# Beyond SAM: Towards More Efficient, Unified, General Views of SAM

Xiangtai Li

2024-3-20

https://lxtgh.github.io/

# Overview

1, SAM overview.

2, Edge-SAM.

3, Open-Vocabulary SAM.

4, OMG-Seg.

5, Close Related Works and Summary.

# Outline

1, SAM overview.

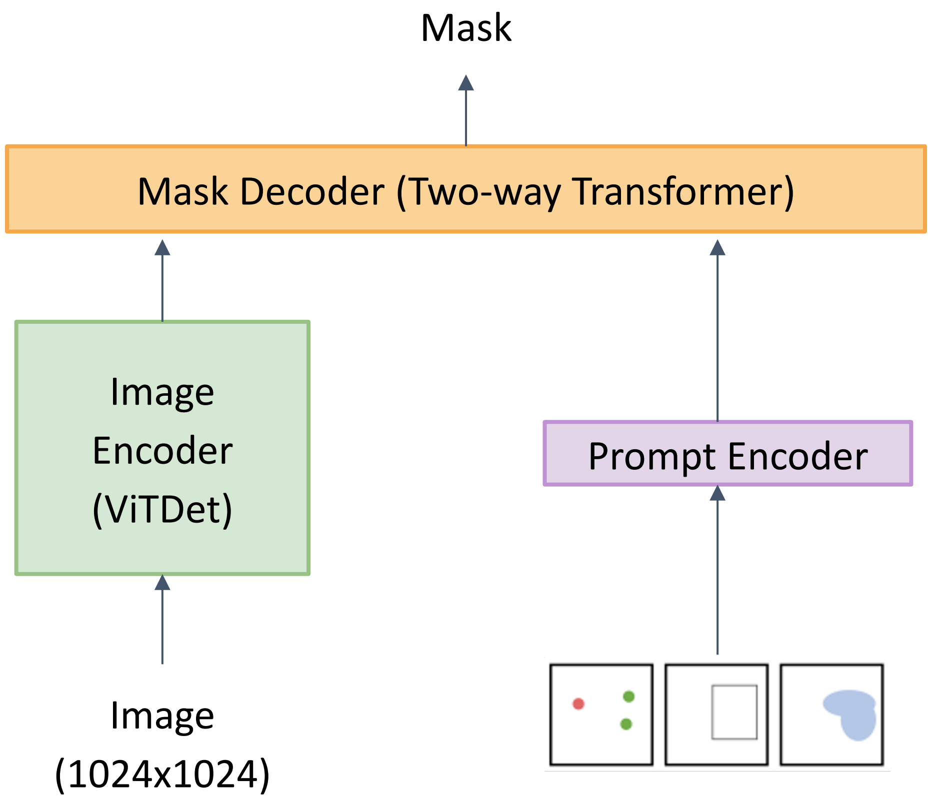
2, Edge-SAM.

3, Open-Vocabulary SAM.

4, OMG-Seg.

5, Close Related Works and Summary.

# 1, SAM overview




**Overall Architecture of SAM**


- **1+ billion masks**
- **11 million images**
- **SAM 1B dataset.**

# 1, SAM overview


SAM - Prompt-based segmentation model.

# 1, SAM overview

**Strong Points:**

- An easy to use interactive segmentation tool.
- Generalization ability with various visual prompts.
- SAM-1B dataset can used for community.
- Multi-granularity masks.

**Problems:**

- No semantic information.
- Not efficient and cannot used on device.
- No temporal association.
- Scale variance problem.

# Outline

1, SAM overview.

2, Edge-SAM.

3, Open-Vocabulary SAM.

4, OMG-Seg.

5, Close Related Works and Summary.

# 2, Edge-SAM

**Current solutions for efficient SAM models:**

1, Training a interactive model using SAM-1B data.

Eg: FastSAM

2, Distillation on the smaller encoder.

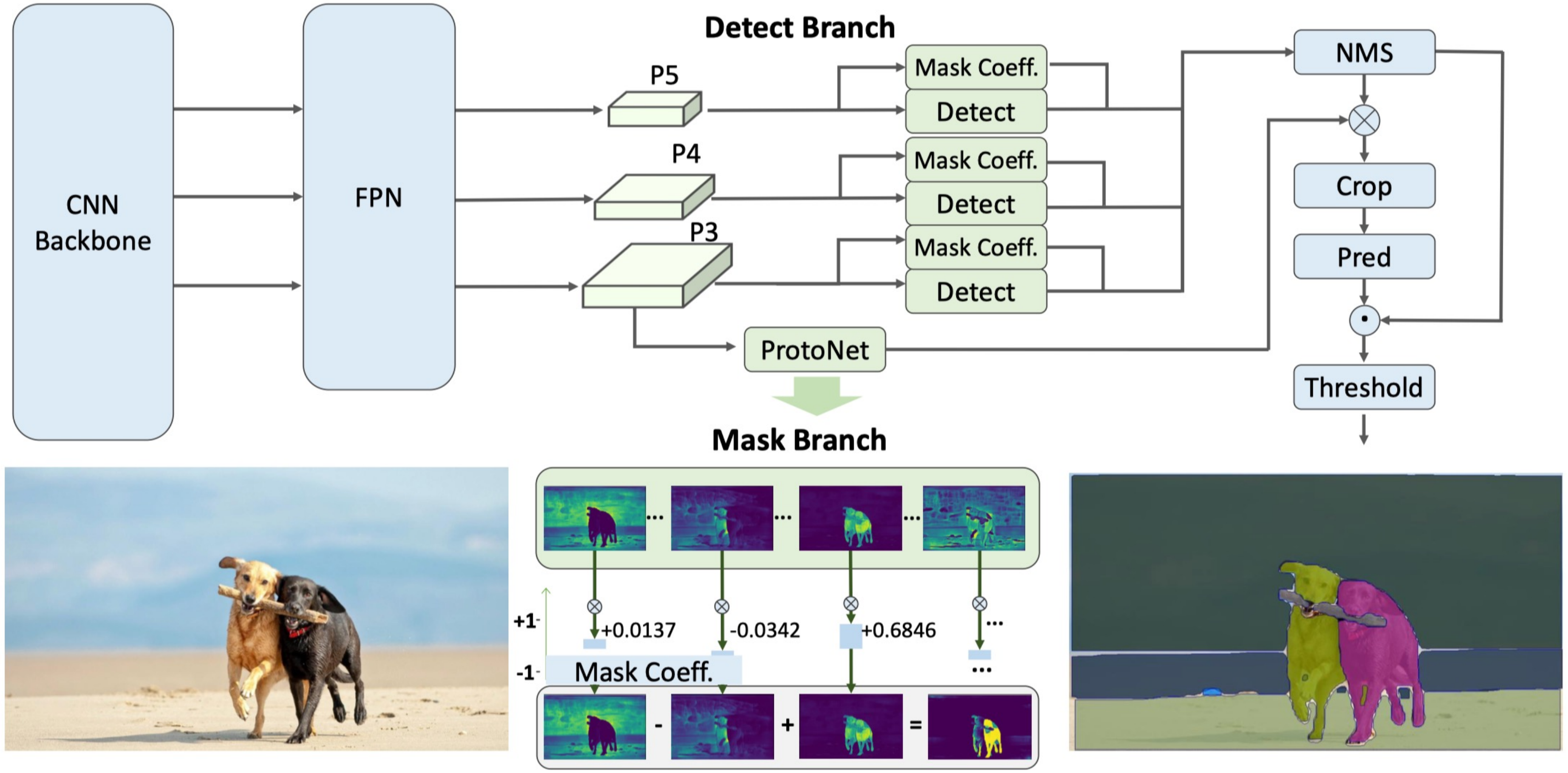
Eg: MobileSAM

3, Combine 1 and 2 together.

Eg: Efficient-SAM

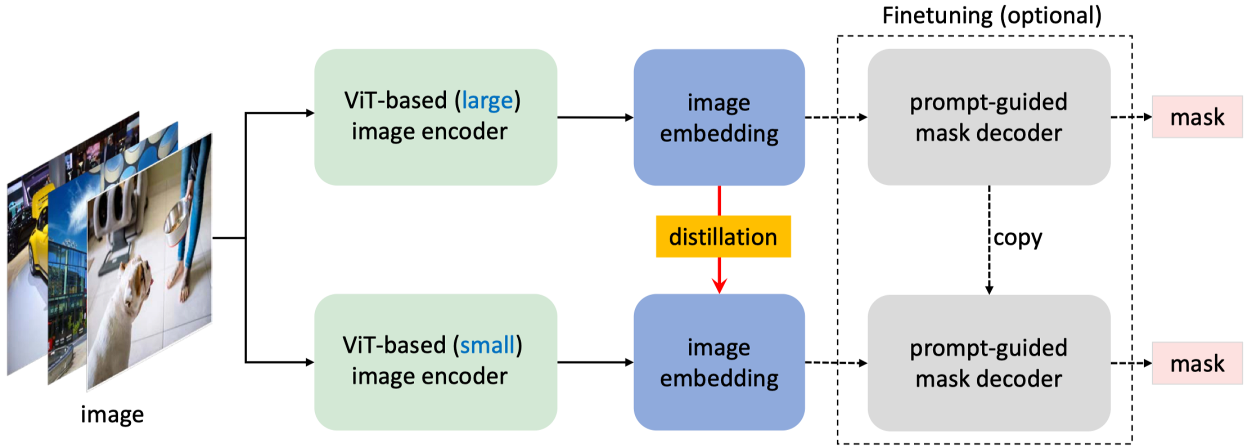
# 2, Edge-SAM

## FastSAM

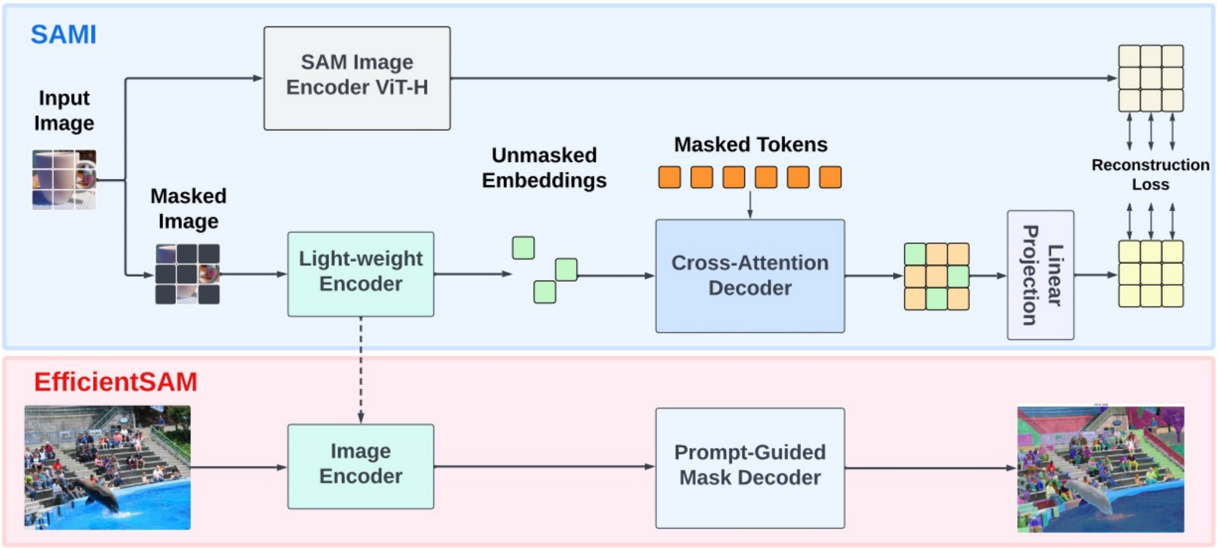



YOLACT + Rule-Based Selection
Train on 10% SA-1B

# 2, Edge-SAM

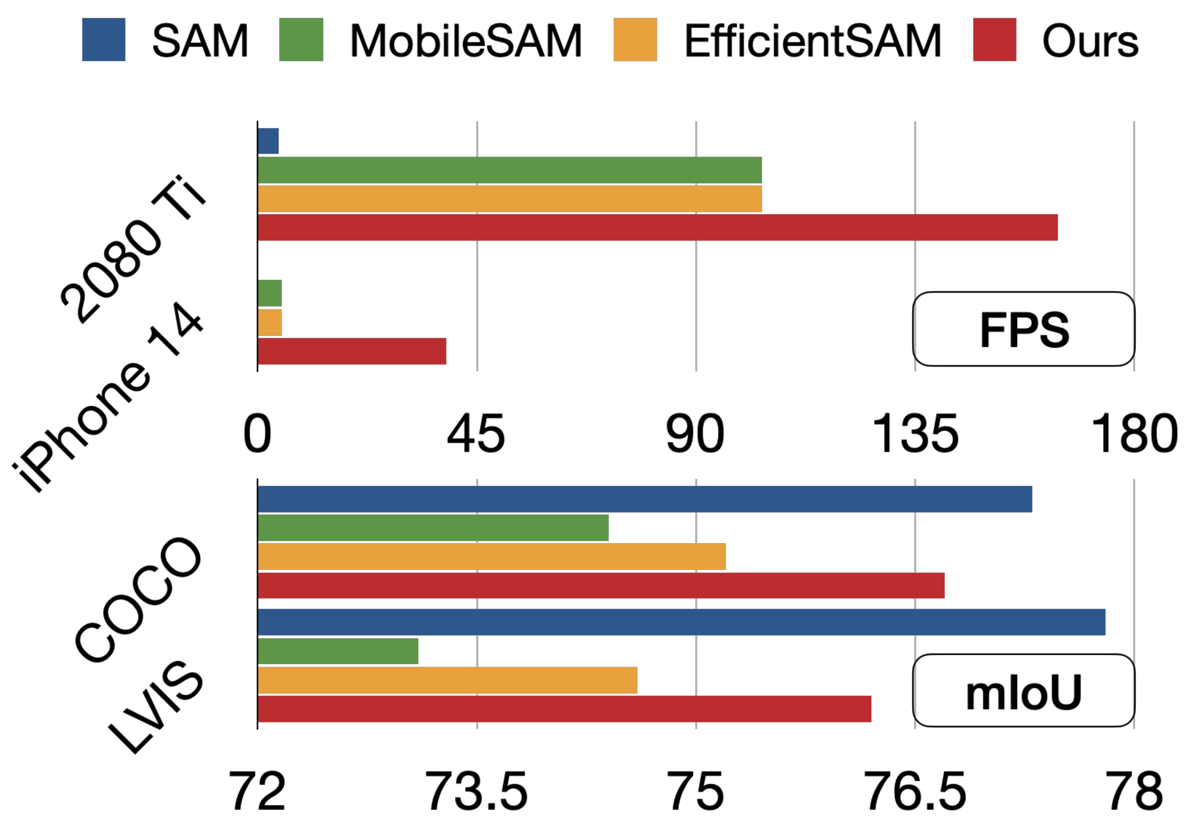



**Feature distillation on SAM encoder. (Mobile-SAM)**




**Knowledge Guided Mask Image Modeling Pre-train and then finetuning. (Efficient-SAM)**

# 2, Edge-SAM




**Our Goals:**

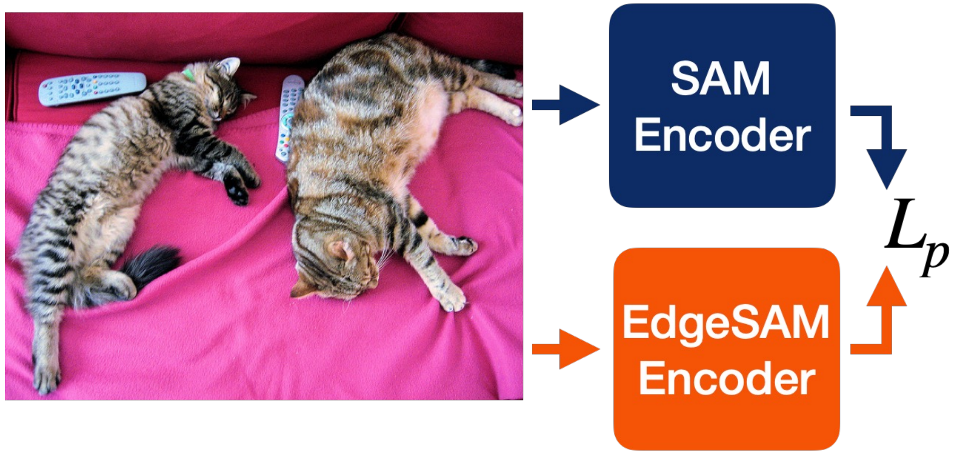
1, Faster and Accurate. (real-time)

2, Running on real device. Such as iPhone

3, Explore interactive property of SAM decoder.

# 2, Edge-SAM

## The first stage：Feature Distillation




$$L_p = \mathrm{MSE}(T_{enc}(I), S_{enc}(I))$$

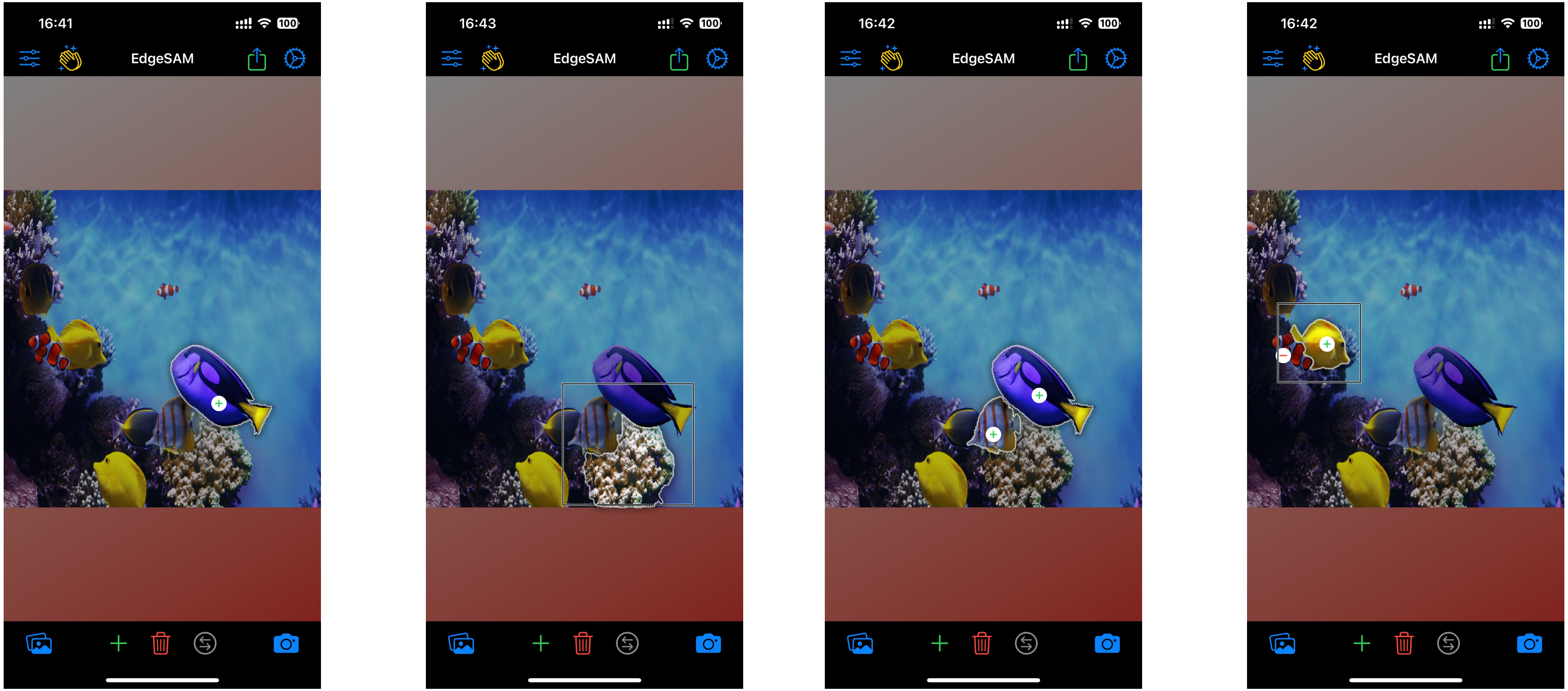
## Choose lightweight backbone

**(c) Choice of the backbone.** We apply encoder-only KD for this ablation.

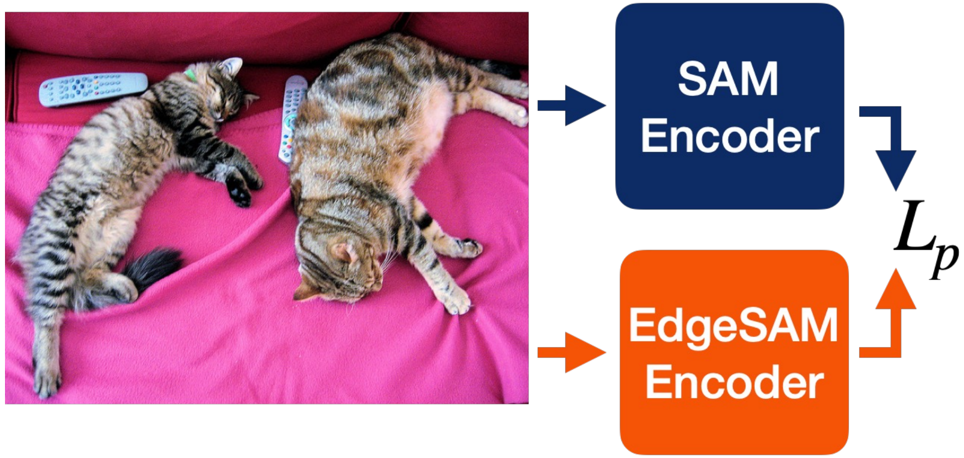
| Method | Res. Align | Type | Box | Center | FPS |
|---|---|---|---|---|---|
| TinyViT-5M | Remove Downsample | ViT | 82.0 | 64.6 | 103.5 |
| EfficientViT-B1 | | Hybrid | 81.6 | 63.7 | 117.0 |
| RepViT-M1 | | CNN | **82.1** | **64.9** | 155.7 |
| TinyViT-5M | FPN | ViT | 81.6 | 63.7 | 114.2 |
| EfficientViT-B1 | | Hybrid | 80.7 | 60.9 | 159.9 |
| RepViT-M1 | | CNN | 82.0 | 64.6 | **164.3** |

