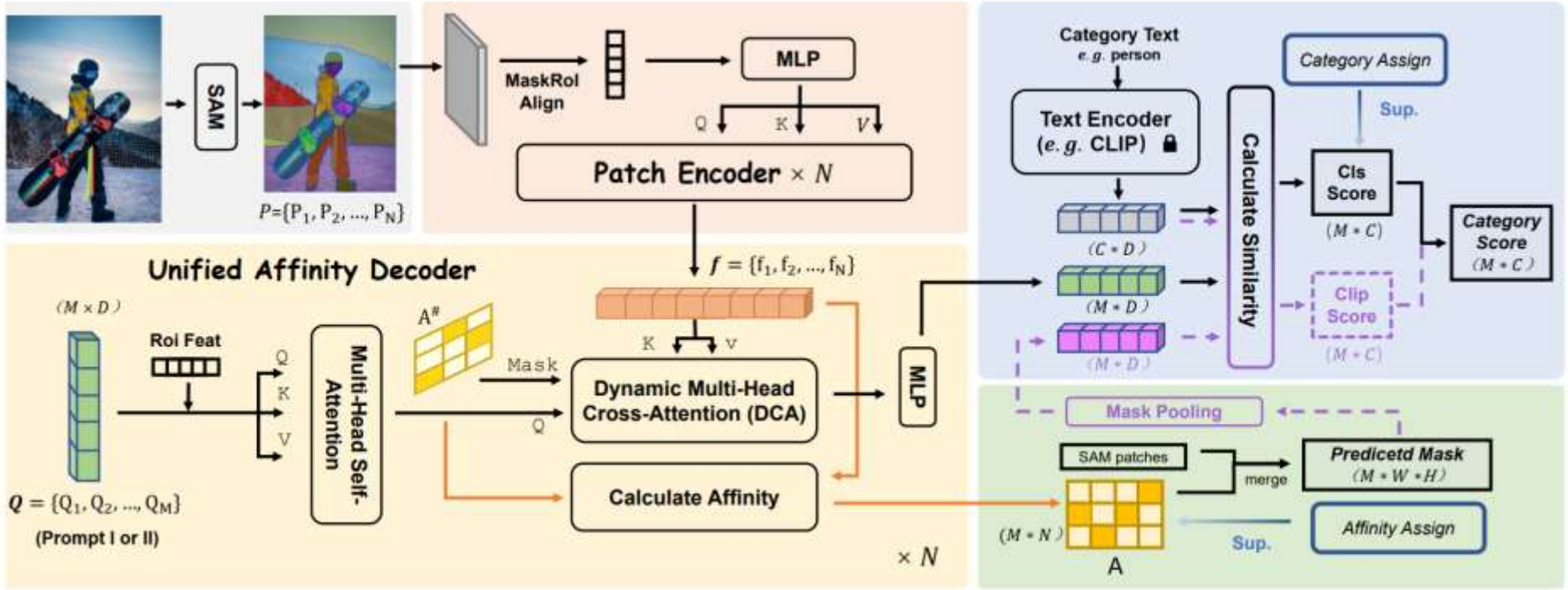
Introduction

- There still exist major challenges in applying SAM to semantic-aware segmentation tasks including semantic, instance, or panoptic segmentation.
- This paper presents a novel approach named SAM-CP where ‘CP’ stands for composable prompts. Different from the existing methods, we establish two types of prompts beyond the patches produced by SAM.

