Unified Open-World Segmentation with Multi-Modal Prompts

Yang Liu
Zhejiang University

Yufei Yin
Hangzhou Dianzi University

Chenchen Jing
Zhejiang University of Technology

Muzhi Zhu
Zhejiang University

Hao Chen
Zhejiang University

Yuling Xi
Zhejiang University

Bo Feng
Apple

Hao Wang
Apple

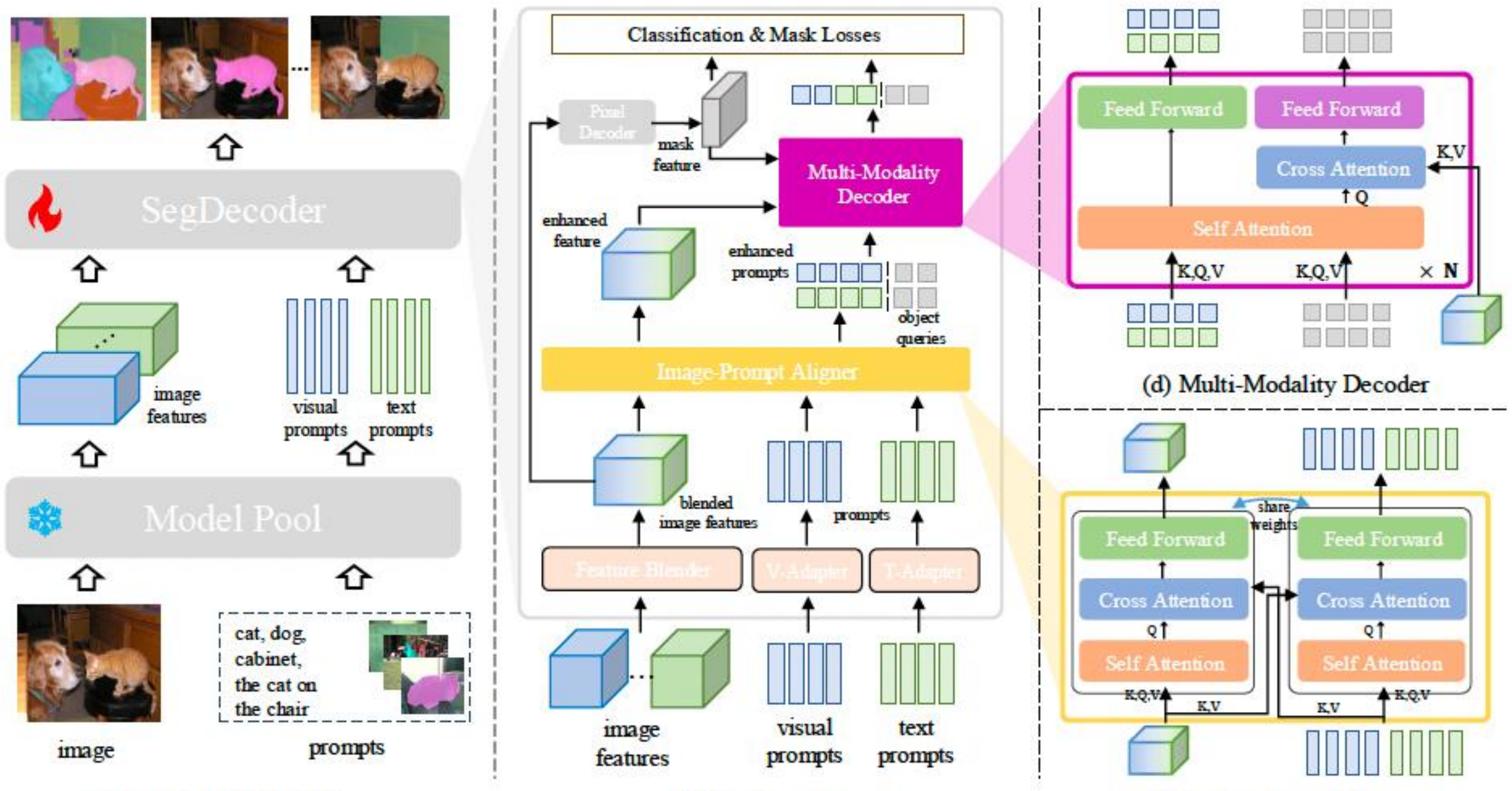
Shiyu Li
Apple

Chunhua Shen
Zhejiang University

Motivation

- Traditional **closed-world** segmentation is restricted to **fixed** categories.
- **Open-world** segmentation enables recognition of **arbitrary** objects guided by prompts.
- Two major paradigms exist:
  - **- Open-Vocabulary Segmentation (text prompts)**
  - **- In-Context Segmentation (image prompts)**
- **Limitation**: Existing works treat them separately, lacking a **unified** framework that leverages **both modalities** together.