# 2, Edge-SAM

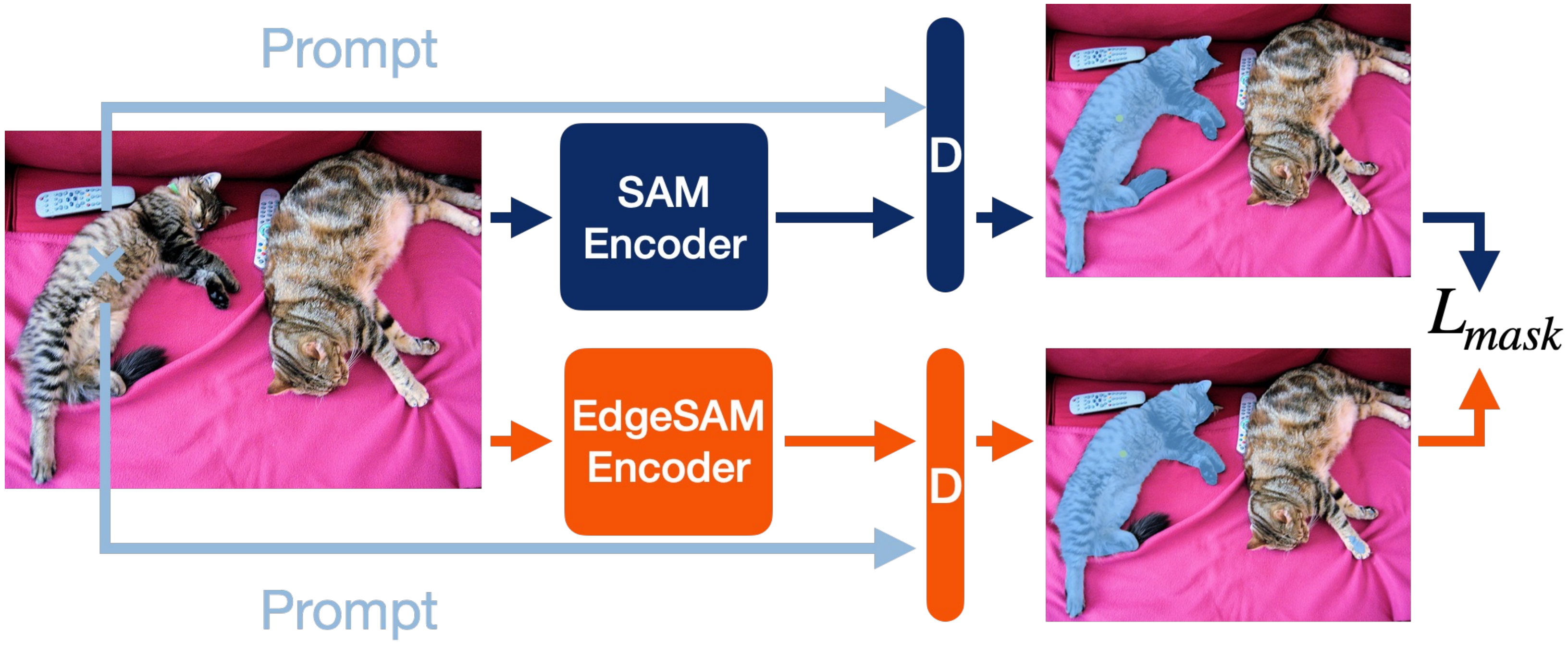
Add prompts during the distillation.



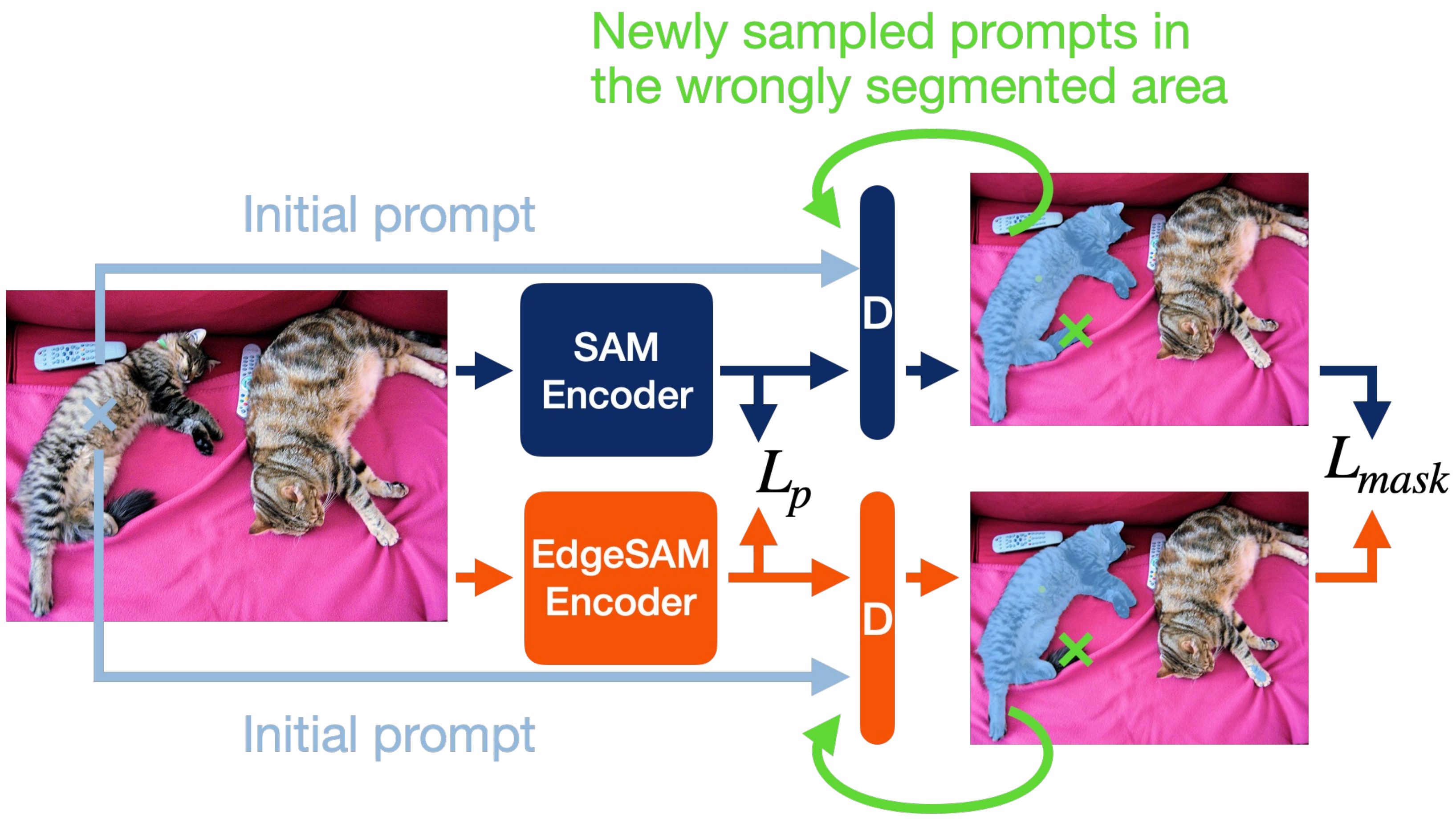
# 2, Edge-SAM



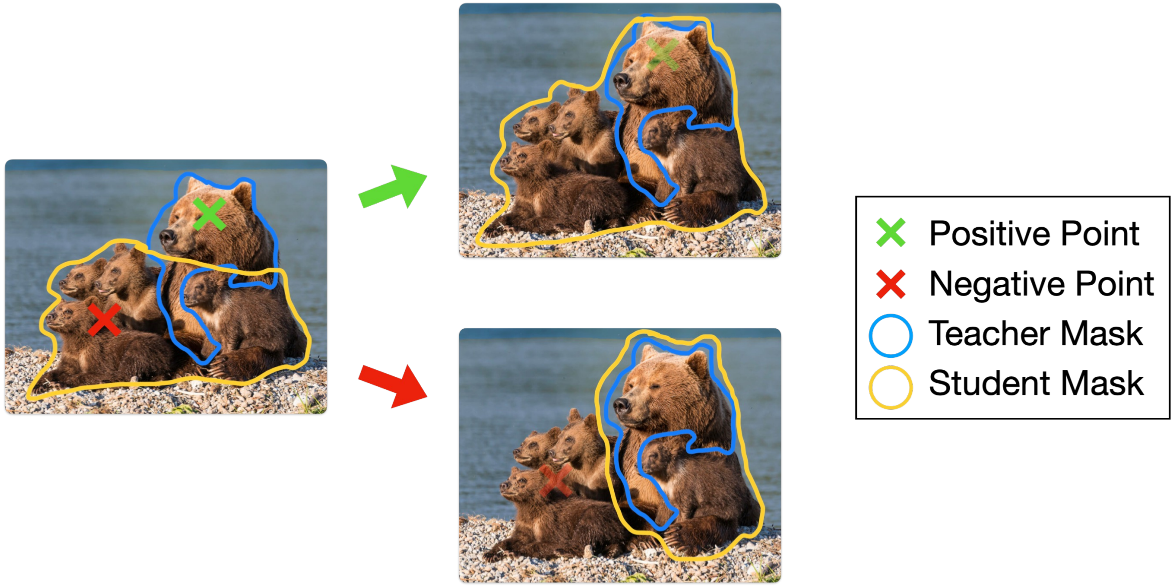
# 2, Edge-SAM



# 2, Edge-SAM


Newly sampled prompts in
the wrongly segmented area
Initial prompt
SAM
Encoder
D
$L_p$
$L_{mask}$
EdgeSAM
Encoder
D
Initial prompt

# 2, Edge-SAM




**Algorithm 1:** Prompt-In-the-Loop Distillation

$T_{enc}, S_{enc} \leftarrow$ SAM / EdgeSAM encoder;
$T_{dec} \leftarrow$ SAM decoder;
$S_{dec} \leftarrow$ EdgeSAM decoder initialized w/ $T_{dec}$;
$m, \mathbf{c} \leftarrow$ shared mask / IoU tokens;
$\mathcal{I}, \mathcal{P} \leftarrow$ images and prompts for training;
$N, M \leftarrow$ training steps, prompt sampling loops;
**for** $i = 1, 2, \ldots, N$ **do**
  $f_t, f_s \leftarrow T_{enc}(\mathcal{I}_i), S_{enc}(\mathcal{I}_i)$;
  $p \leftarrow$ select the box or point prompt in $\mathcal{P}_i$;
  $m_t, m_s \leftarrow T_{dec}(f_t, p, m, \mathbf{c}), S_{dec}(f_s, p, m, \mathbf{c})$;
  $L \leftarrow L_{\text{mask}}(m_t, m_s)$;
  **for** $j = 1, 2, \ldots, M$ **do**
    $\hat{p} \leftarrow \text{sample\_in\_disagree}(m_t, m_s)$;
    $p \leftarrow \hat{p}$ appends to $p$;
    $m_t, m_s \leftarrow$
     $T_{dec}(f_t, p, m, \mathbf{c}), S_{dec}(f_s, p, m, \mathbf{c})$;
    $L \leftarrow L + L_{\text{mask}}(m_t, m_s)$;
  **end**
  $S_{enc}, S_{dec} \leftarrow$ SGD model update;
**end**

# 2, Edge-SAM

**Table 4: Performance with boxes from an external object detector as prompts.** We report the mask mAP and boundary IoU on the COCO dataset. The box mAPs of Detic and ViTDet-H are 47.4 and 58.7 respectively.

| Method | Detic | | | | | ViTDet-H | | | | Train Set | FPS |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | AP | $AP_S$ | $AP_M$ | $AP_L$ | BIoU | AP | $AP_S$ | $AP_M$ | $AP_L$ | | |
| SAM | **38.8** | **26.9** | **44.1** | **50.3** | **26.8** | **46.1** | **33.6** | **51.9** | **57.7** | SA-1B | 4.3 |
| FastSAM | - | - | - | - | - | 37.9 | 23.9 | 43.4 | 50.0 | 2% SA-1B | <103.5 |
| MobileSAM | 33.1 | 21.7 | 37.8 | 44.8 | 20.2 | 39.4 | 26.9 | 44.4 | 52.2 | 1% SA-1B | <u>103.5</u> |
| EfficientSAM-Ti | - | - | - | - | - | <u>42.3</u> | 26.7 | 46.2 | <u>57.4</u> | SA-1B+IN | <u>103.5</u> |
| **EdgeSAM** | <u>35.2</u> | <u>23.5</u> | <u>40.3</u> | <u>46.6</u> | <u>22.5</u> | 42.2 | <u>29.6</u> | <u>47.6</u> | 53.9 | 1% SA-1B | **164.3** |

**Table 2: Performance with GT boxes as prompts.** We report the mIoU across all instances in the test set. *+1 pt.* denotes appending an additional refinement point as the prompt. **Bold** marks the best while <u>underline</u> marks the second best. Since EfficientSAM is trained on the entire SA-1B dataset, we do not evaluate it on SA-1K.

| Method | SA-1K | | | COCO | | | LVIS | | |
|---|---|---|---|---|---|---|---|---|---|
| | Box | +1 pt. | +2 pt. | Box | +1 pt. | +2 pt. | Box | +1 pt. | +2 pt. |
| SAM | **86.7** | **86.7** | **87.1** | **77.3** | <u>77.7</u> | <u>78.1</u> | **77.8** | **78.3** | **78.5** |
| MobileSAM | 82.0 | 82.4 | 82.7 | 74.4 | 74.8 | 75.1 | 73.1 | 73.7 | 74.0 |
| EfficientSAM-Ti | - | - | - | 75.2 | 76.0 | 76.6 | 74.6 | 75.2 | 75.1 |
| **EdgeSAM** | <u>83.0</u> | <u>83.7</u> | <u>84.1</u> | <u>76.7</u> | **78.1** | **79.0** | <u>76.2</u> | <u>77.3</u> | <u>78.0</u> |

**Table 3: Performance with center points as prompts.** Similar to Tab. 2 but using the mask center point as the initial prompt.

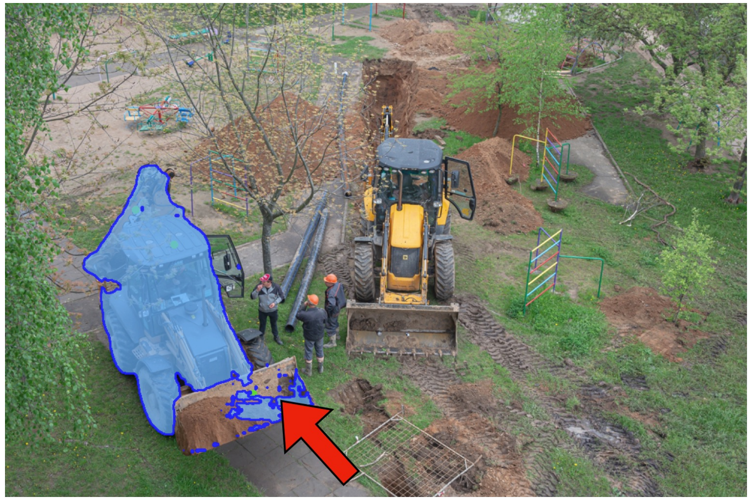
| Method | SA-1K | | | COCO | | | LVIS | | |
|---|---|---|---|---|---|---|---|---|---|
| | Center | +1 pt. | +2 pt. | Center | +1 pt. | +2 pt. | Center | +1 pt. | +2 pt. |
| SAM | **76.5** | **83.4** | **85.1** | **53.6** | **67.4** | **71.7** | **60.5** | **68.1** | **70.7** |
| MobileSAM | 64.6 | 73.4 | 76.2 | <u>50.9</u> | <u>63.0</u> | 66.8 | 52.1 | 59.9 | 63.0 |
| EfficientSAM-Ti | - | - | - | 49.8 | 60.5 | 65.7 | <u>56.4</u> | 62.5 | 65.4 |
| **EdgeSAM** | <u>67.5</u> | <u>76.1</u> | <u>79.0</u> | 48.0 | 61.8 | <u>68.7</u> | 53.7 | <u>63.4</u> | <u>67.7</u> |
| **EdgeSAM-RPN** | | | | 54.3 | | | | | |