Method

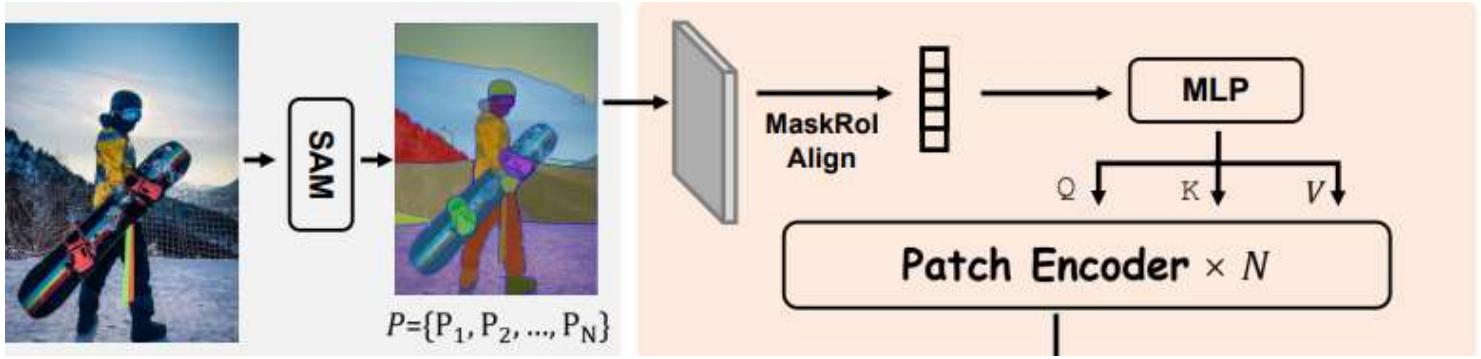


- The core idea is to establish two types of prompts beyond SAM.
- Prompt I – semantic labeling. Given a text label T and one patch P, judge if P can be classified as T.
- Prompt II – instance merging. Given a text label T and two patches P1 and P2 classified as T, judge if P1 and P2 belong to the same instance of T.



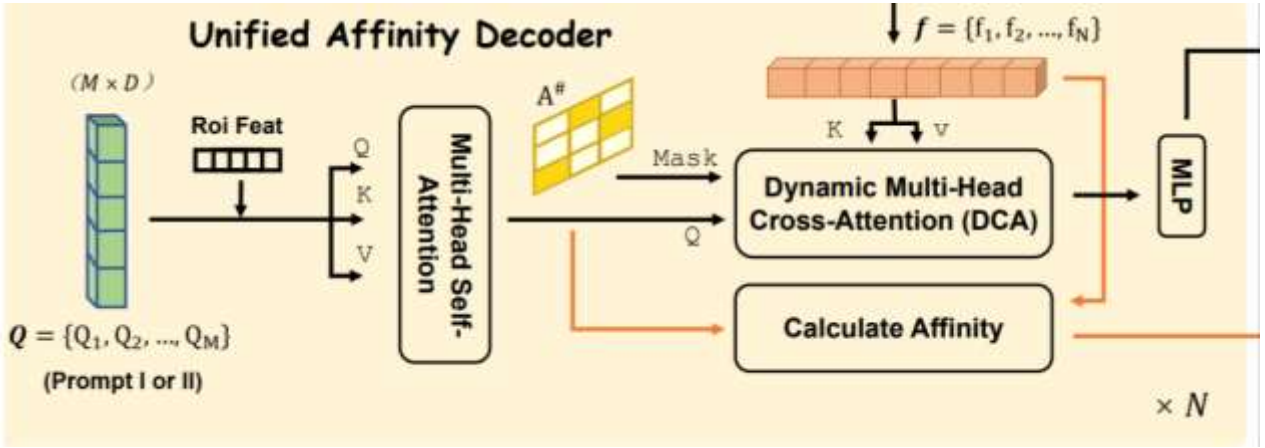
- The input image with SAM patches is fed into a patch encoder. Type-I and Type-II prompts appear as two sets of queries. Affinity values are computed and the SAM patches are merged according to the affinity values. Semantic and instance level supervision are added to the merged patches.

Patch Encoder



- For each patch Pn, apply a regular backbone equipped with a RoIAlign operator to obtain a basic feature vector ˜fn.
- These features are propagated through a multi-layer perceptron (MLP) and fed into ω multi-head self-attention layers

Unified Affinity Decoder



- The affinity is mathematically defined as a matrix A sized M × N. Each entry of A, Am,n, denotes the probability that the patch Pn belongs to the query Qm

Label Assignment and Supervision

- Each query, regardless of its type (semantic or instance), is expected to occupy a set of (one or more) patches and be assigned a class label. So, two sources of supervision are required.
- Semantic-level supervision: build a vision-language classifier upon the semantic queries. Following GLIP
- Instance-level supervision:

Experiments

Datasets

- COCO-Panoptic: 118K training and 5K validation images.
- ADE20K: 20,210 images, use the 150 most common object categories