Method



(a) Overview of COSINE (b) SegDecoder (d) Multi-Modality Decoder (c) Image-Prompt Aligner

- **Model Pool**: Pretrained CLIP (vision & text), DINOv2 extract multi-modal features.
- **SegDecoder**: - *Feature Blender* - *Image-Prompt Aligner* - *Pixel Decoder* - *Multi-Modality Decoder*
- **Training**: Only SegDecoder is trained → efficient, unleashes foundation models.
- **Inference**: Supports image prompts, text prompts, or both collaboratively.

Experiments

Methods	Venue	few-shot sem. LVIS-92^i		few-shot ins. LVIS		open-voc. pano. ADE20K			open-voc. pano. Cityscapes		open-voc. sem. A-847	open-voc. sem. PC-459
		one-shot	few-shot	AP	APr	PQ	AP	mIoU	PQ	mIoU	mIoU	mIoU
few-shot model												
HSNet [34]	ICCV'21	17.4	22.9	-	-	-	-	-	-	-	-	-
VAT [15]	ECCV'22	18.5	22.7	-	-	-	-	-	-	-	-	-
DiffewS [63]	NeurIPS'24	31.4	35.4	-	-	-	-	-	-	-	-	-
in-context model												
SegGPT [44]	ICCV'23	18.6	25.4	-	-	-	-	-	-	-	-	-
PerSAM-F [57]	ICLR'24	18.4	-	-	-	-	-	-	-	-	-	-
Matcher [26]	ICLR'24	33.0	40.0	-	-	-	-	-	-	-	-	-
SINE [27]	NeurIPS'24	31.2	35.5	8.6	7.1	-	-	-	-	-	-	-
open-vocabulary model												
ODISE [48]	CVPR'23	-	-	-	-	23.4	13.9	28.7	23.9	-	11.1	14.5
FC-CLIP [55]	NeurIPS'23	-	-	-	-	26.8	16.8	34.1	44.0	56.2	14.8	18.2
HIPIE [42]	NeurIPS'23	-	-	-	-	22.9	19.0	29.0	-	-	9.7	14.4
SED [46]	CVPR'24	-	-	-	-	-	-	-	-	-	13.9	22.6
universal model												
X-Decoder [65]	CVPR'23	-	-	-	-	21.8	13.1	29.6	38.1	52.0	9.2	16.1
UNINEXT* [51]	CVPR'23	-	-	-	-	8.9	14.9	6.4	-	-	1.8	5.8
OpenSeeD [56]	ICCV'23	-	-	-	-	19.7	15.0	23.4	41.4	47.8	-	-
DINOv [20]	CVPR'24	-	-	15.4	14.5	23.2	15.1	25.3	-	-	-	-
OMG-Seg [21]	CVPR'24	-	-	-	-	27.9	-	-	-	-	-	-
PSALM [58]	ECCV'24	-	-	-	-	-	13.9	24.4	-	-	-	14.0
COSINE†	this work	34.2	39.1	17.4	23.3	28.1	16.7	35.2	37.1	53.4	15.2	19.6
COSINE		35.2	40.7	20.3	25.8	31.0	21.1	35.7	42.0	56.1	15.6	19.2

Table 1. Results of different open world segmentation tasks including few-shot semantic segmentation, open-vocabulary panoptic segmentation and semantic segmentation. * We report the performance evaluated in [42]. † indicates the single-scale variant of COSINE.

Experiments

Method	Venue	refCOCO			refCOCO+			refCOCOg	
		val	testA	testB	val	testA	testB	val(U)	test(U)
MAttNet [54]	CVPR'18	56.5	62.4	51.7	46.7	52.4	40.1	47.6	48.6
MCN [30]	CVPR'20	62.4	64.2	59.7	50.6	55.0	44.7	49.2	49.4
VLT [9]	ICCV'21	67.5	70.5	65.2	56.3	61.0	50.1	55.0	57.7
LAVT [53]	CVPR'22	72.7	75.8	68.8	62.1	68.4	55.1	61.2	62.1
CRIS [45]	CVPR'22	70.5	73.2	66.1	62.3	68.1	53.7	59.9	60.4
ReLA [25]	CVPR'23	73.8	76.5	70.2	66.0	71.0	57.7	65.0	66.0
X-Decoder [65]	CVPR'23	-	-	-	-	-	-	64.6	-
SEEM [66]	NeurIPS'23	-	-	-	-	-	-	65.7	-
LISA [19]	CVPR'24	74.9	79.1	72.3	65.1	70.8	58.1	67.9	70.6
COSINE	this work	77.2	80.7	71.1	66.4	73.2	56.4	67.4	68.5

Table 2. Results of referring segmentation on refCOCO, refCOCO+ and RefCOCOg. We report the metric of cIoU.