# 2, Edge-SAM

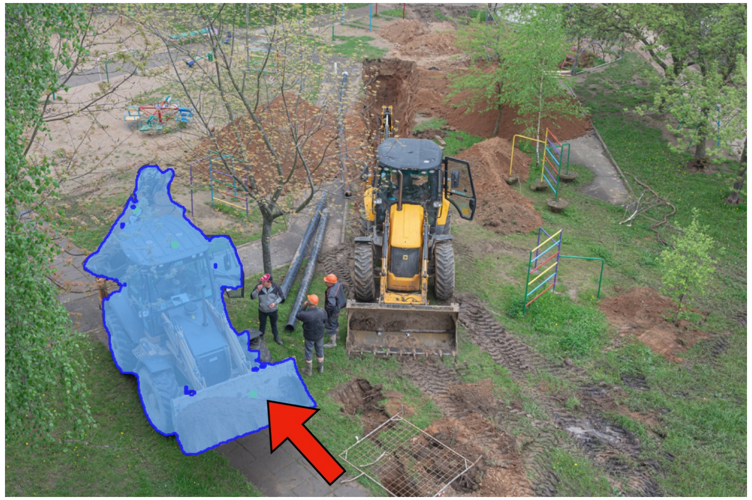
Without our KD

With our KD

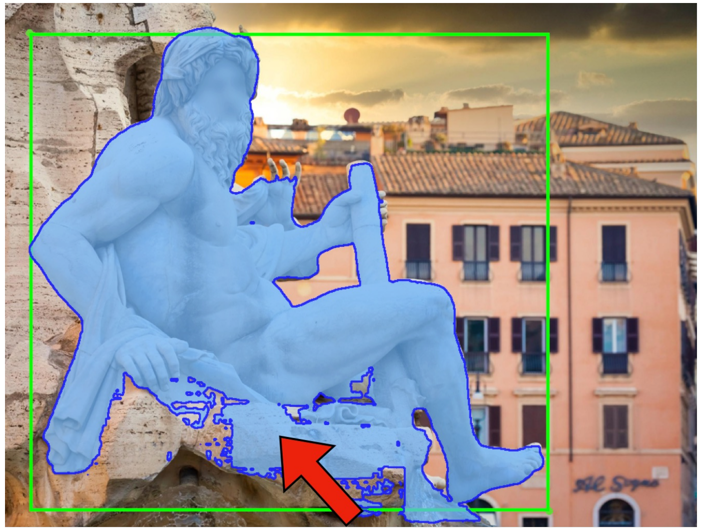
# 2, Edge-SAM



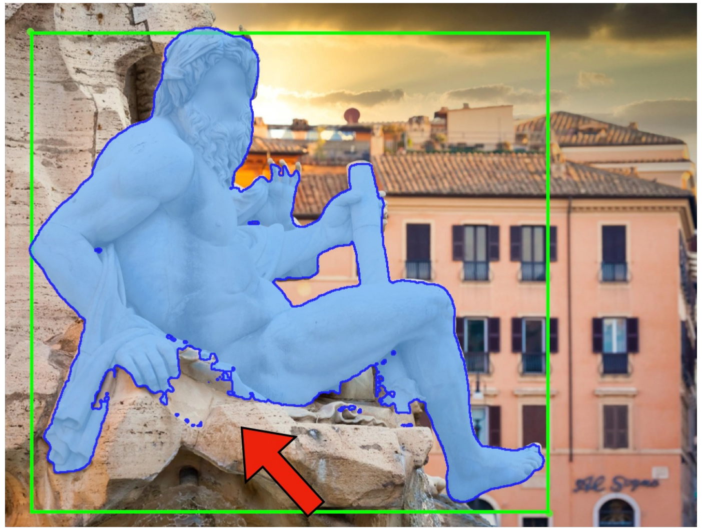
Without our KD



With our KD

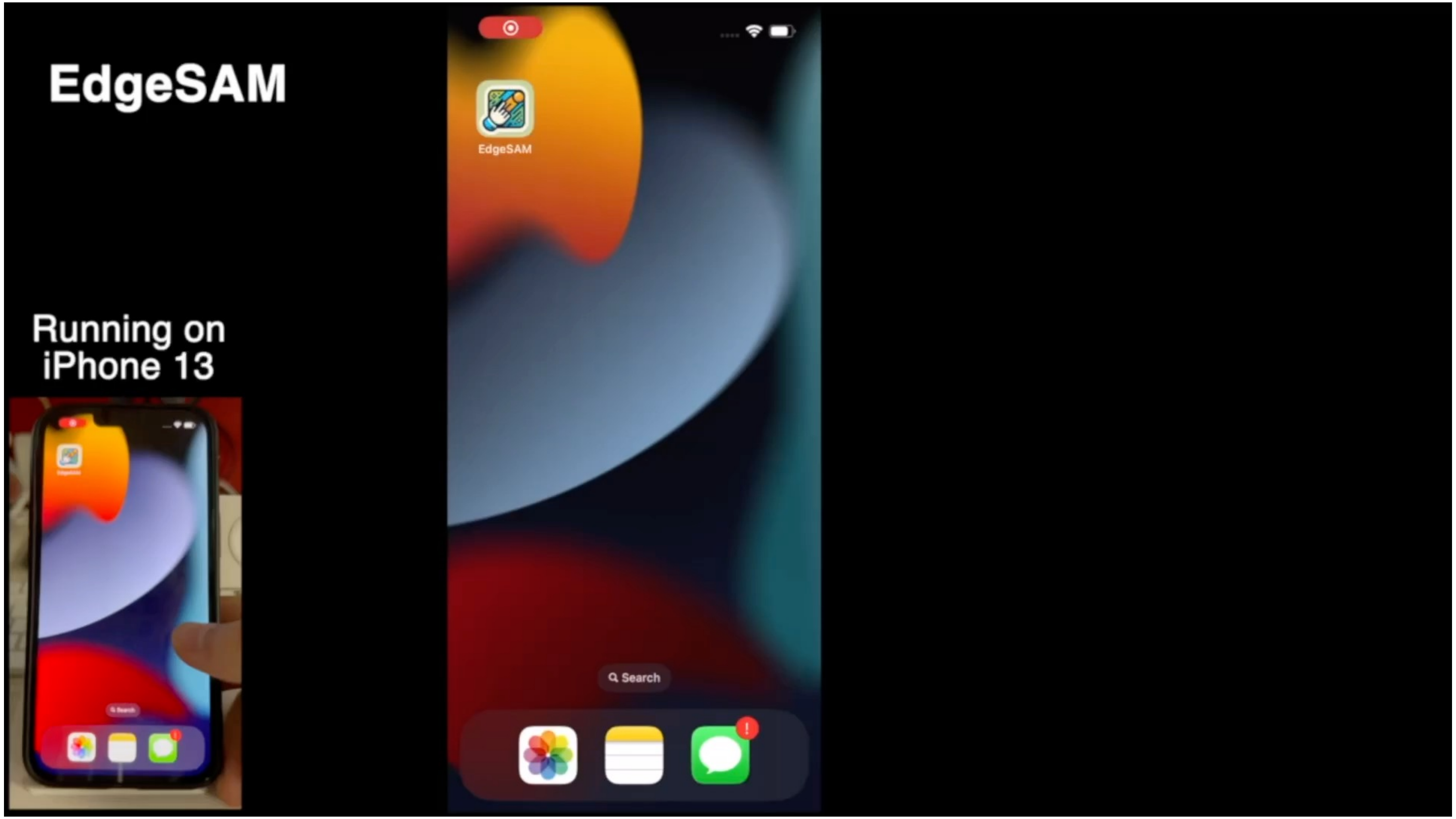


Without our KD



With our KD

# 2, Edge-SAM



# Outline

1, SAM overview.

2, Edge-SAM.

3, Open-Vocabulary SAM.

4, OMG-Seg.

5, Close Related Works and Summary.

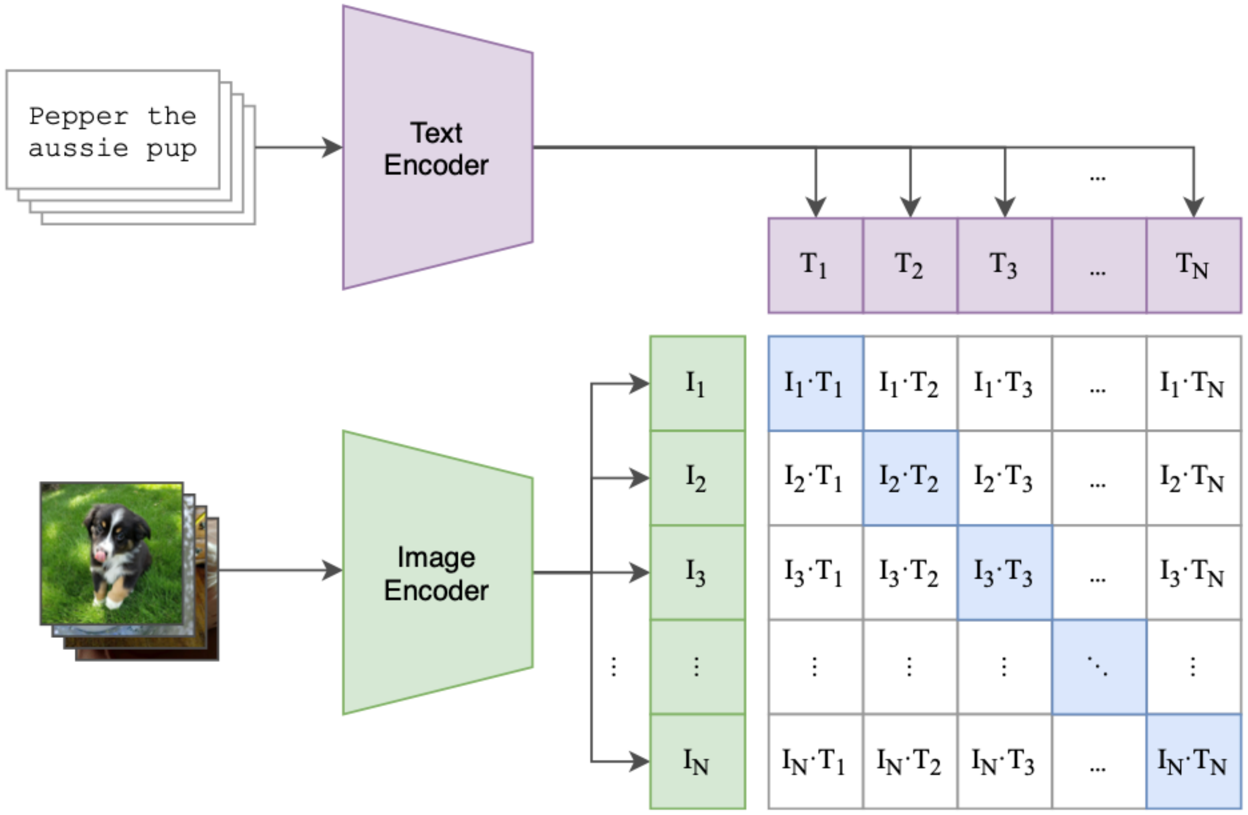
# 3, Open-Vocabulary SAM

CLIP: Learning Transferable Visual Models From Natural Language Supervision

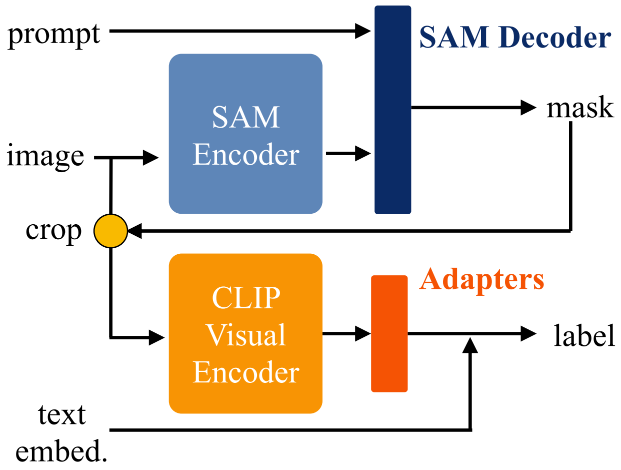
**SAM cannot recognize and label selected objects!**

**One Simple Solution:**

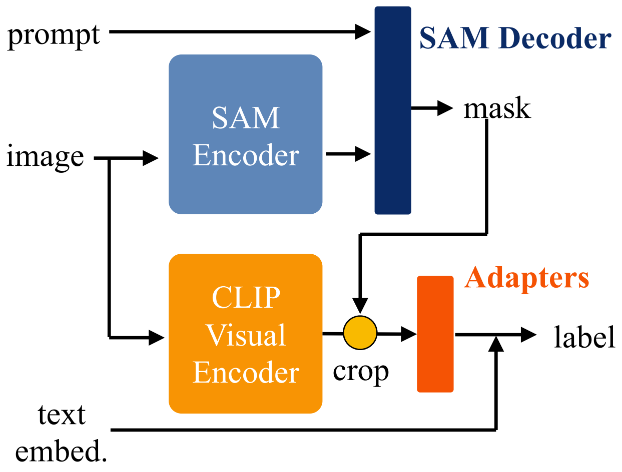
Combine VLMs, such as CLIP.



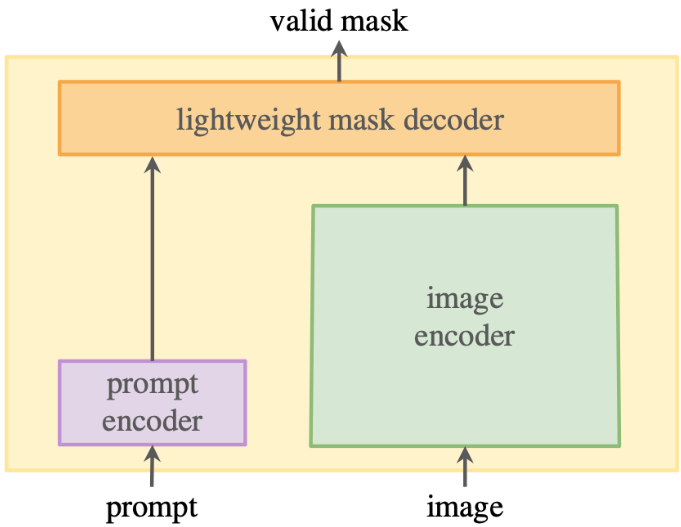
# 3, Open-Vocabulary SAM



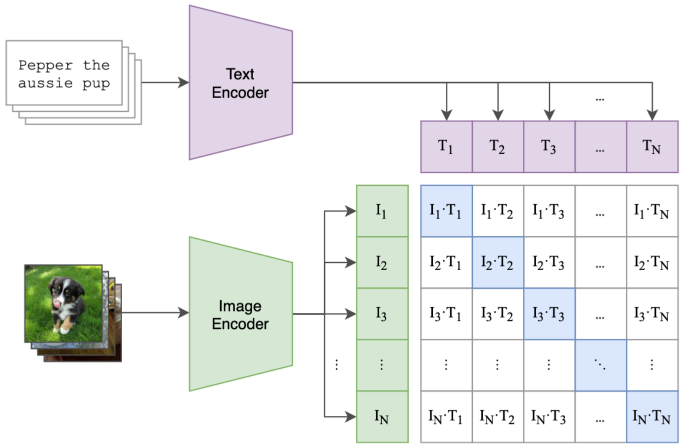

(a) Image Cropping Baseline




(a) Feature Cropping Baseline

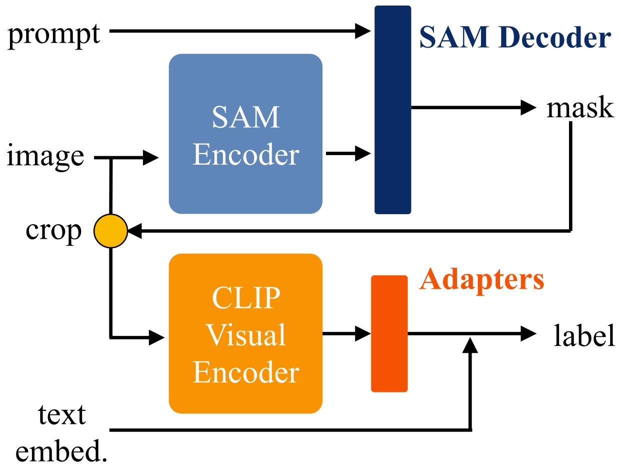



SAM

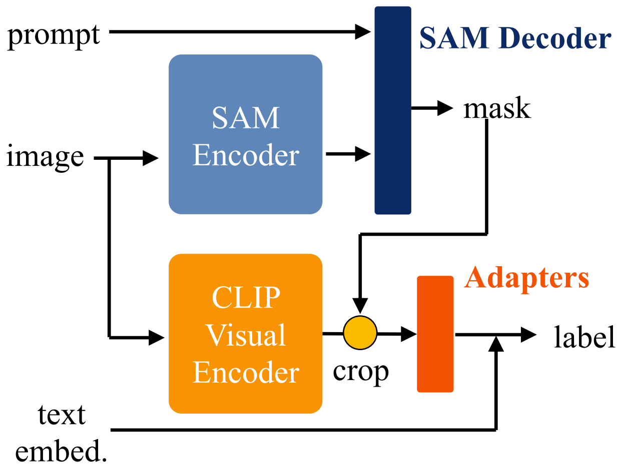



CLIP

# 3, Open-Vocabulary SAM




(a) Image Cropping Baseline




(a) Feature Cropping Baseline

## Problems:

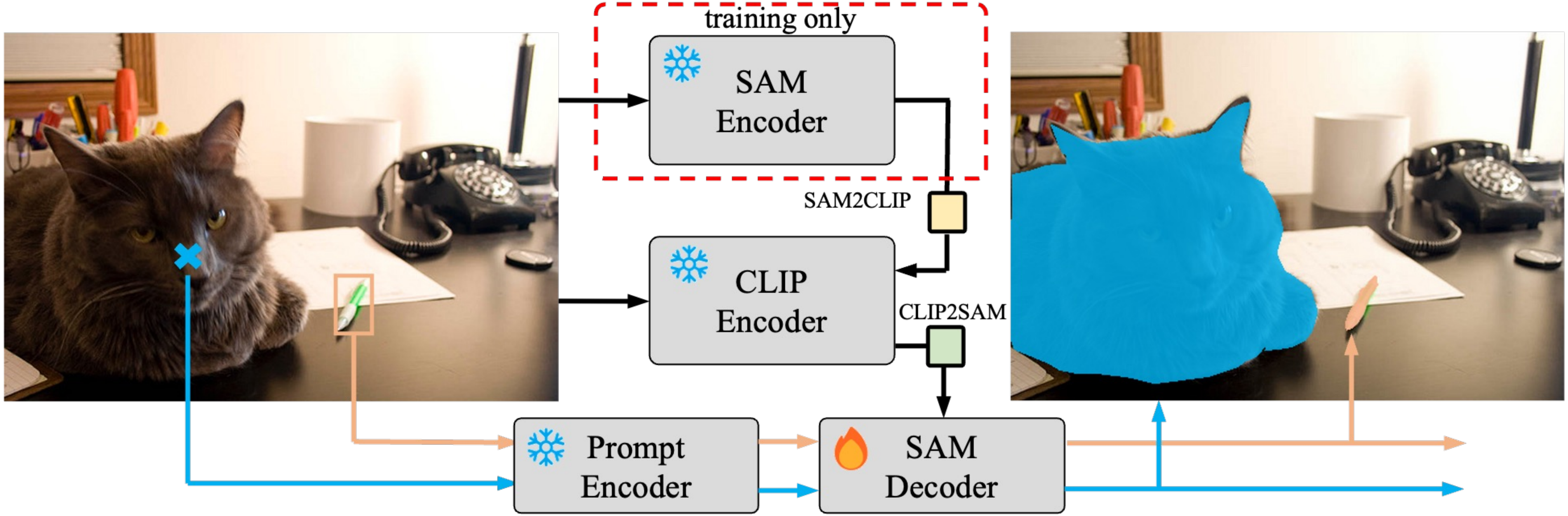
Problem-1： Two independent backbones. (Need extra costs).

Problem-2： Two different knowledge distribution. (CLIP vs SAM)

Problem-3： Smaller object recognition.

Problem-4： How to scale up the data using combined SAM and CLIP models?

