NEURAL INFORMATION PROCESSING SYSTEMS

Uncertainty-aware fine-tuning of segmentation foundation models

Kangning Liu[1,2], Brian Price[2], Jason Kuen[2], Yifei Fan[2], Zijun Wei[2], Luis Figueroa[2], Krzysztof J. Geras[1], Carlos Fernandez-Granda[1]

1. New York University 2. Adobe

NYU DATA SCIENCE

Adobe

Kangning Liu
Nov 10, 2024

Segment Anything Model (SAM) is a large-scale foundation model that has revolutionized segmentation methodology

SAM features robust zero-shot capabilities and flexible prompting options (e.g. point, box, mask)

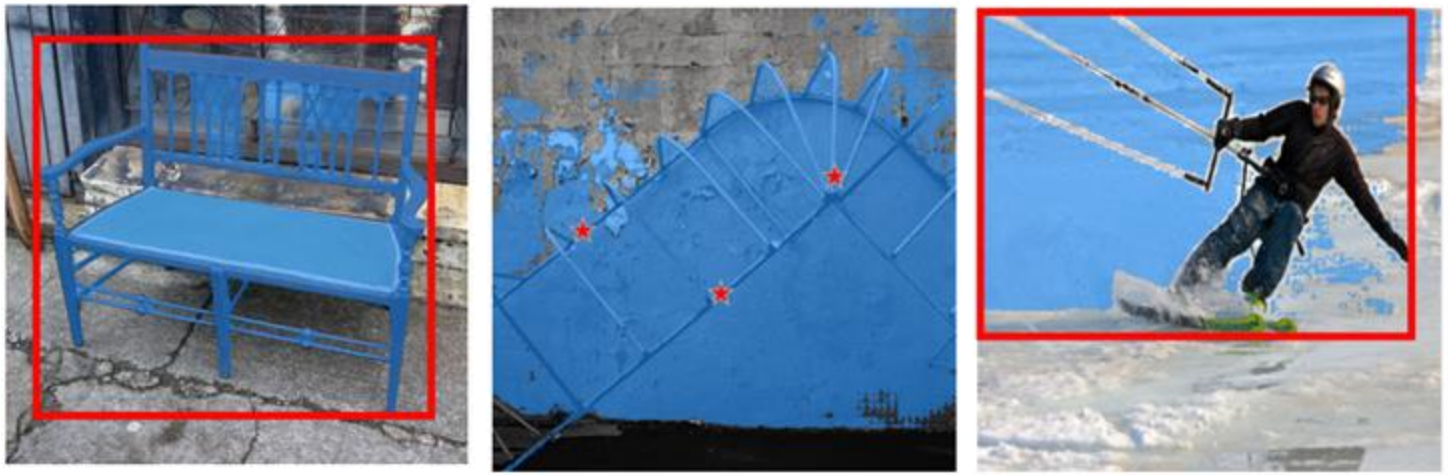
Interactive segmentation

Automatic segmentation

Zero-shot capacity

Example from https://segment-anything.com/

However, the prediction of SAM is unsatisfactory in many cases, especially for intricate structures



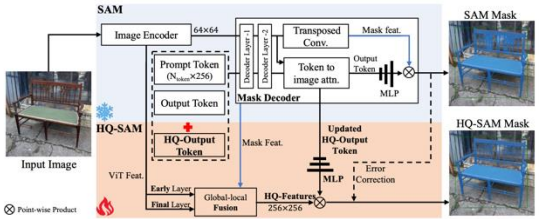
"SAM is not optimized for the very high IoU regime" – SAM paper

Existing methods suggest fine-tuning SAM with high quality annotations for enhanced performance

SAM-Adapter:

HQ-SAM:

Tuning 0.5% parameters



It claims to preserve the zero-shot capabilities and flexibility of SAM by such lightweight fine-tuning.

Is the zero-shot performance really preserved? Let's find out

- Previously, HQ-SAM was assessed on closely related datasets, all focused on object segmentation tasks.

- We have constructed a more comprehensive evaluation set containing a broader array of tasks for thorough analysis.