Quantitative Results

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

Open-vocabulary segmentation

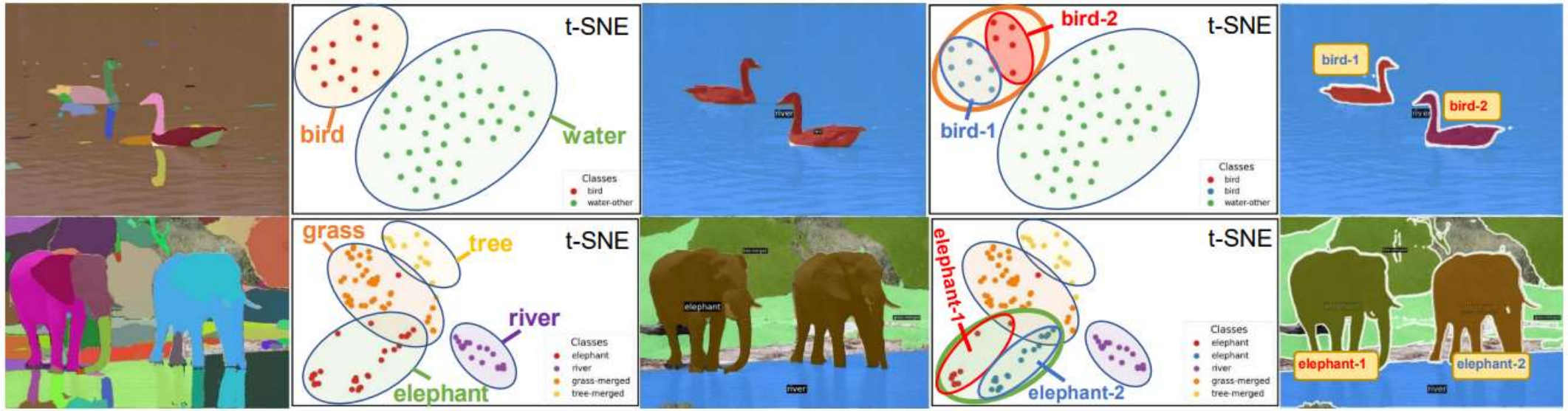
- Accuracy (%) of Open-vocabulary panoptic segmentation (in PQ, SQ and RQ), instance segmentation (in AP) and semantic segmentation (in mIoU). CN-L means ConvNext-L.

Closed-domain segmentation

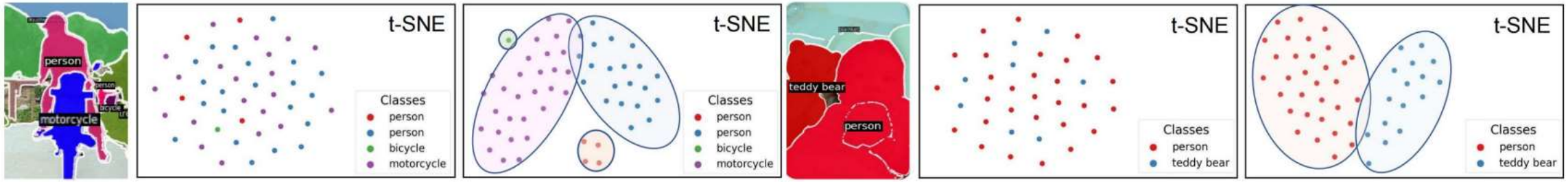
Method	Backbone	Seg. Style	COCO					ADE20K				
			Epoch	PQ	AP^{det}	AP	mIoU	Epoch	PQ	AP^{det}	AP	mIoU
DETR[3]	R50	reg.+seg.	50+25e	–	–	31.1	–	–	–	–	–	–
MaskFormer [12]	R50	seg.	300e	46.5	–	33.9	57.8	128e	34.7	–	–	–
Mask2Former [11]	R50	seg.	50e	51.5	–	41.7	61.7	128e	39.7	–	26.4	47.7
Mask2Former [11]	Swin-L	seg.	100e	57.8	–	48.6	67.4	128e	48.1	–	34.2	56.1
Mask DINO [27]	R50	reg.+seg.	50e	53.0	48.8	44.3	60.6	–	–	–	–	–
Mask DINO [27]	Swin-L	reg.+seg.	50e	58.3	56.2	50.6	67.3	128e	–	–	–	56.6
X-Decoder [63]	Focal-T	seg.	50e	52.6	–	41.3	62.4	128e	41.6	–	27.7	51.0
X-Decoder [63]	Focal-L	seg.	50e	56.9	–	46.7	67.5	128e	49.6	–	35.8	58.1
SAM-CP	R50	SAM*	36e	48.6	46.1	41.7	55.6	128e	38.5	28.7	25.1	42.4
SAM-CP	Swin-L	SAM*	36e	52.7	50.4	45.2	61.8	128e	44.4	34.6	30.3	49.4

- SAM-CP reports higher PQ and AP but a lower mIoU, which implies its advantageous performance in instance-level recognition.