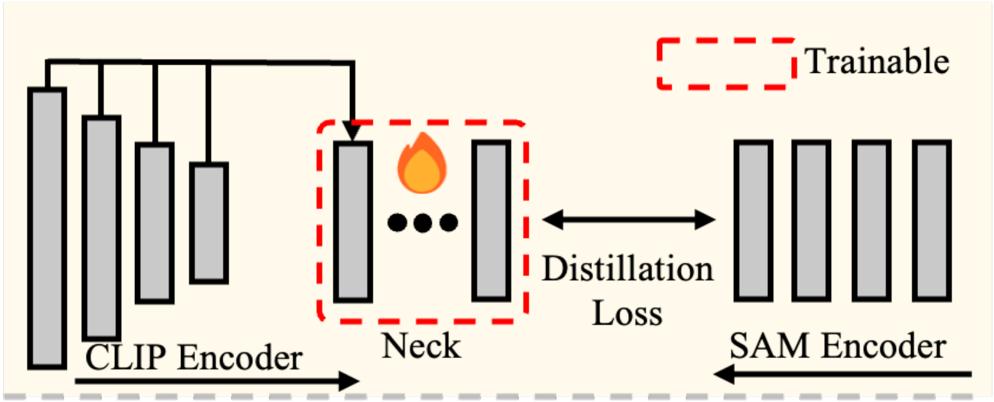
# 3, Open-Vocabulary SAM

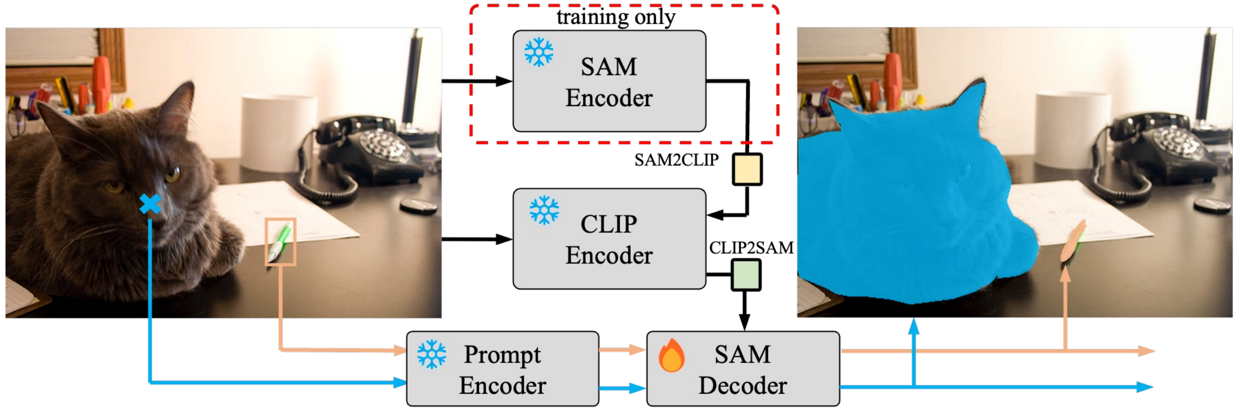



One visual backbone, using CLIP ->Problem-1

# 3, Open-Vocabulary SAM

SAM2CLIP: Transfer SAM Knowledge to CLIP

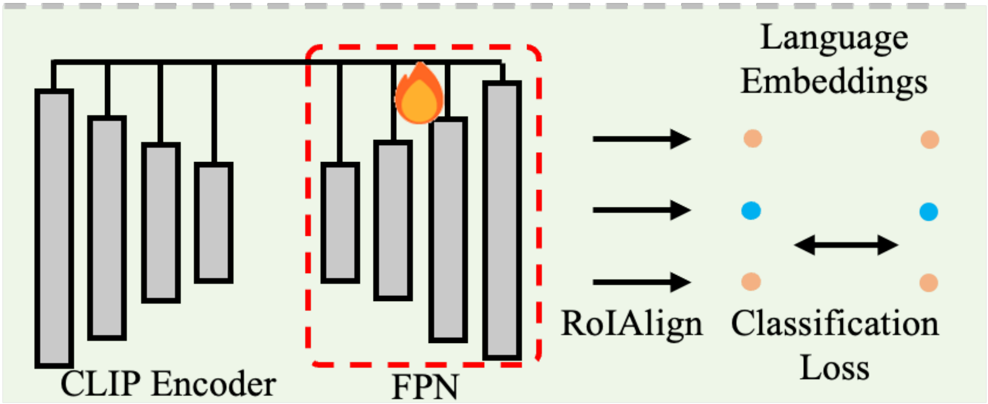


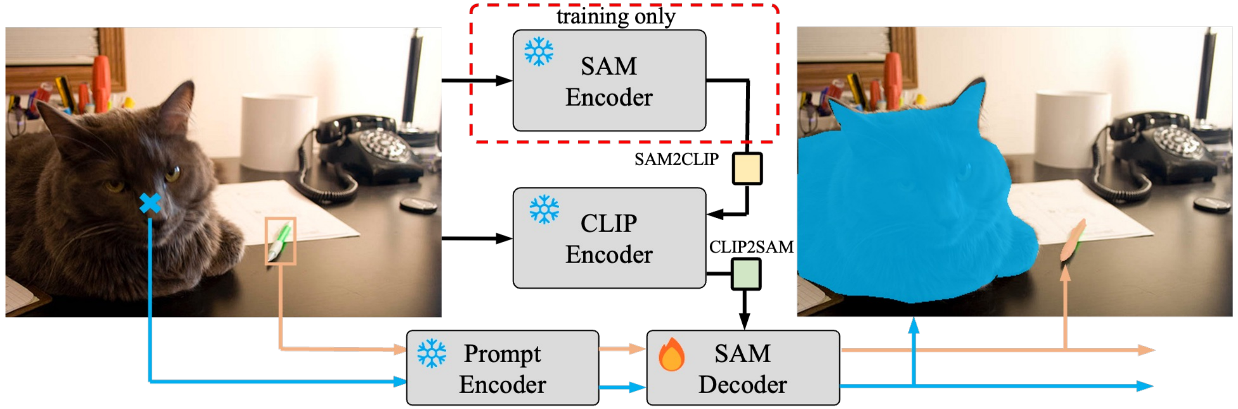




We explore one transformer architecture for adapting SAM's knowledge to CLIP. -> Problem-2

# 3, Open-Vocabulary SAM

CLIP2SAM: Transfer CLIP Knowledge to SAM decoder.



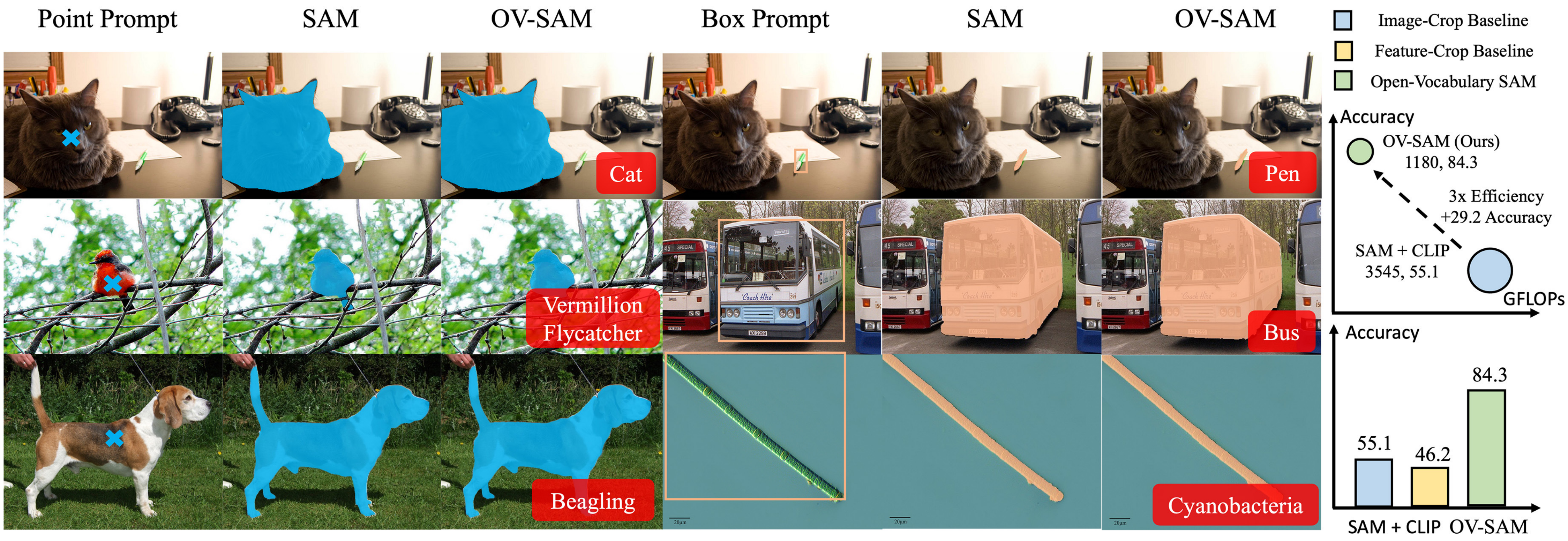




We explore combined CLIP features and RoI-Align operation. Problem-3.

We explore merged dataset co-training: imagenet,coco,lvis,v3det. Problem-4.

# 3, Open-Vocabulary SAM

Open-Vocabulary SAM performance



# 3, Open-Vocabulary SAM

| Method | Venue | Constrained | | Generalized | | |
|---|---|---|---|---|---|---|
| | | Base | Novel | Base | Novel | All |
| XPM [24] | CVPR'22 | 42.4 | 24.0 | 41.5 | 21.6 | 36.3 |
| MaskCLIP [13] | ICML'23 | 42.8 | 23.2 | 42.6 | 21.7 | 37.2 |
| MasQCLIP [74] | ICCV'23 | 40.9 | 30.1 | 40.7 | 28.4 | 37.5 |
| Open-Vocabulary SAM | (Ours) | 41.7 | 37.5 | 39.3 | 39.8 | 39.4 |

OVSAM vs Open-Vocabulary Segmentation Methods

| Method | Detectors | mAP | AP50 | AP75 | APS | APM | APL | #Params | FLOPs |
|---|---|---|---|---|---|---|---|---|---|
| SAM-Huge | Faster-RCNN (R50) | 35.6 | 54.9 | 38.4 | 17.2 | 39.1 | 51.4 | 641M | 3,001G |
| SAM-Huge (finetuned) | Faster-RCNN (R50) | 35.8 | 55.0 | 38.4 | 16.5 | 38.6 | 53.0 | 641M | 3,001G |
| Open-Vocabulary SAM | Faster-RCNN (R50) | 35.8 | 55.6 | 38.3 | 16.0 | 38.9 | 53.1 | 304M | 1,180G |
| SAM-Huge | Detic (swin-base) | 36.4 | 57.1 | 39.4 | 21.4 | 40.8 | 54.6 | 641M | 3,001G |
| SAM-Huge (finetuned) | Detic (swin-base) | 36.8 | 57.4 | 39.8 | 20.8 | 40.6 | 55.1 | 641M | 3,001G |
| Open-Vocabulary SAM | Detic (swin-base) | 36.7 | 57.2 | 39.7 | 20.7 | 40.8 | 54.9 | 304M | 1,180G |
| SAM-Huge | ViTDet (Huge) | 46.3 | 72.0 | 49.8 | 25.2 | 45.5 | 59.6 | 641M | 3,001G |
| SAM-Huge (finetuned) | ViTDet (Huge) | 46.5 | 72.3 | 50.3 | 25.2 | 45.8 | 60.1 | 641M | 3,001G |
| Open-Vocabulary SAM | ViTDet (Huge) | 48.8 | 73.8 | 52.9 | 24.8 | 46.3 | 64.2 | 304M | 1,180G |

OVSAM vs SAM

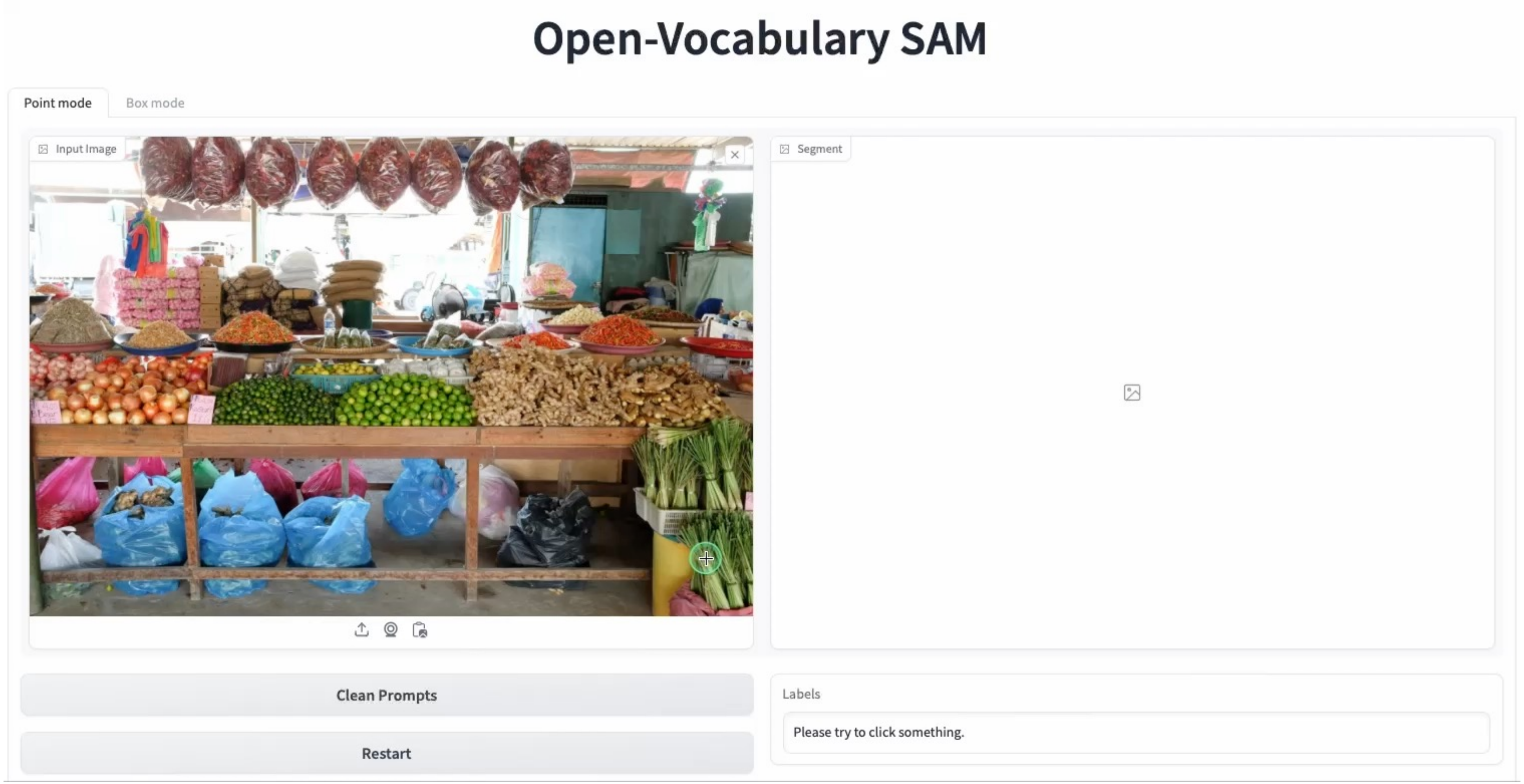
| Datasets | Accuracy (COCO) | #vocaulary | #images |
|---|---|---|---|
| LVIS | 83.1 | 1,203 | 99K |
| V3Det | 78.7 | 13,204 | 183K |
| I-21k | 44.5 | 19,167 | 13M |
| V3Det + LVIS | 82.7 | 13,844 | 282K |
| V3Det + LVIS + I-21k | 83.3 | 25,898 | 13M |
| V3Det + LVIS + I-21k + Object365 | 83.0 | 25,970 | 15M |

Scale-up Training

# 3, Open-Vocabulary SAM

# 3, Open-Vocabulary SAM

# 3, Open-Vocabulary SAM



# Outline

1, SAM overview.

2, Edge-SAM.

3, Open-Vocabulary SAM.

4, OMG-Seg.

5, Close Related Works and Summary.

# 4, OMG-Seg

# OMG-Seg: Is One Model Good Enough For All Segmentation?

## CVPR-2024

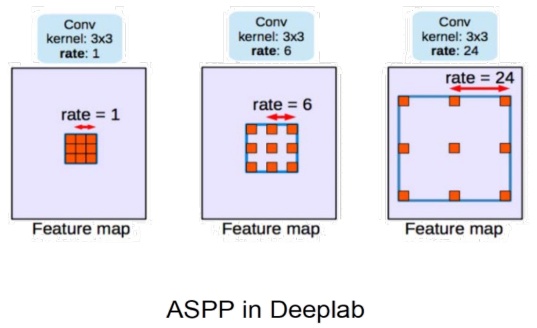
Xiangtai Li · Haobo Yuan . Wei Li . Henghui Ding · Size Wu · Wenwei Zhang ·
Yining Li . Kai Chen . Chen Change Loy*

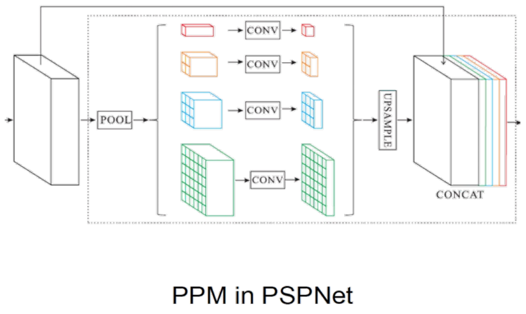
S-Lab, MMlab@NTU, Shanghai AI Laboratory

**A Baseline of One  Model For Segmentation Tasks.**

# 4, OMG-Seg

## Single Expert Models



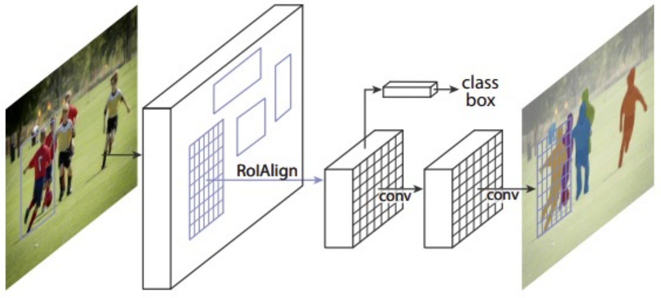
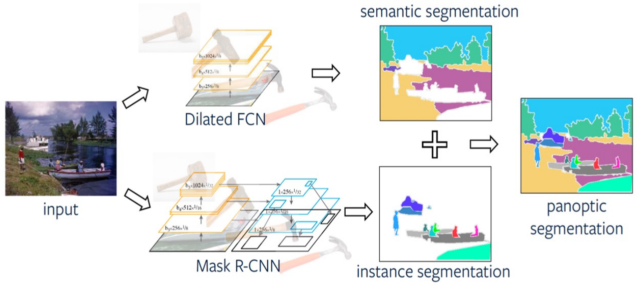



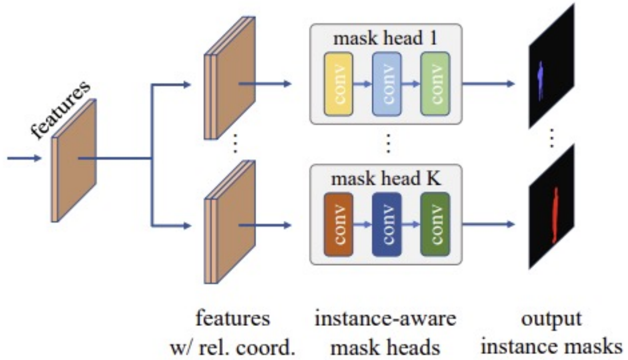

Figure 1. The **Mask R-CNN** framework for instance segmentation.