Comprehensive evaluation across segmentation tasks

1. Salient Object:

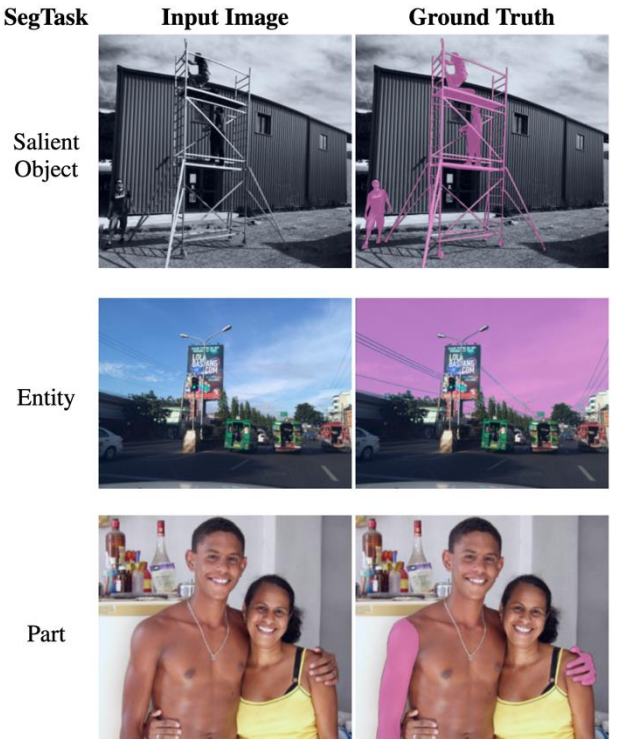
   - COIFT/ DIS-VD/ ThinObject5K/ HR-SOD/ VOCEVAL/BIG

2. Entity:

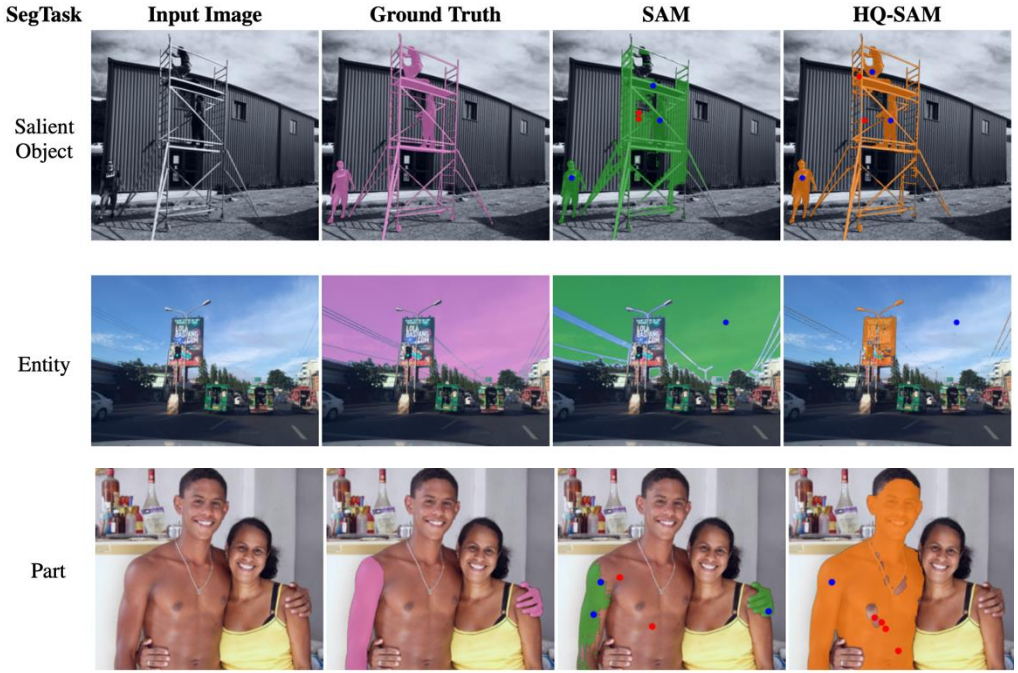
   - EntitySeg validation sets with 454, 459 and 401 images, detailing both foreground and background.

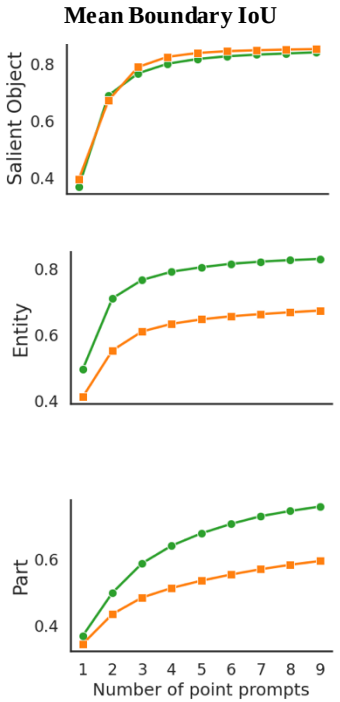
3. Part:

   - Fashionpedia (1,148 images)
   - Fashionpedia subpart (868 images)
   - Multi-Human Parsing (1,000 images)
   - Easyportrait (1,000 images)
   - Paco (1,000 images)