Quantitative Results



- The leftmost column shows the input image with SAM patches; the middle and right parts show the semantic and instance segmentation results, respectively.



- The t-SNE visualization upon the visual features of SAM and SAM-CP. The points with the same color belong to the same semantic class (according to the ground truth).

Ablative Studies

Loss	Label Assignment	Closed-domain (COCO)				Open-domain (COCO→ADE20K)				
		PQ	AP^{det}	AP	mIoU	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 $\mathcal{L}_{mfl}$	w/o mfl	0.0	3.5	0.0	0.0	0.6	22.0	0.9	0.0	3.4
w/o $\mathcal{L}_{dice}$	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

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

DCA	AR	MaskRoI	QE	BG	Closed-domain (COCO)				Open-domain (COCO→ADE20K)				
					PQ	AP^{det}	AP	mIoU	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		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

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

Strategy	PQ	AP^{det}	AP	mIoU
patch-level	47.0	45.8	41.4	54.2
patch-level, w/o PE	45.8	45.0	40.6	51.7
image-level	46.2	45.6	40.9	52.3

- Accuracy (%) on COCO (on R50, 12 epochs) with different strategies of dynamic cross-attention (DCA), where ‘patch-level’ is the best option, ‘PE’ denotes the patch encoder.

Learnable	CLIP	PQ	PQ^{th}	PQ^{st}	AP	mIoU
✓		17.5	15.7	21.1	11.9	19.1
	✓	16.9	14.1	22.6	6.7	22.5
✓	✓	27.2	27.0	27.7	17.0	31.8

- Accuracy (%) on the setting of COCO→ADE20K (on ConvNext-L, 12 epochs) with different classifier types for open domain. ‘Learnable’ means classifier trained on COCO.

- A showcase of our results on COCO for closed domain. In each group, from left to right: input, SAM patches, panoptic segmentation by SAM-CP (ours).

τ	PQ	AP^{det}	AP	mIoU
0.9	43.2	42.8	38.2	49.9
0.8	45.0	45.4	40.7	52.6
0.7	43.3	43.3	39.9	50.3
0.5	42.0	34.3	37.2	48.9
0.8 (w/o IoP_{box})	42.1	31.7	35.4	48.4
0.8 (w/o IoP_{mask})	39.7	42.3	32.5	44.1
0.8 (w/low quality match)	47.0	45.8	41.4	54.2

- Accuracy (%) on COCO (with R50, 12 epochs) with different definitions of parts in the training stage.

Proposal types	PQ	AP^{det}	AP	mIoU
SAM*	48.6	46.1	41.7	55.6
SAM* + MD	51.4	51.6	45.8	57.3

- Accuracy (%) on COCO (with R50, 36 epochs) by adding proposals.

κ	0.0	0.3	0.4	0.5	0.8	1.0
PQ	17.5	26.9	27.2	26.2	22.9	16.9

- COCO→ADE20K accuracy (%) with different coefficients, κ.

	mIoU ↑	$\text{mIoU}_{>0.5}$ ↑	$\text{MR}_{0.25}$ ↓	$\text{MR}_{0.5}$ ↓	$\text{MR}_{0.75}$ ↓
Mask DINO	76.3	83.0	4.2%	10.1%	32.3%
SAM	71.1	79.5	8.9%	16.7%	39.3%
SAM+Merging	73.3	81.2	9.0%	15.0%	33.3%

- A comparison between the mIoU and missing rates with respect to different IoUs for COCO (val2017) instance segmentation.

Conclusion

- We propose SAM-CP, a novel approach that equips SAM with semantic and instance segmentation abilities.
- The inference speed of SAM-CP is bound by that of SAM; once a more efficient vision foundation model is available, our framework can be seamlessly transplanted and achieve faster inference.

Thank You!

Avenida da Universidade, Taipa, Macau, China
Tel : (853) 8822 8833 Fax : (853) 8822 8822
Email : info@um.edu.mo Website : www.um.edu.mo