ASPP: Deeplab v3+ (ECCV-2018)

PPM: PSPNet (CVPR-2017)

Mask R-CNN-ICCV-2017



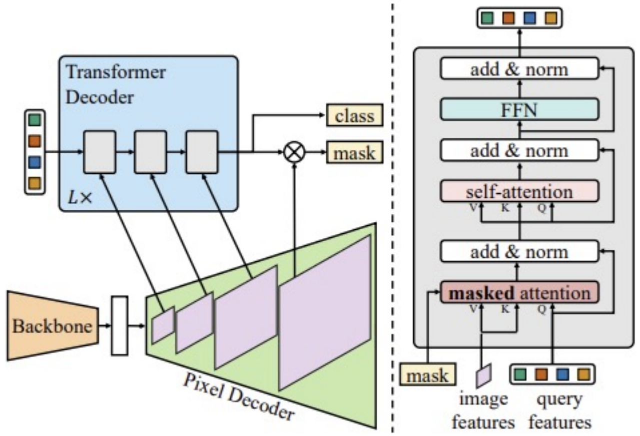




Panoptic Segmentation-CVPR-2019

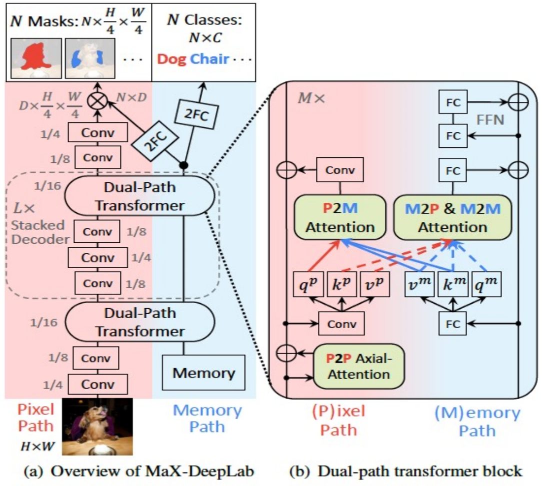
CondInst-ECCV-2020

# 4, OMG-Seg

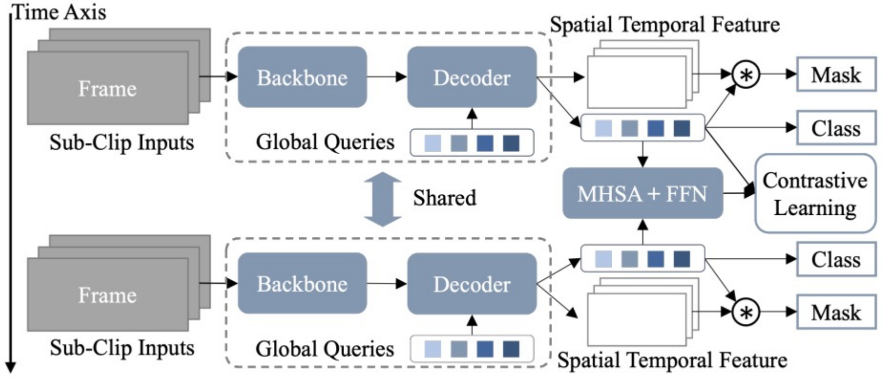
## Unified Models For Image and Video Segmentation




Mask2Former-CVPR-2022




Max-Deeplab-CVPR-2021




Tube-Link-ICCV-2023

# 4, OMG-Seg

## Partially Unified Models

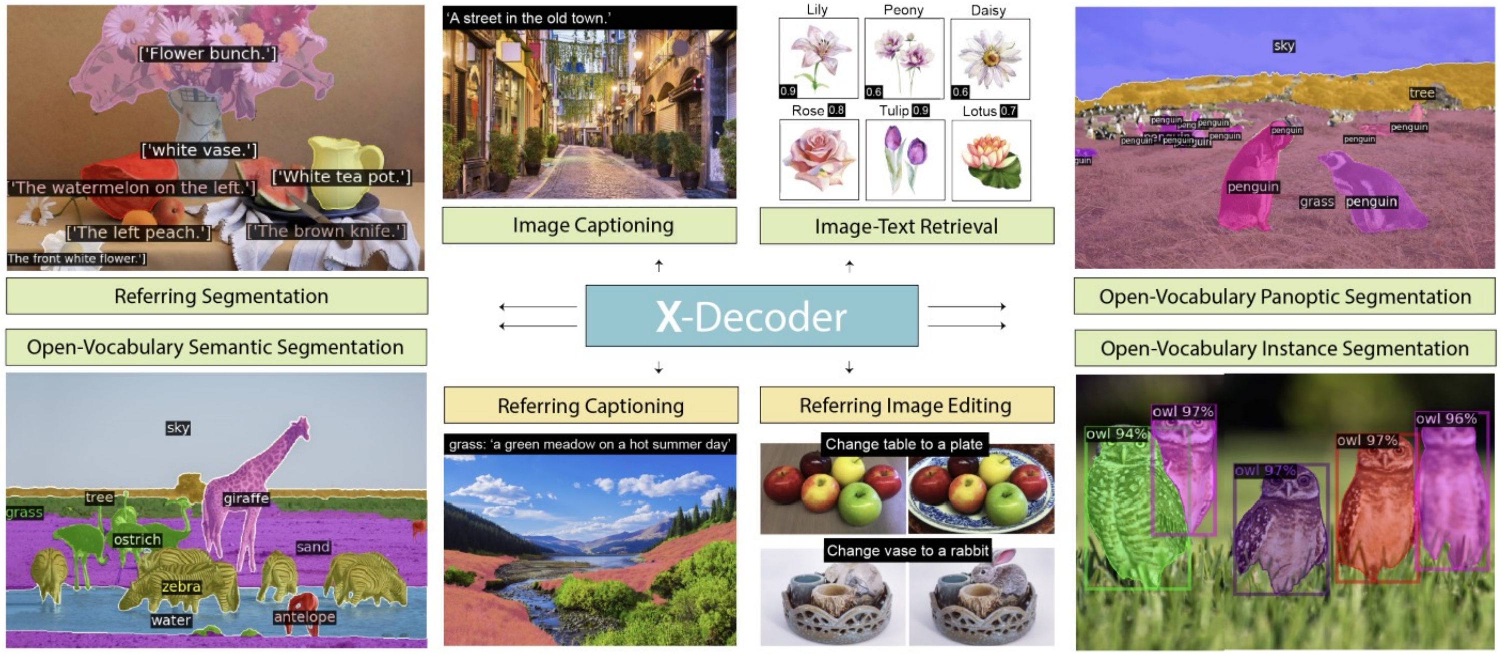



**Figure 1.** With one suite of parameters, X-Decoder after pretraining supports all types of image segmentation tasks ranging from open-vocabulary instance/semantic/panoptic segmentation to referring segmentation, and vision-language tasks including image-text retrieval, and image captioning (labeled in green boxes). It further empowers composite tasks like referring captioning using X-Decoder itself and image editing that combines with generative models such as Stable Diffusion [66] (labeled in yellow boxes).

### X-Decoder-CVPR-2023

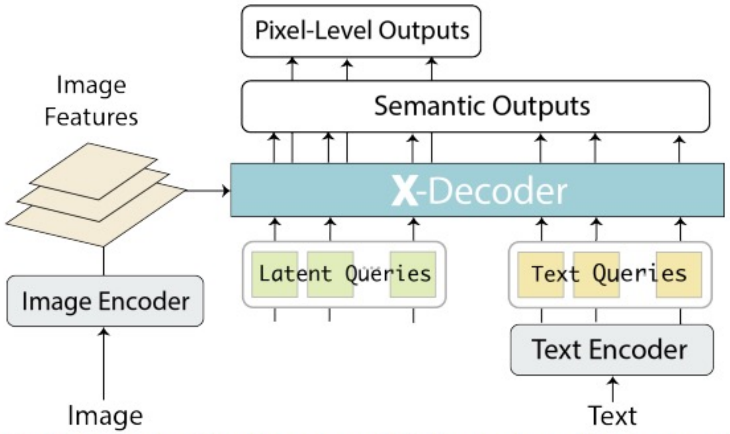



**Figure 2.** Overall pipeline for our model. It consists of an image encoder, a text encoder and our own designed X-Decoder.

# 4, OMG-Seg

## Partially Unified Models

Semantic-SAM, arxiv-23-7-10

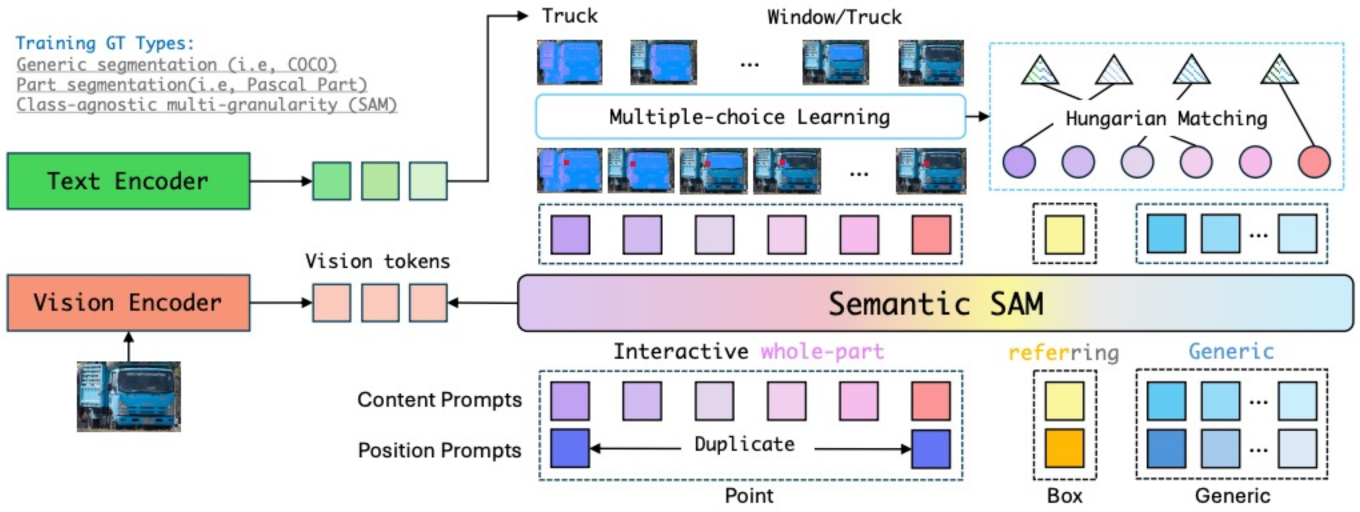



Figure 3: Semantic-SAM is a universal segmentation framework that can take multiple types of segmentation data including generic, part, and class-agnostic segmentation data. The Vision Encoder is used to extract image features. The mask decoder can do both generic segmentation and promptable segmentation with various types of prompts. For point and box, we input them via anchor boxes to the mask decoder. Since there is an ambiguity of granularity for a point input, we duplicate each point 6 times and give them different levels of embeddings. The output masks of point prompts match with multiple GT masks of different granularities.

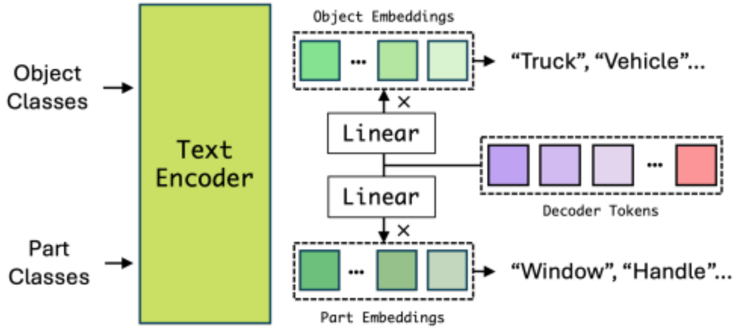



Figure 4: Decoupled object and part classification.

# 4, OMG-Seg

## Partially Unified Models

TarViS-CVPR-2023

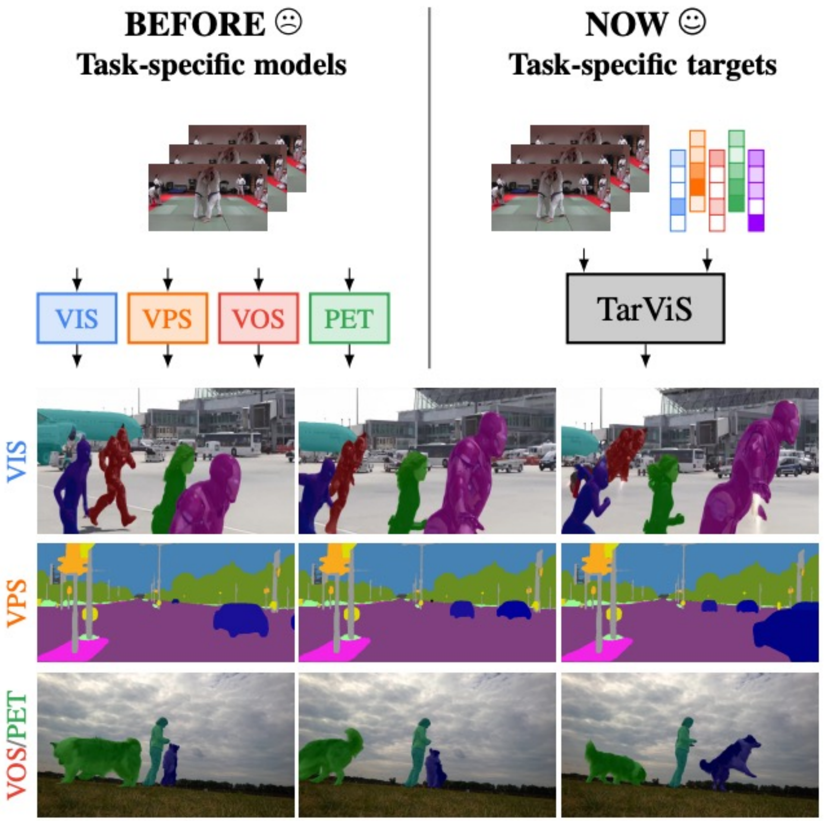



Figure 1. Predicted results from a jointly trained TarViS model for four different video segmentation tasks.

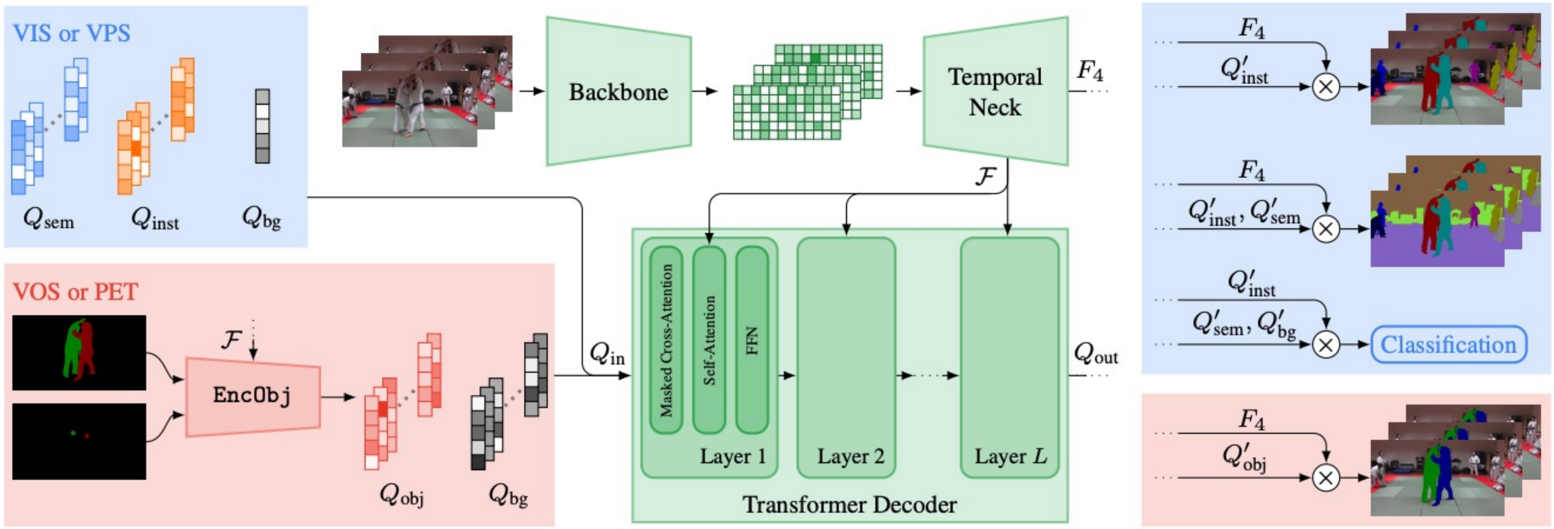
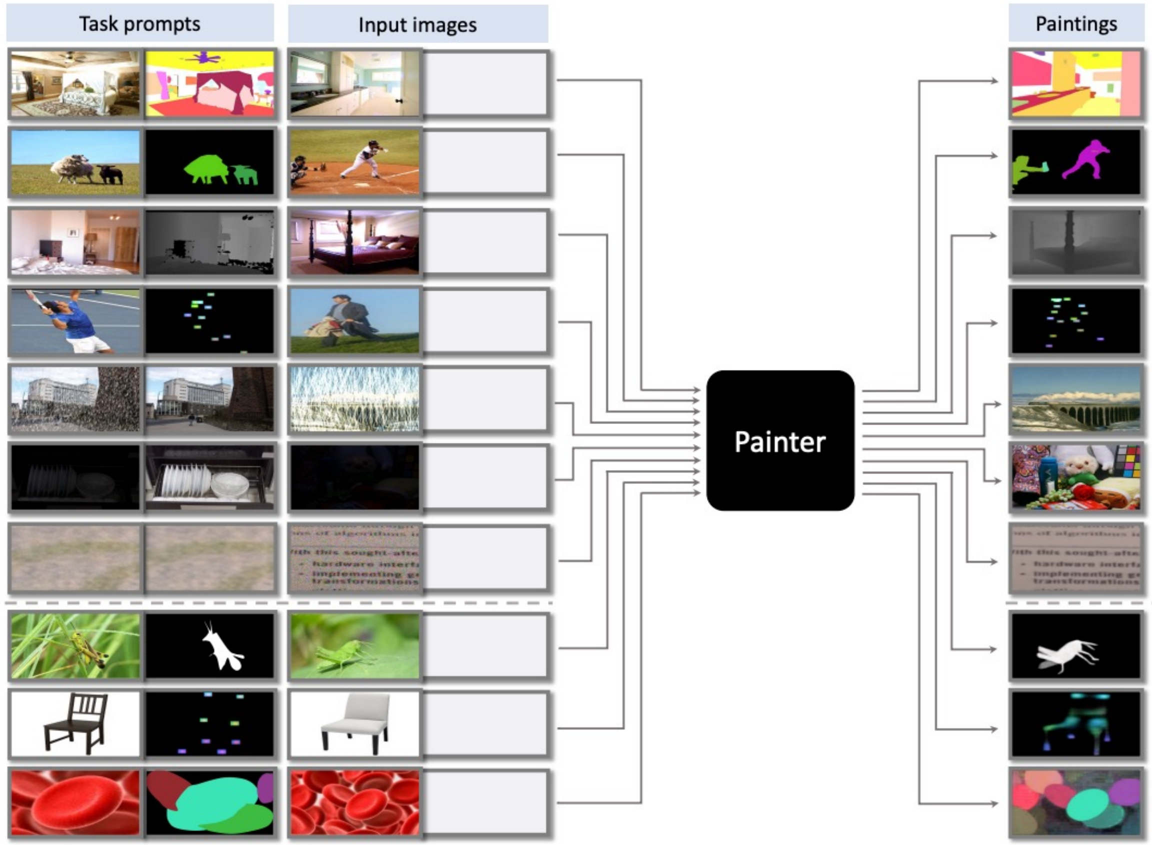



Figure 2. **TarViS Architecture.** Segmentation targets for different tasks are represented by a set of abstract target queries $Q_{in}$. The core network (in green) is agnostic to the task definitions. The inner product between the output queries $Q_{out}$ and video feature $F_4$ yields segmentation masks as required by the task.

# 4, OMG-Seg

## General Visual Models

Images Speak in Images: A Generalist Painter for In-Context Visual Learning, CVPR-2023



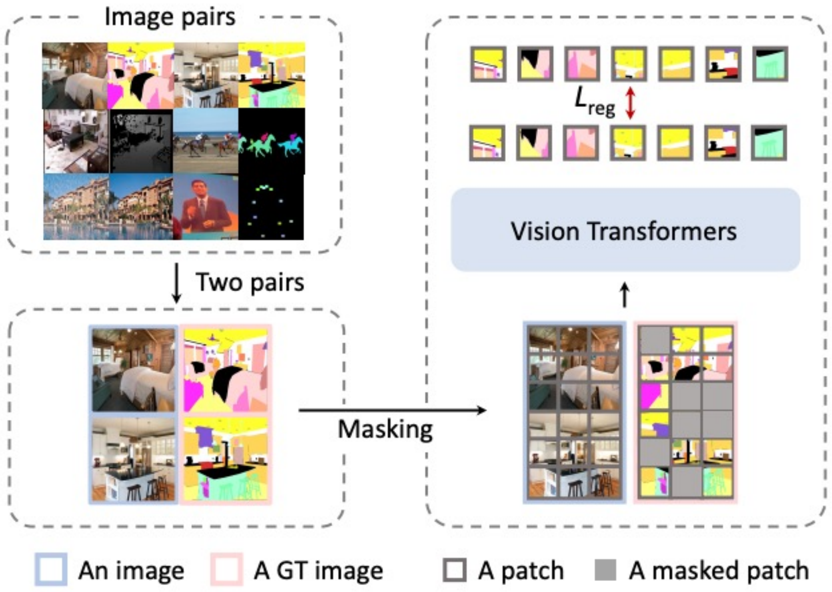




**Figure 2. The training pipeline of the masked image modeling (MIM) framework.**

# 4, OMG-Seg

**Summary**

1, Visual segmentation problems are traditionally tackled by distinct or partially unified models.

2, Unified Image / Video / Open-Vocabulary / Interactive Models are proposed, most of them are foundation models. No works combine them all.