Our experiments show that HQ-SAM forgets how to “segment anything”





Our proposal:

Leverage **human labeled data** to improve segmentation quality :

- Pro: high-quality annotation
- Con: limited quantity and variety

Leverage **unlabeled data** to prevent overfitting

- Pro: diverse and large quantity
- Con: SAM pseudo labels can be noisy

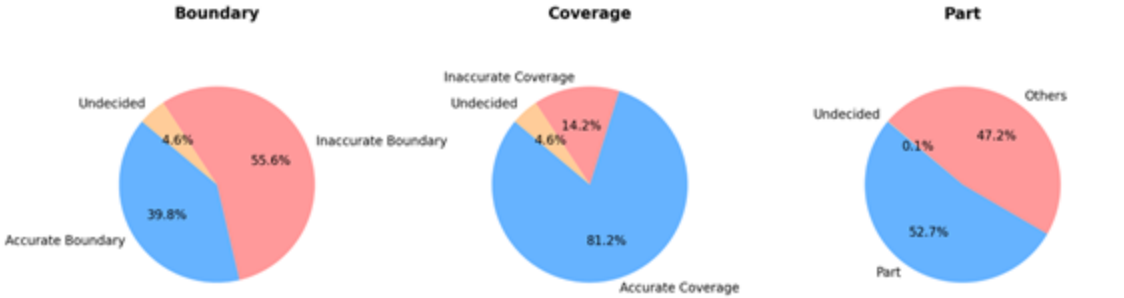
Challenges

Human annotated data and pseudo labels diverge significantly.

Simply merging them without distinction can lead to issues:

- Pseudo-labels are inaccurate
- Human annotations focus on different tasks, resulting in systematic differences between labels

Challenge 1: the pseudo-labels are inaccurate

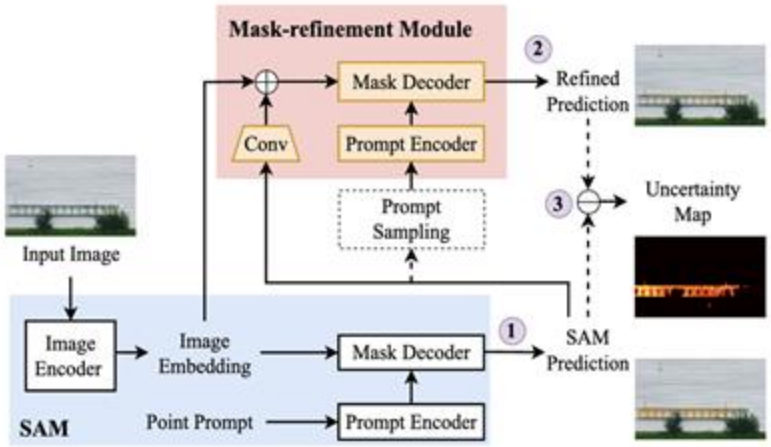


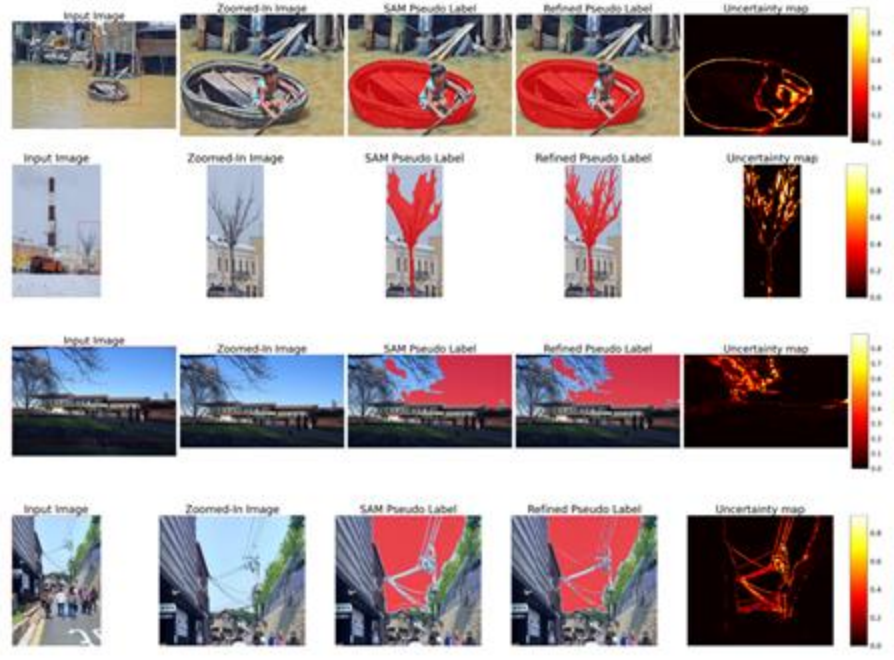
(1) Inaccurate segmentation is reinforced;

(2) Point prompts sampled from the pseudo labels can be completely incorrect.