Experiments

Methods	Venue	DAVIS 2017	YT-VOS 2019
		$J\&F$	G
with video data			
AOT [52]	NeurIPS'21	85.4	85.3
XMem [4]	ECCV'22	87.7	85.5
DEVA [5]	ICCV'23	86.8	85.5
Cutie [6]	CVPR'24	88.8	86.1
without video data			
Painter [43]	CVPR'23	34.6	20.6
SegGPT [44]	ICCV'23	75.6	73.1
SEEM [66]	NeurIPS'23	58.9	-
DINOv [20]	CVPR'24	73.3	52.0
PerSAM-F [57]	ICLR'24	76.1	46.6
SINE [27]	NeurIPS'24	77.0	66.4
COSINE	this work	76.7	66.0
COSINE-FT		80.2	70.0

Table 3. Results of video object segmentation on DAVIS 2017, and YouTube-VOS 2019. Gray indicates the model is trained on target datasets with video data.

Experiments

Prompt		LVIS-92^i		ADE20K		
vision	text	1-shot	5-shot	PQ	AP	mIoU
✓		24.5	27.8	-	-	-
	✓	-	-	13.2	7.6	30.2
✓	✓	27.7	32.1	17.7	8.1	30.4

Table 4. Effect of the interaction between visual and textual branches during Training. All models are trained for 10k steps.

Prompt		LVIS-92^i		ADE20K		
vision	text	1-shot	5-shot	PQ	AP	mIoU
✓		35.2	40.7	23.8	15.8	26.3
	✓	37.8	-	31.0	21.1	35.7
✓	✓	43.1	45.9	31.4	21.3	36.3

Table 5. Effect of the interaction between visual and textual branches during inference.