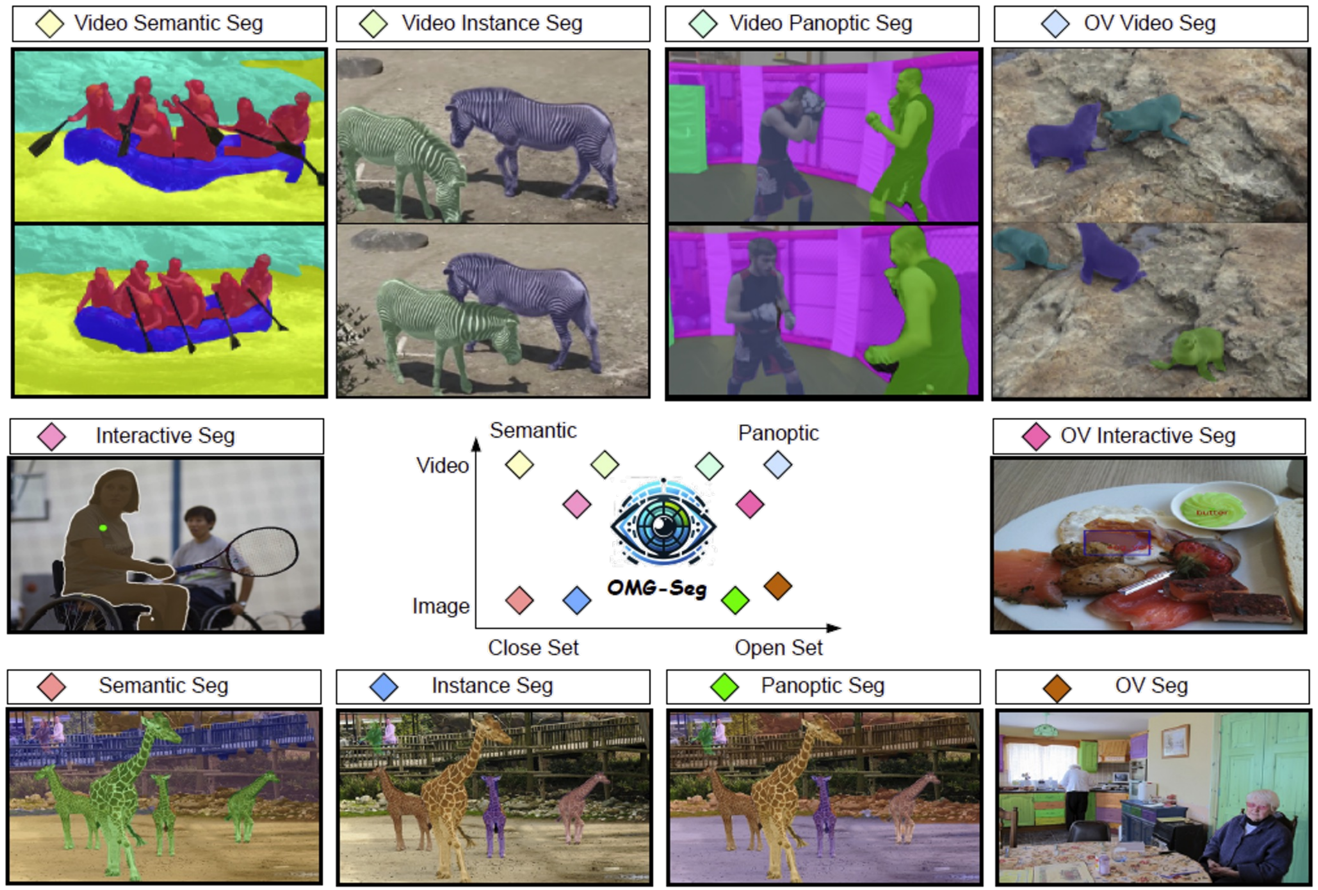
3, The performance gaps are large between vision generalist and segmentation experts.

4, Is there one model to solve all these task with extremely parameter saving and handcraft saving?

**OMG-Seg is all your need!!**

**The OMG-Seg logo is generated by DALLE-3.**

# 4, OMG-Seg




## Key Features:

- One shared model for all segmentation.

- Good enough performance on various segmentation tasks and datasets.

- Enable task association and sharing.

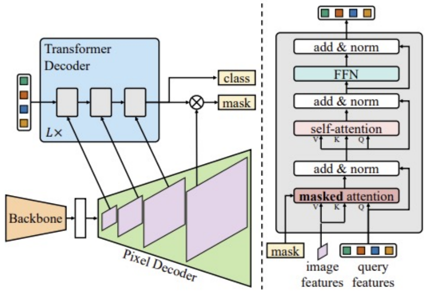
- Enable open-vocabulary and interactive segmentation.

- The first work to unify image, video, open-vocabulary and interactive segmentation in one share model.

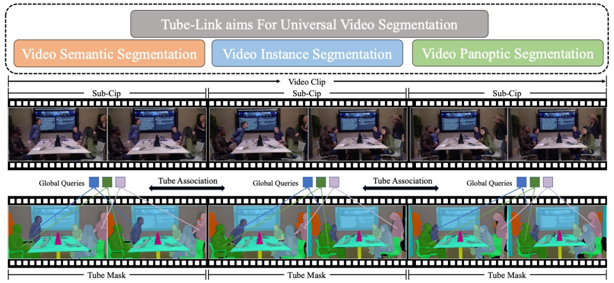
# 4, OMG-Seg

## How Do we perform Unified Task Representation

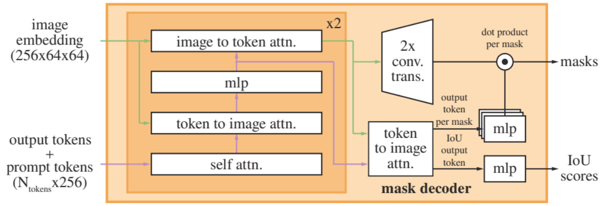
1, Image Segmentation:



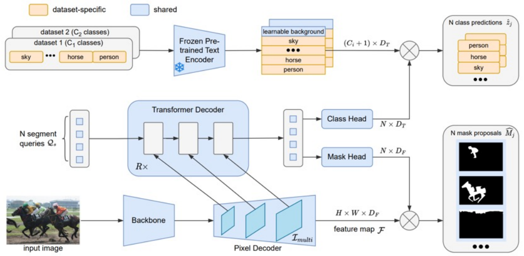

2, Video Segmentation:




3, Interactive Segmentation:.




4, Open-Vocabulary and Multi-Dataset Segmentation:



1, Decoder -> Cross Attention.

2, Query Representation -> Each Entity.

3, Classification -> Mask Classification.

4, Instance Matching -> Match Tube/Stuff/Thing Masks.

# 4, OMG-Seg

## Unified Task Representation

1, Image Segmentation: one query -> one mask and one label.

2, Video Segmentation: one query -> one tube mask, one tube label and one ID.

3, Interactive Segmentation: one visual prompt -> one query -> one mask.

4, Open-Vocabulary and Multi-Dataset Segmentation: replace the class label into CLIP text embedding and adopt frozen CLIP visual backbone.

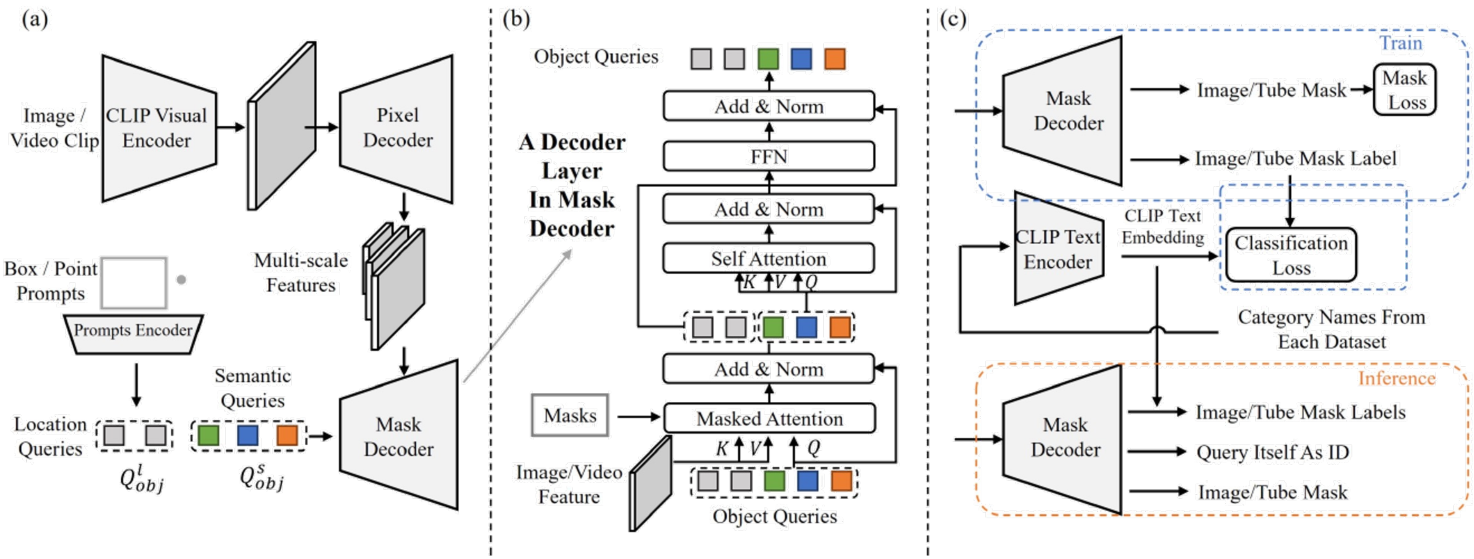
**Put them all together in one model!**

# 4, OMG-Seg

Table 1. Setting Comparison For Different Models. We include several representative methods here. Our proposed OMG-Seg can perform various segmentation tasks in one model.

| Methods | SS | IS | PS | VSS | VIS | VPS | VOS | Open-Set | Multi dataset training | Interactive | Shared model |
|---|---|---|---|---|---|---|---|---|---|---|---|
| DeeplabV3+ [11] | ✓ | | | | | | | | | | |
| MaskRCNN [30] | | ✓ | | | | | | | | | |
| PanopticFPN [36] | | | ✓ | | | | | | | | |
| DERT [6] | | | ✓ | | | | | | | | |
| DetectorRS [64] | | ✓ | ✓ | | | | | | | | |
| TCB [58] | | | | ✓ | | | | | | | |
| VisTR [78] | | | | | ✓ | | | | | | |
| VPSNet [34] | | | | | | ✓ | | | | | |
| STM [61] | | | | | | | ✓ | | | | |
| K-Net [97] | ✓ | ✓ | ✓ | | | | | | | | |
| Mask2Former [18] | ✓ | ✓ | ✓ | | | | | | | | |
| Video K-Net [51] | | | | ✓ | ✓ | ✓ | | | | | |
| Tube-Link [49] | | | | ✓ | ✓ | ✓ | | | | | |
| TubeFormer [35] | | | | ✓ | ✓ | ✓ | | | | | |
| OneFormer [33] | ✓ | ✓ | ✓ | | | | | | | | ✓ |
| TarViS [2] | | | | ✓ | ✓ | ✓ | ✓ | | | | ✓ |
| MSeg [40] | ✓ | | | | | | | | ✓ | | ✓ |
| UNINEXT [89] | | ✓ | | | ✓ | | ✓ | | ✓ | | |
| OpenSeg [26] | ✓ | | | | | | | ✓ | ✓ | | ✓ |
| SAM [38] | | | | | | | | ✓ | | ✓ | |
| Semantic-SAM [42] | ✓ | ✓ | ✓ | | | | | ✓ | ✓ | ✓ | ✓ |
| SEEM [104] | ✓ | ✓ | ✓ | | | | | ✓ | ✓ | ✓ | ✓ |
| OPSNet [14] | | | ✓ | | | | | ✓ | | | |
| FreeSeg [65] | ✓ | ✓ | ✓ | | | | | ✓ | | | ✓ |
| **OMG-Seg** | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

# 4, OMG-Seg



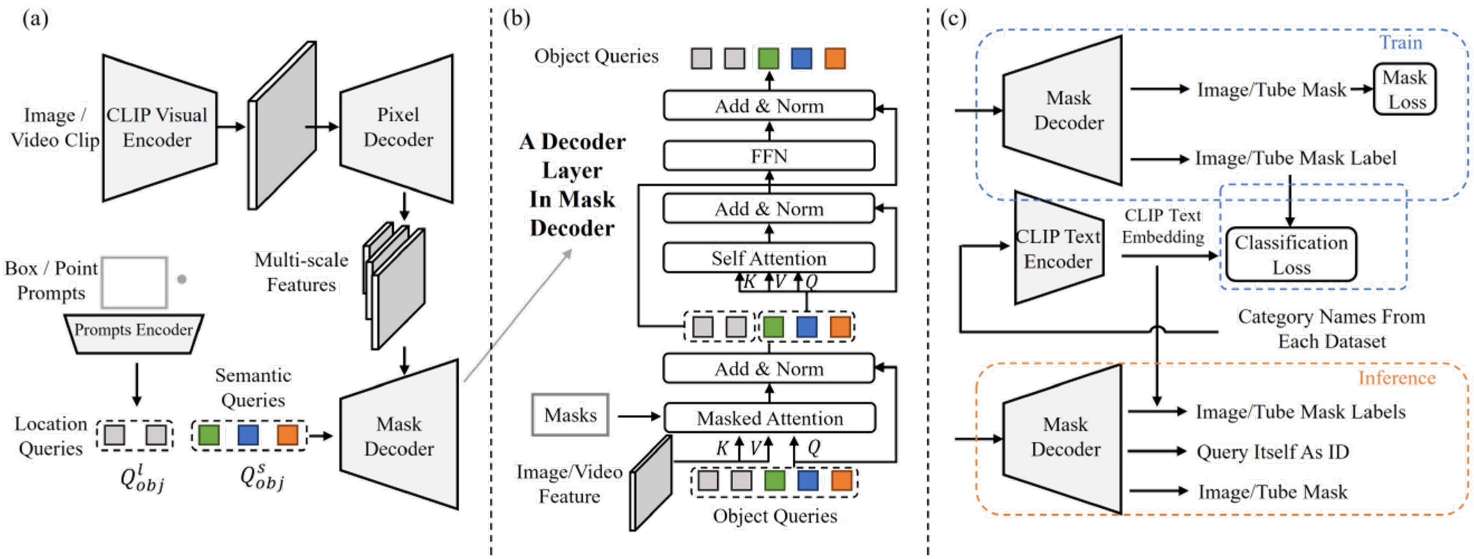

1, Simple Encoder and Pixel Decoder as Mask2Former.

2, Adopt the frozen CLIP backbone.

3, Location Queries and Semantic Queries are used as input of shared decoder.

4, The Decoder decode image masks, tube masks, binary masks, and image labels according to the queries' tasks.

# 4, OMG-Seg



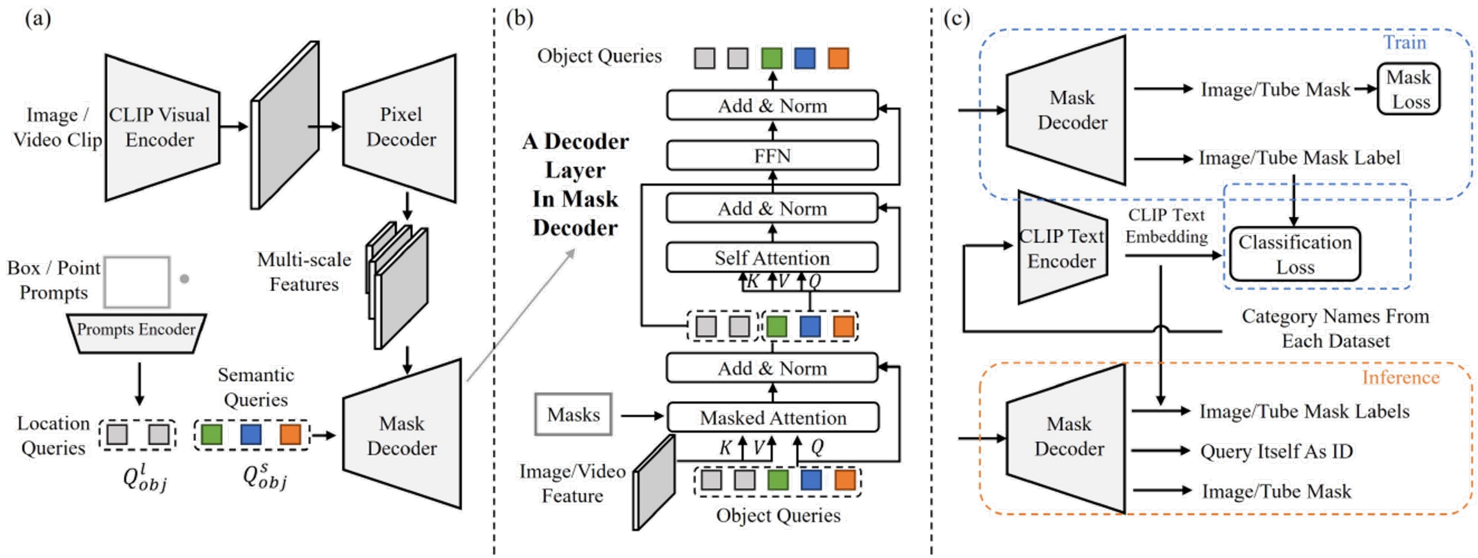

5, The CLIP text embeddings are used to supervise the classification to avoid label conflicts.

6, For open-vocabulary inference, we also adopt fused CLIP visual features score and learned text score to achieve better zero-shot classification.

7, For video instance, we adopt query embedding itself for association.

8, For interactive mode, we directly output binary mask as SAM.

# 4, OMG-Seg




## Training:

We jointly co-training the image/video data in one shot. Two mages are treated as one clips.

## Inference:

We inference each task and datasets according to different prompts and class embeddings.

# 4, OMG-Seg

Table 2. Experiment results of OMG-Seg on image, video, open-vocabulary, and SAM-like settings. * denotes models are pre-trained on the Object365 dataset [68]. We only list representative methods due to the page limit. Refer to the supplementary material for more methods. Our results are the averaged results of five different experiments.

| Methods | Backbone | COCO-PS PQ | Cityscapes-PS PQ | COCO-IS mAP | VIPSeg-VPS VPQ | YT-VIS-19 mAP | YT-VIS-21-OV mAP | ADE-OV PQ | DAVIS-17-VOS-OV J&F | COCO-SAM mIoU | Share Model - |
|---|---|---|---|---|---|---|---|---|---|---|---|
| DetectorRS [64] | ResNet50 | - | - | 42.1 | - | - | - | - | - | - | - |
| HTC [8] | ResNet50 | - | - | 38.4 | - | - | - | - | - | - | - |
| STM [62] | ResNet101 | - | - | - | - | - | - | - | 79.2 | - | - |
| K-Net [97] | ResNet50 | 47.1 | - | 38.6 | - | - | - | - | - | - | - |
| Mask2Former [18] | ResNet50 | 51.9 | 62.1 | 43.7 | - | - | - | - | - | - | - |
| Mask2Former [18] | Swin-Large | 57.8 | 66.6 | 50.1 | - | - | - | - | - | - | - |
| k-Max Deeplab [92] | ResNet50 | 53.0 | 64.3 | - | - | - | - | - | - | - | - |
| k-Max Deeplab [92] | ConvNeXt-Large | 58.1 | 68.4 | - | - | - | - | - | - | - | - |
| SeqFormer [80] | ResNet50 | - | - | - | - | 47.4 | - | - | - | - | - |
| IDOL [83] | Swin-Large | - | - | - | - | 64.3 | - | - | - | - | - |
| MinVIS [31] | Swin-Large | - | - | - | - | 61.6 | - | - | - | - | - |
| Video K-Net [51] | ResNet50 | - | - | - | 26.1 | 40.5 | - | - | - | - | - |
| Tube-Link [49] | ResNet50 | - | - | - | 41.2 | 52.8 | - | - | - | - | - |
| Tube-Link [49] | Swin-base | - | - | - | 54.5 | - | - | - | - | - | - |
| OneFormer [33] | Swin-Large | 58.0 | 67.2 | 49.2 | - | - | - | - | - | - | ✓ |
| TarViS [2] | Swin-Large | - | - | - | 48.0 | - | - | - | - | - | ✓ |
| fc-clip [91] | ConvNeXt-Large | 54.4 | - | 44.6 | - | - | - | 26.8 | - | - | ✓ |
| ODISE [86] | ViT-Large | 55.4 | - | 46.0 | - | - | - | 22.6 | - | - | ✓ |
| DaTaSeg [27] | ViT-L | 53.5 | - | - | - | - | - | - | - | - | ✓ |
| X-Decoder [103] | DaViT | 56.9 | - | 46.7 | - | - | - | 21.8 | - | - | ✓ |
| SEEM [104] * | DaViT | 57.5 | - | 47.7 | - | - | - | - | 58.9 | 83.4 | ✓ |
| UNINEXT [89] * | ConvNeXt-L | - | - | 49.6 | - | 64.3 | - | - | 77.2 | - | ✓ |
| HIPIE [75] * | ViT-H | 58.0 | - | 51.9 | - | - | - | 20.6 | - | - | ✓ |
| OpenSeeD [96] * | Swin-L | 59.5 | - | 53.2 | - | - | - | 19.7 | - | - | ✓ |
| SAM [38] | ViT-H | - | - | - | - | - | - | - | - | 55.3 | ✓ |
| Semantic-SAM [42] | Swin-T | 55.2 | - | 47.4 | - | - | - | - | - | 53.0 | ✓ |
| Painter [76] | ViT-L | 43.4 | - | - | - | - | - | - | - | - | ✓ |
| OMG-Seg | ConvNeXt-Large (frozen) | 53.8 | 65.7 | 44.5 | 49.8 | 56.4 | 50.5 | 27.9 | 74.3 | 58.0 | ✓ |
| OMG-Seg | ConvNeXt-XX-Large (frozen) | 55.4 | 65.3 | 46.5 | 53.1 | 60.3 | 55.2 | 27.8 | 76.9 | 59.3 | ✓ |

# 4, OMG-Seg

Table 3. Experiment results of OMG-Seg on multiple dataset settings. We use five different datasets for balanced joint co-training for only 12 epochs. We also implement compared baselines in the same codebase.

| Methods / Settings | Backbone | COCO-PS | COCO-IS | ADE-PS | VIPSeg-VPS | YT-VIS-19 | YT-VIS-21 | Params(M) | Share Model |
|---|---|---|---|---|---|---|---|---|---|
| K-Net [97] | ConvNeXt-Large (trained) | 50.5 | 42.3 | 40.2 | - | - | - | | - |
| Mask2Former [18] | ConvNeXt-Large (trained) | 53.2 | 45.2 | 43.2 | - | - | - | | - |
| Mask2Former-VIS [16] | ConvNeXt-Large (trained) | - | - | - | - | 45.8 | 42.3 | | |
| single dataset baseline | ConvNeXt-Large (frozen) | 52.5 | 45.6 | 41.2 | 42.3 | 45.3 | 44.3 | 1326 | - |
| OMG-Seg | ConvNeXt-Large (frozen) | 52.9 | 44.3 | 28.2 | 46.9 | 48.8 | 46.2 | 221 | ✓ |
| OMG-Seg | ConvNeXt-Large (trained) | 55.0 | 45.3 | 36.8 | 45.8 | 47.2 | 45.2 | 221 | ✓ |

Table 4. Ablation on joint co-training. (a), COCO-PS. (b), VIPSeg-VPS. (c). YT-VIS-19.

| Setting | COCO-PS | VIPSeg-VPS | YT-VIS-19 | ADE-OV | YT-VIS-21-OV |
|---|---|---|---|---|---|
| a | 53.4 | 32.2 | 34.2 | 25.5 | 30.3 |
| a + b | 52.9 | 49.0 | 45.2 | 26.2 | 39.6 |
| a + b + c | 53.0 | 48.5 | 56.8 | 26.1 | 50.3 |

Table 5. Ablation on shared decoder design.

| Setting | COCO-PS | VIPSeg-VPS | Param | GFlops |
|---|---|---|---|---|
| shared | 53.0 | 48.5 | 221 | 868 |
| decoupled image/video | 53.6 | 46.2 | 243 | 868 |

# 4, OMG-Seg

Panoptic Segmentation

COCO-dataset

Interactive Segmentation

COCO-dataset

Video Instance Segmentation

Youtube-VIS-2019-dataset

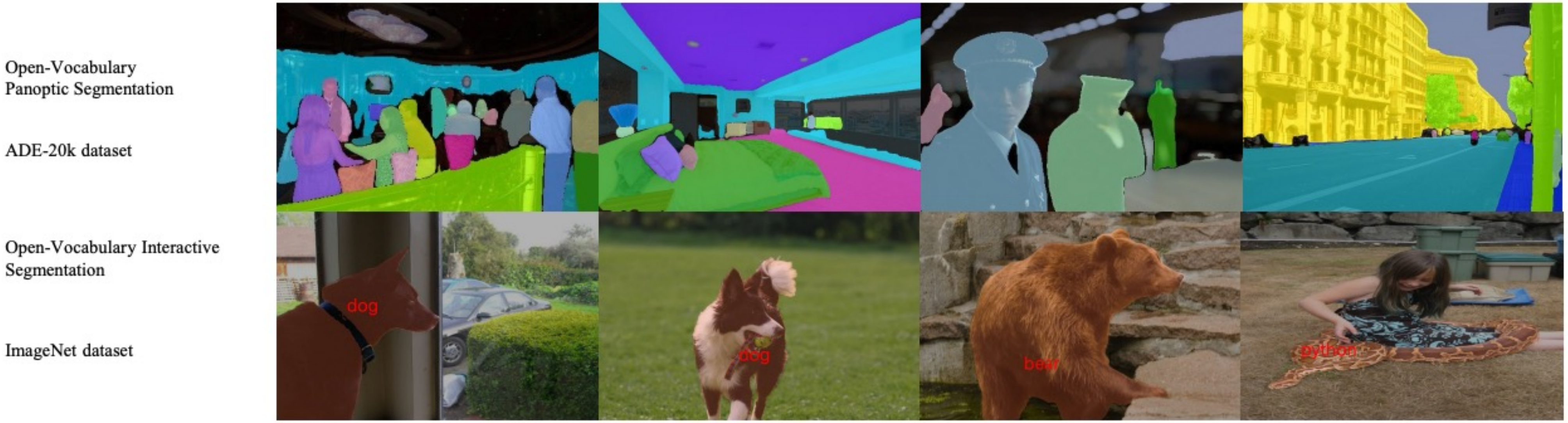
Video Panpotic Segmentation

VIP-Seg-dataset

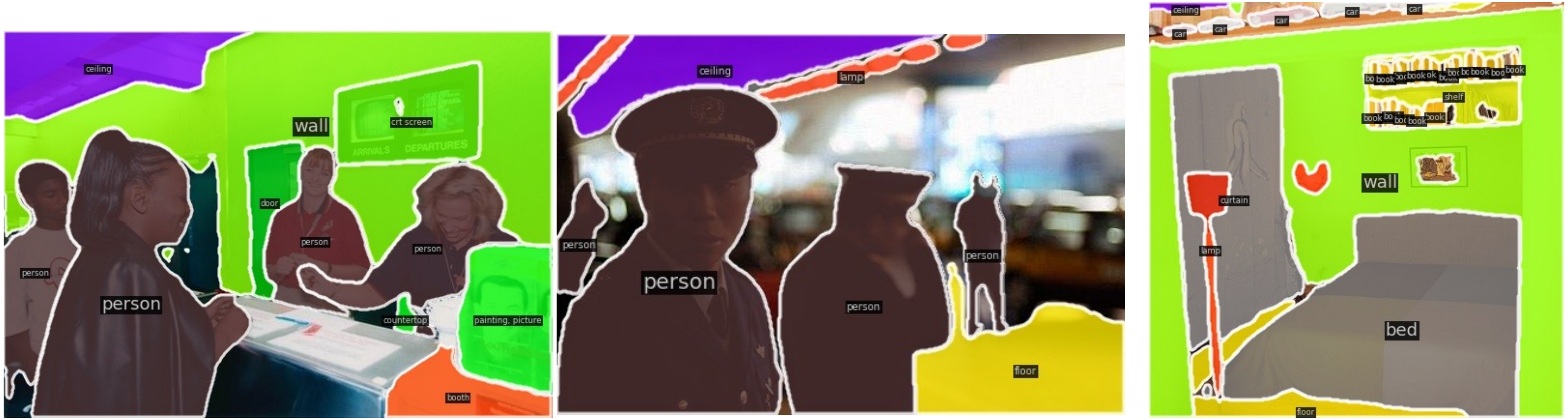
Open-Vocabulary Video Instance Segmentation

Youtube-VIS-2021 dataset

# 4, OMG-Seg



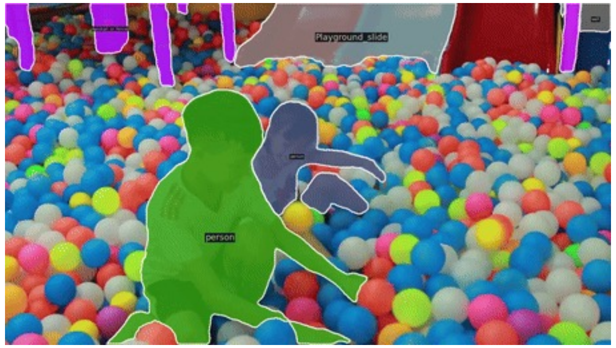
Open-Vocabulary
Panoptic Segmentation
ADE-20k dataset
Open-Vocabulary Interactive
Segmentation
ImageNet dataset
dog
dog
bear
python

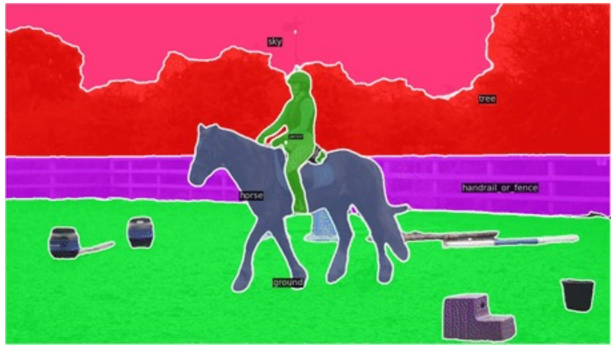
ceiling
wall
crt screen
door
person
person
person
person
countertop
painting, picture
booth
ceiling
lamp
person
person
person
person
floor
ceiling
car
shelf
wall
curtain
lamp
bed
floor

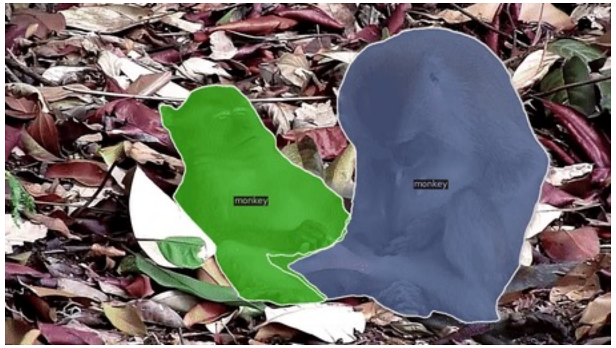
# 4, OMG-Seg

Video Demo.





# Outline

1, SAM overview.

2, Edge-SAM.

3, Open-Vocabulary SAM.

4, OMG-Seg.

5, Close Related Works and Summary.

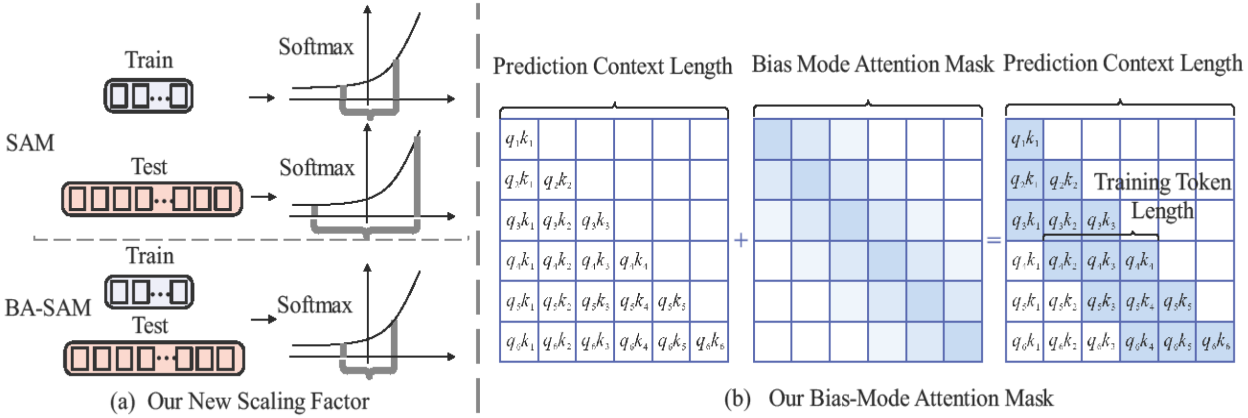
# 5, Close Related Works.

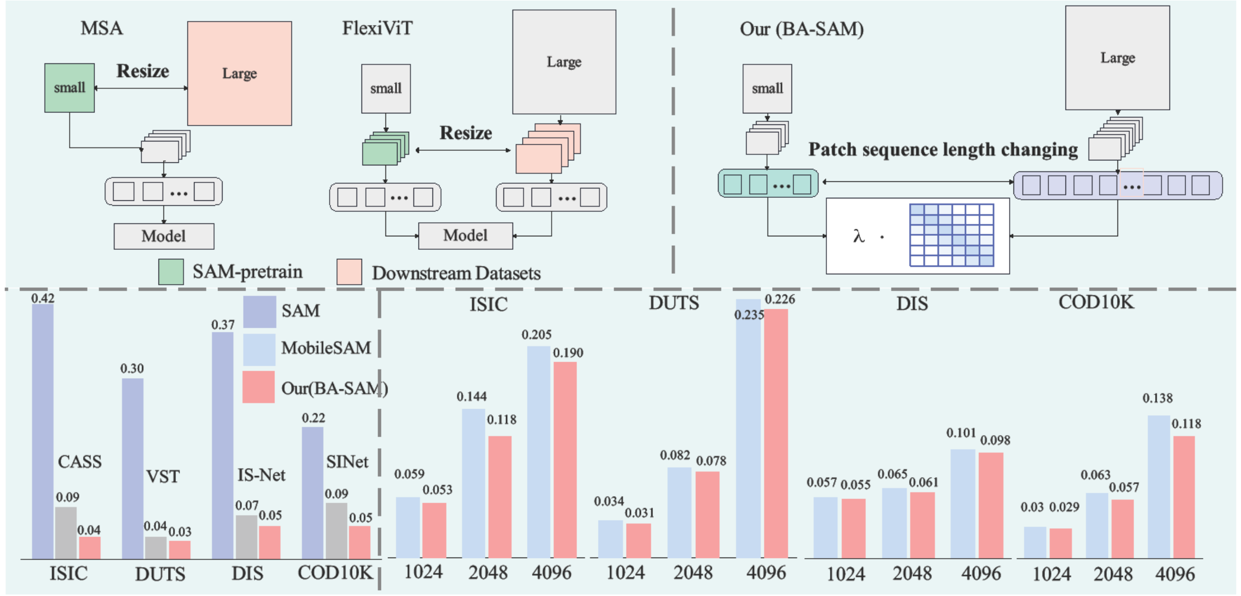
BA-SAM: Scalable Bias-Mode Attention Mask for Segment Anything Model

Yiran Song[1]*, Qianyu Zhou[1]*, Xiangtai Li[2], Deng-Ping Fan[3], Xuequan Lu[4]†, Lizhuang Ma[1]†
[1]Shanghai Jiao Tong University; [2] Nanyang Technological University;
[3] Nankai University; [4] La Trobe University
[1]{songyiran,zhouqianyu,lzma}@sjtu.edu.cn,
[2]xiangtai94@gmail.com, [3]dengpfan@gmail.com [4]xuequan.lu@deakin.edu.au

CVPR-2024

SAM: scale problems on high-resolution segmentation.

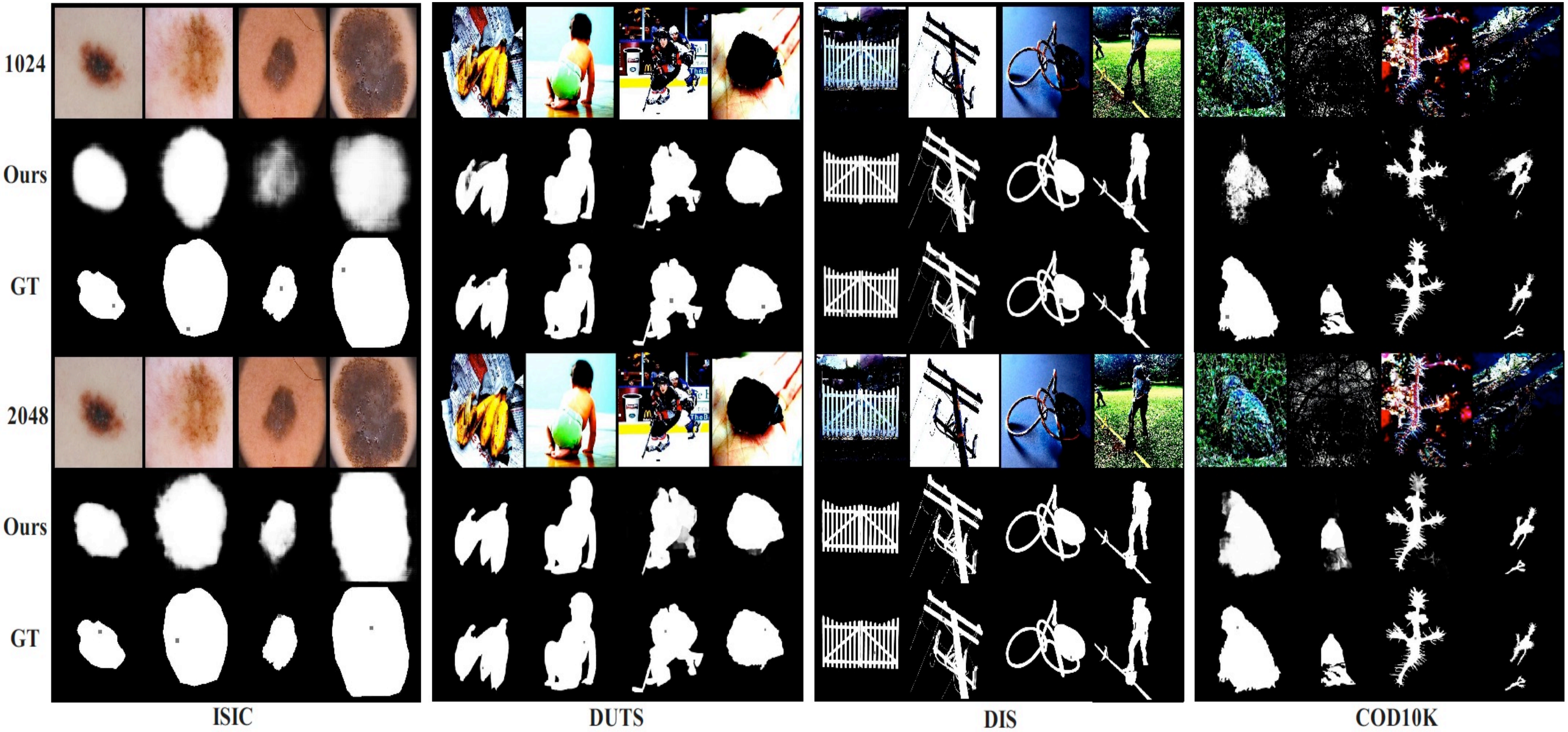



**Scalable Bias-mode Attention Mask (BA-SAM)**

- New Scaling Factor (Left)
- Bias-Mode Attention Mask (Right)

# 5, Close Related Works.



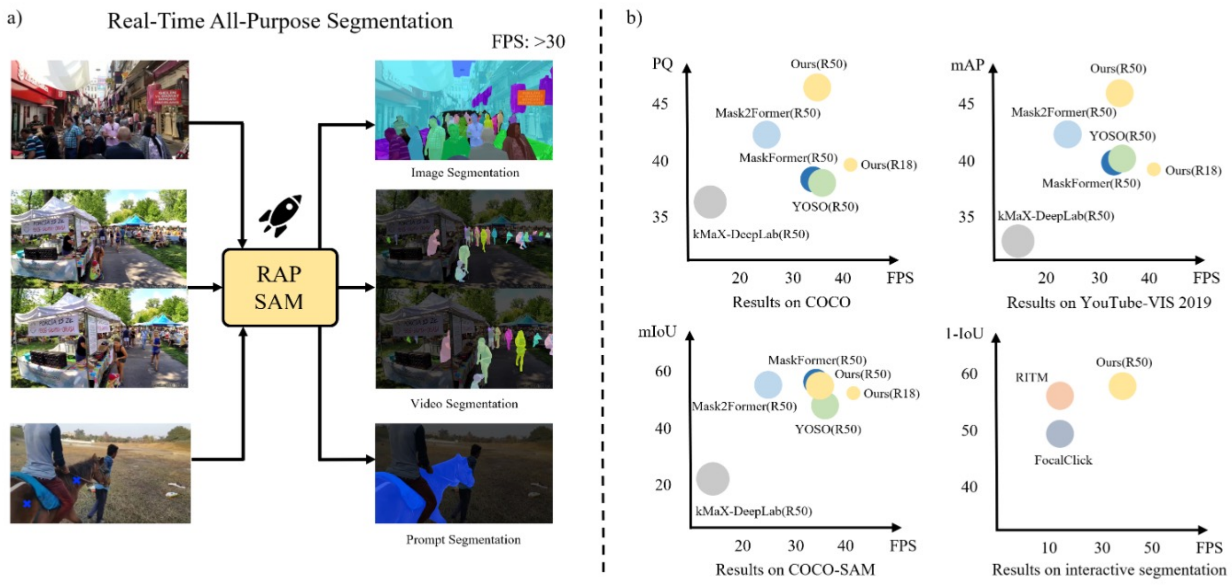
# 5, Close Related Works.