Experiments

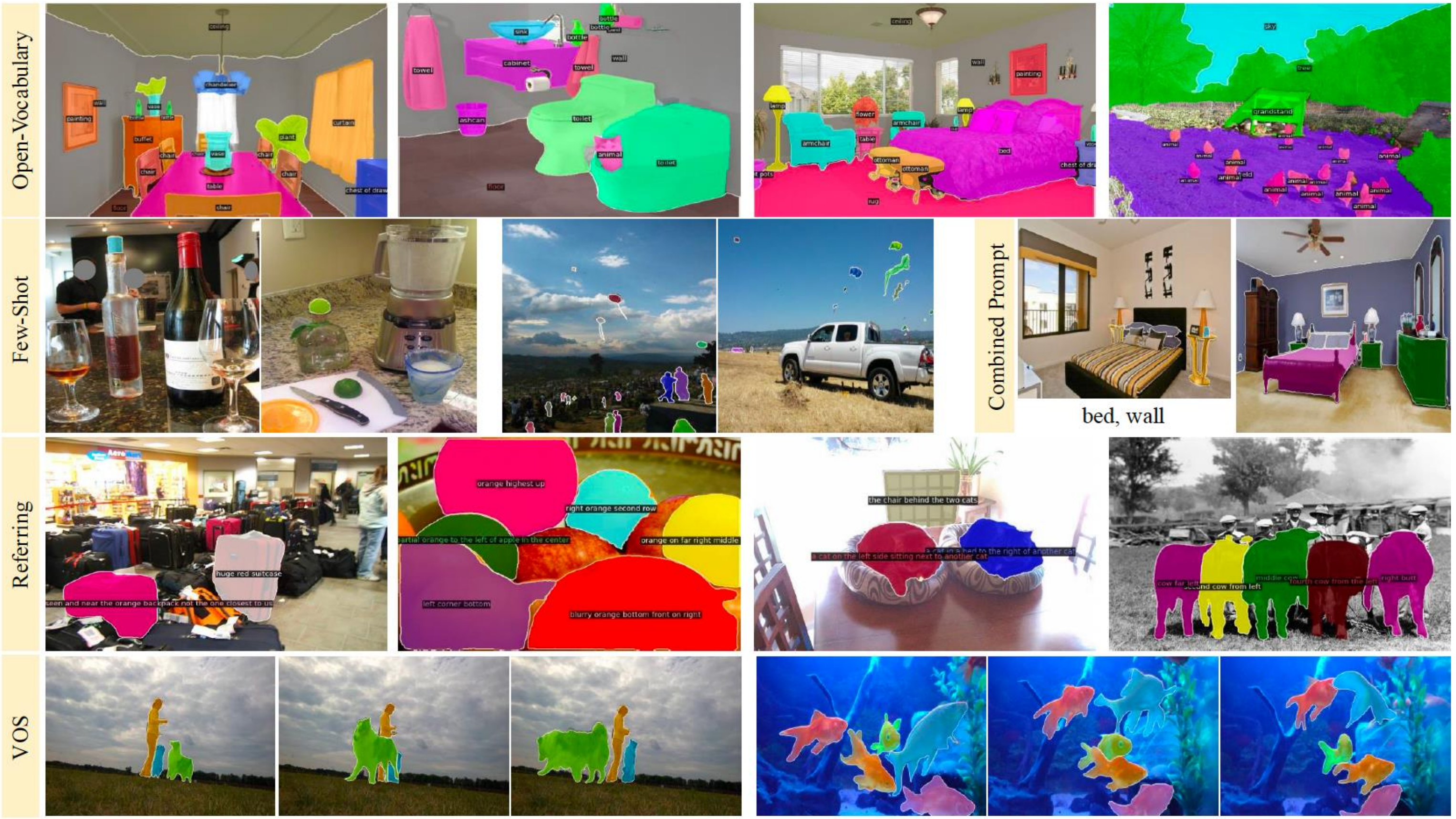


Figure 3. Qualitative results. COSINE can perform various open-world segmentation tasks with different modal prompts (image and text). For few-shot segmentation, the left image is the example image and the right is the result.

Experiments

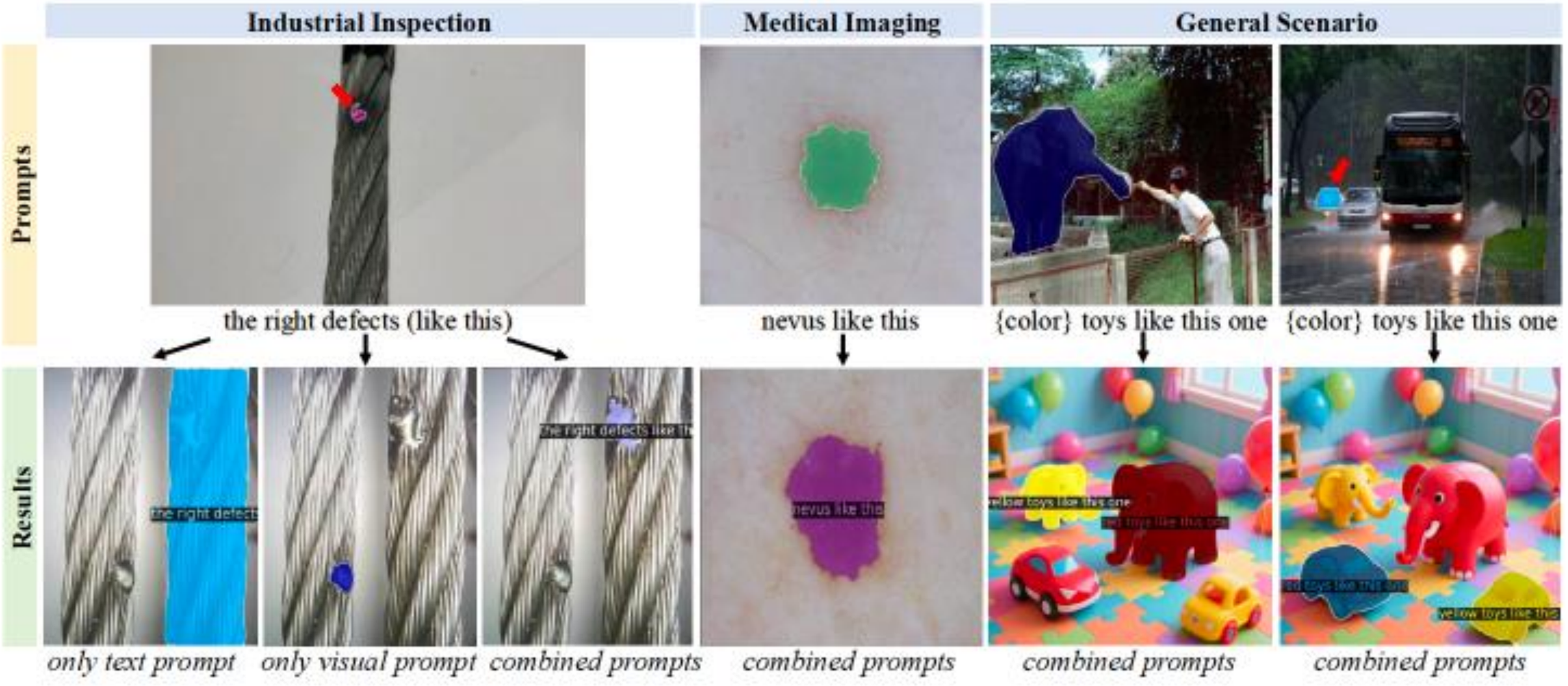


Figure 4. Visualization of prompt synergy. The top row shows the input prompts, the bottom row presents the corresponding outputs.

Thanks.