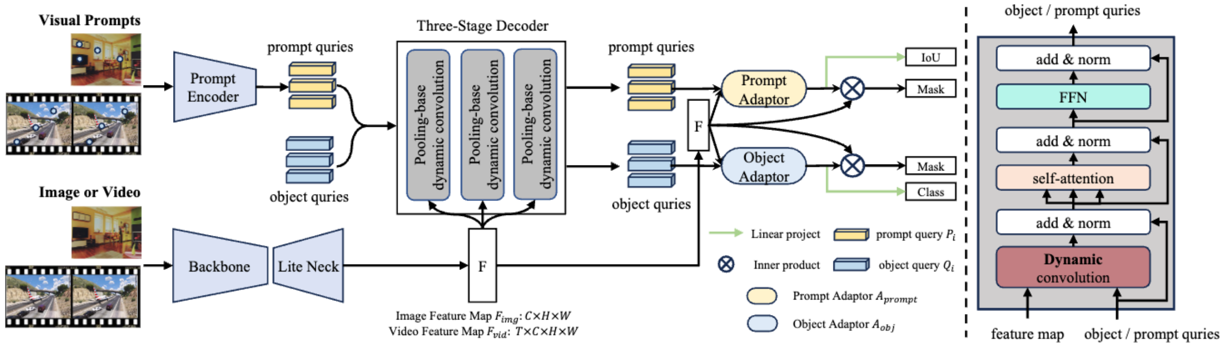
## RAP-SAM:Towards Real-Time All-Purpose Segment Anything


Shilin Xu[1,4*], Haobo Yuan[2*], Qingyu Shi[1], Lu Qi[3], Jingbo Wang[4], Yibo Yang[5], Yining Li[4], Kai Chen[4], Yunhai Tong[1], Bernard Ghanem [5], Xiangtai Li[2,4†], Ming-Hsuan Yang[3,6]

[1]Peking University, [2]Nanyang Technology University, [3]UC, Merced, [4]Shanghai AI Laboratory, [5]KAUST, [6]Google Research

*Equally contribution †Project Leader


| Methods | SS | PS | VIS | Interactive | Multi-Task in One Model | Real Time |
|---|---|---|---|---|---|---|
| ICNet [93] | ✓ | ✗ | ✗ | ✗ | ✗ | ✓ |
| Bi-Seg [82] | ✓ | ✗ | ✗ | ✗ | ✗ | ✓ |
| YOSO [25] | ✓ | ✓ | ✗ | ✗ | ✗ | ✓ |
| Mobilie-VIS [89] | ✗ | ✗ | ✓ | ✗ | ✗ | ✓ |
| SAM [34] | ✗ | ✗ | ✗ | ✓ | ✗ | ✗ |
| Mask2Former [10] | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ |
| Video K-Net [47] | ✗ | ✓ | ✓ | ✗ | ✗ | ✗ |
| OneFormer [29] | ✓ | ✓ | ✗ | ✗ | ✓ | ✗ |
| RAP-SAM (Ours) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

1, New tasks: Real-Time All-Purpose Segmentation.

2, SAM-like model but using a convolution encoder and dynamic convolution decoder.

3, SOTA performance on speed and accuracy trade-off.

# 5, Close Related Works.

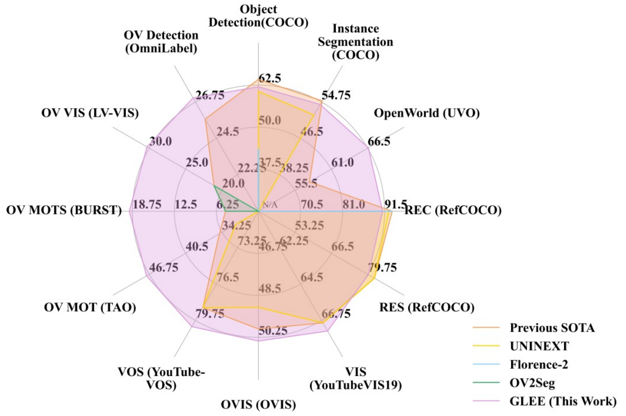


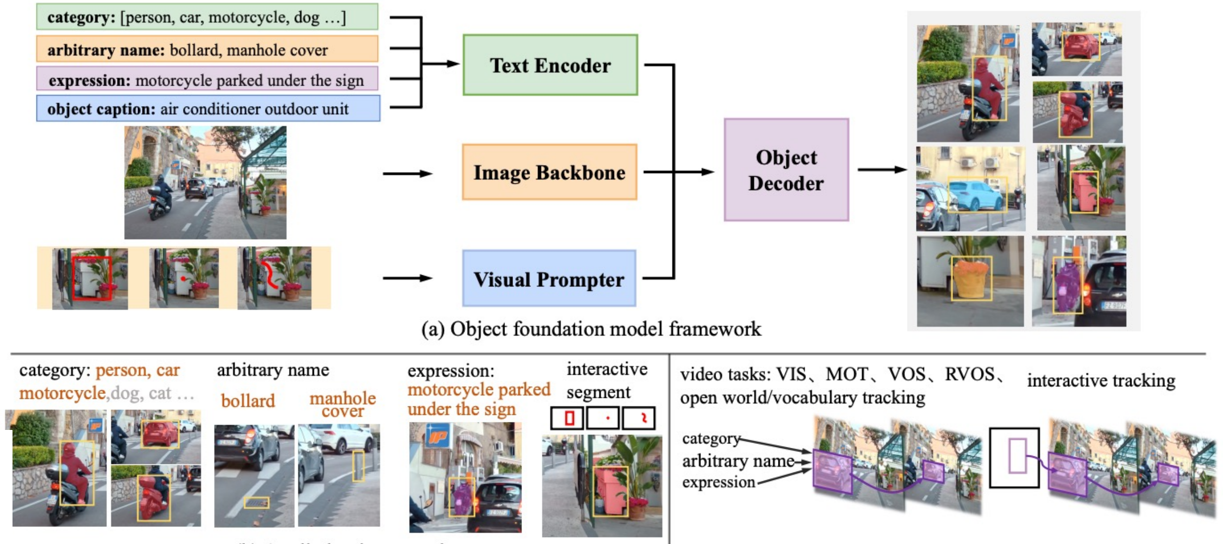

Figure 1. The performance of GLEE on a broad range of object-level tasks compared with existing models.

| Method | Type | Generic Detection & Segmentation | | | | | | | | Referring Detection & Segmentation | | | | | | OpenWorld |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | COCO-val | | COCO-test-dev | | LVIS | | | | RefCOCO | | RefCOCO+ | | RefCOCOg | | UVO |
| | | $AP_{box}$ | $AP_{mask}$ | $AP_{box}$ | $AP_{mask}$ | $AP_{box}$ | $AP_{r-box}$ | $AP_{mask}$ | $AP_{r-mask}$ | P@0.5 | oIoU | P@0.5 | oIoU | P@0.5 | oIoU | $AR_{mask}$ |
| MDETR [42] | Specialist Models | - | - | - | - | - | - | - | - | 87.5 | - | 81.1 | - | 83.4 | - | - |
| SeqTR [131] | | - | - | - | - | - | - | - | - | 87.0 | 71.7 | 78.7 | 63.0 | 82.7 | 64.7 | - |
| PolyFormer (L) [62] | | - | - | - | - | - | - | - | - | 90.4 | 76.9 | 85.0 | 72.2 | 85.8 | 71.2 | - |
| ViTDet-L [55] | | 57.6 | 49.8 | - | - | 51.2 | - | 46.0 | 34.3 | - | - | - | - | - | - | - |
| ViTDet-H [55] | | 58.7 | 50.9 | - | - | 53.4 | - | 48.1 | 36.9 | - | - | - | - | - | - | - |
| EVA-02-L [26] | | 64.2 | 55.0 | 64.5 | 55.8 | 65.2 | - | 57.3 | - | | | | | | | |
| ODISE [107] | | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 57.7 |
| Mask2Former (L) [16] | | - | 50.1 | - | 50.5 | - | - | - | - | - | - | - | - | - | - | - |
| MaskDINO (L) [50] | | - | 54.5 | - | 54.7 | - | - | - | - | - | - | - | - | - | - | - |
| UniTAB (B) [114] | Generalist Models | - | - | - | - | - | - | - | - | 88.6 | - | 81.0 | - | 84.6 | - | - |
| OFA (L) [94] | | - | - | - | - | - | - | - | - | 90.1 | - | 85.8 | - | 85.9 | - | - |
| Pix2Seq v2 [15] | | 46.5 | 38.2 | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Uni-Perceiver-v2 (B) [51] | | 58.6 | 50.6 | - | - | - | - | - | - | - | - | - | - | - | - | - |
| Uni-Perceiver-v2 (L) [51] | | 61.9 | 53.6 | - | - | - | - | - | - | - | - | - | - | - | - | - |
| UNINEXT (R50) [112] | | 51.3 | 44.9 | - | - | 36.4 | - | - | - | 89.7 | 77.9 | 79.8 | 66.2 | 84.0 | 70.0 | - |
| UNINEXT (L) [112] | | 58.1 | 49.6 | - | - | - | - | - | - | 91.4 | 80.3 | 83.1 | 70.0 | 86.9 | 73.4 | - |
| UNINEXT (H) [112] | | 60.6 | 51.8 | - | - | - | - | - | - | 92.6 | 82.2 | 85.2 | 72.5 | 88.7 | 74.7 | - |
| GLIPv2 (B) [123] | | - | - | 58.8 | 45.8 | - | - | - | - | - | - | - | - | - | - | - |
| GLIPv2 (H) [123] | | - | - | 60.6 | 48.9 | - | - | - | - | - | - | - | - | - | - | - |
| X-Decoder (B) [134] | | - | 45.8 | - | 45.8 | - | - | - | - | - | - | - | - | - | - | - |
| X-Decoder (L) [134] | | - | 46.7 | - | 47.1 | - | - | - | - | - | - | - | - | - | - | - |
| Florence-2 (L) [106] | | 43.4 | - | - | - | - | - | - | - | 93.4 | - | 88.3 | - | 91.2 | - | - |
| GLEE-Lite | Foundation Models | 55.0 | 48.4 | 54.7 | 48.3 | 44.2 | 36.7 | 40.2 | 33.7 | 88.5 | 77.4 | 78.3 | 64.8 | 82.9 | 68.8 | 66.6 |
| GLEE-Plus | | 60.4 | 53.0 | 60.6 | 53.3 | 52.7 | 44.5 | 47.4 | 40.4 | 90.6 | 79.5 | 81.6 | 68.3 | 85.0 | 70.6 | 70.6 |
| GLEE-Pro | | 62.0 | 54.2 | 62.3 | 54.5 | 55.7 | 49.2 | 49.9 | 44.3 | 91.0 | 80.0 | 82.6 | 69.6 | 86.4 | 72.9 | 72.6 |

Table 1. Comparison of GLEE to recent specialist and generalist models on object-level image tasks. For REC and RES tasks, we report Precision@0.5 and overall IoU (oIoU). For open-world instance segmentation task, we reported the average recall of 100 mask proposals (AR@100) on the UVO [96].

## General Object Foundation Model for Images and Videos at Scale


Junfeng Wu[1*], Yi Jiang[2*], Qihao Liu[3], Zehuan Yuan[2], Xiang Bai[1†], Song Bai[2†]
[1]Huazhong University of Science and Technology, [2]ByteDance Inc., [3]Johns Hopkins University


(a) Object foundation model framework

(b) Applied to image tasks

(c) Applied to video tasks

1, Unify the all object centered datasets and tasks in one training format.

2, Train one transformer model on such format.

3, SOTA performance.

# 5, Close Related Works.

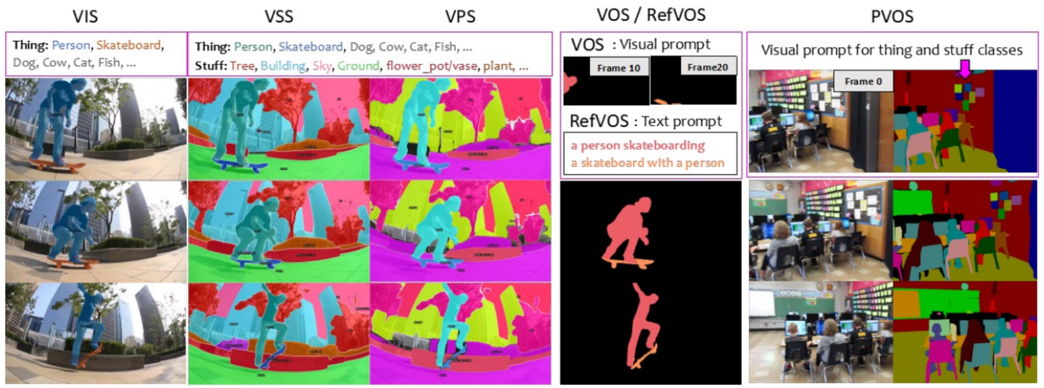
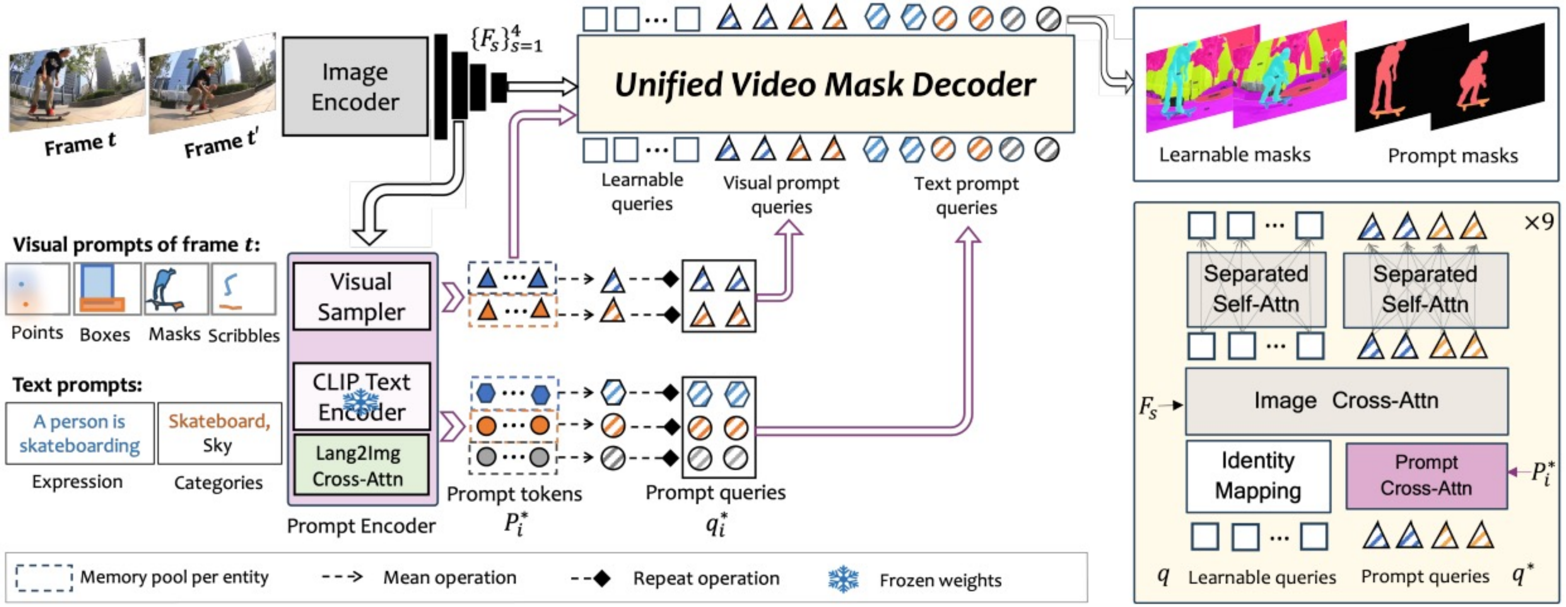



Figure 1. Illustration of different video segmentation (VS) tasks. Category-specified VS includes VIS, VSS and VPS tasks, while prompt-specified VS consists of VOS, RefVOS and PVOS tasks. Please find more video demos on our project page https://sites.google.com/view/unified-video-seg-univs.

## UniVS: Unified and Universal Video Segmentation with Prompts as Queries


Minghan Li[1,2]*, Shuai Li[1,2]*, Xindong Zhang[2] and Lei Zhang[1,2]†
[1]The Hong Kong Polytechnic University [2]OPPO Research Institute
liminghan0330@gmail.com, xindongzhang@foxmail.com, {csshuaili, cslzhang}@comp.polyu.edu.hk

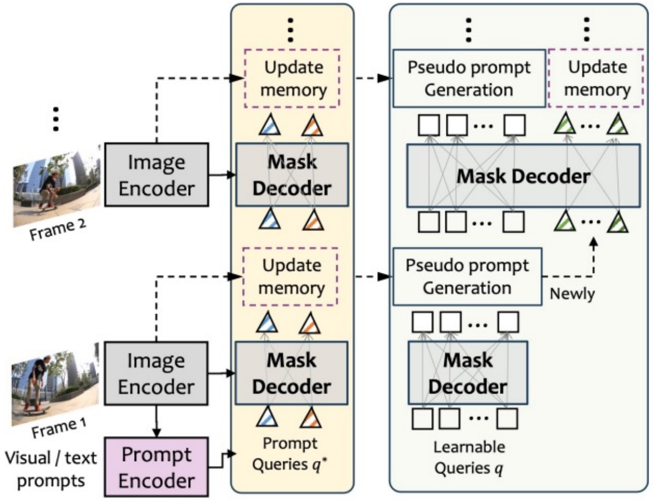




Figure 4. Inference process of our UniVS on prompt-specified and category-specified video segmentation tasks, respectively.

# 5, Summary

**Summary:**

1, Unified architecture for multiple datasets and tasks is one research trend.

2, Efficient modeling for SAM need specific designs.

3, Knowledge transfer and combination of foundation models are important for downstream application.

**Future Work Direction:**

1, Scale up model training with SAM-1B datasets.

2, Unify generation model and segmentation model.

3, Combining SAM-Like model with LLMs.

# 5, Summary and Q & A

| OMG-Seg | Edge-SAM | OV-SAM | BA-SAM | RAP-SAM |
|---|---|---|---|---|

Our code and models are available to the community.

Welcome to start and use it. We also support Hugging Face Models

## Acknowledgement for co-authors:

Haobo Yuan, Chong Zhou, Yiran Song, Shilin Xu, Qianyu Zhou, Wei Li, Yining Li, Henghui Ding, Chen Change Loy