Solution: quantify and leverage uncertainty in the SAM pseudo labels

This inaccuracy is systematic and predictable.





Challenge 2: human annotations focus on different tasks, resulting in systematic differences between labels

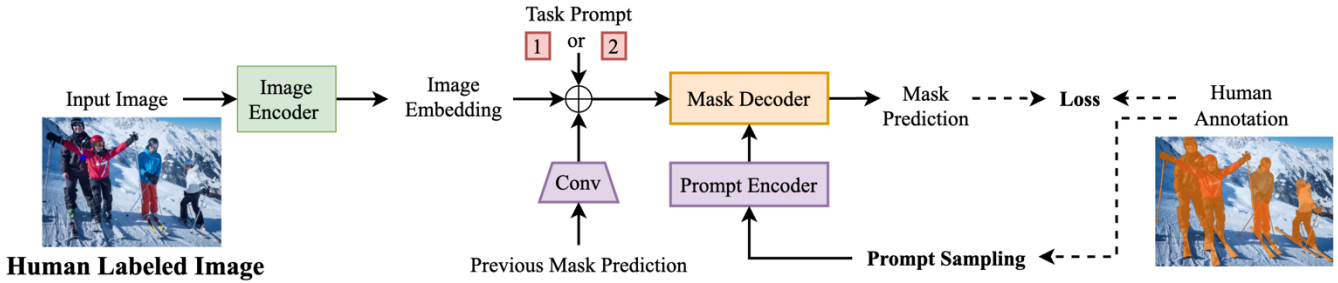
- Human-annotated masks may include **multiple entities** in a complex arrangement.


- SAM pseudo labels, mostly correspond to **entity segmentation** or **part segmentation**


A high degree of ambiguity regarding the segmentation mask that should be predicted by the model after the initial prompt.

Solution: Incorporate a **task prompt**, indicating the segmentation task relevant to each example.

Segment with Uncertainty Model (SUM) framework



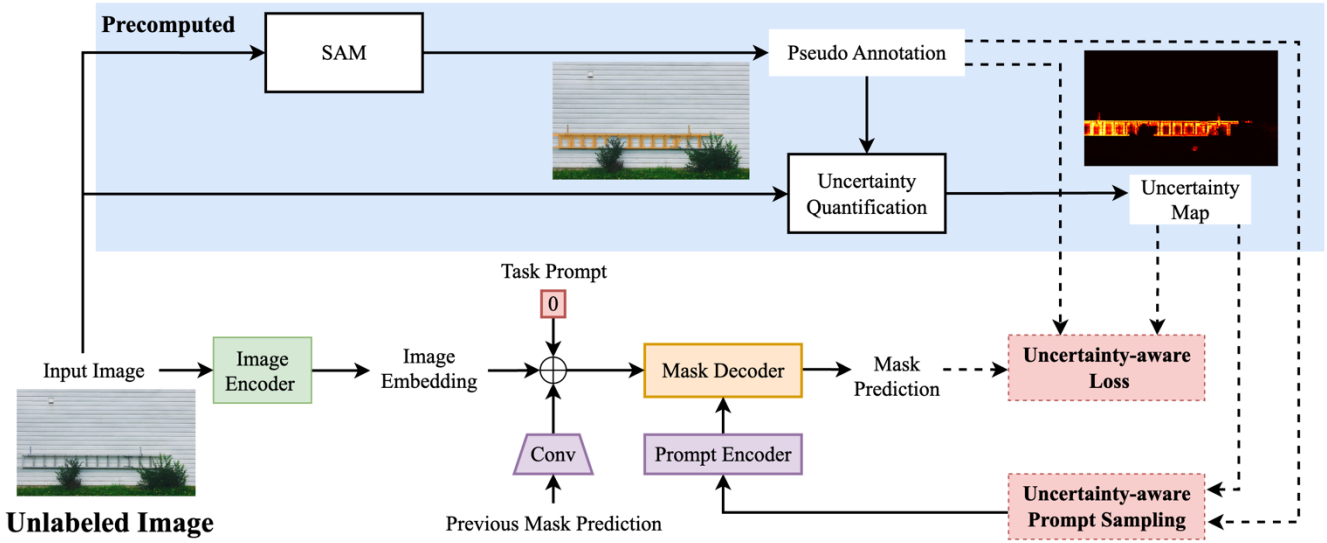
Salient Object

1

Entity

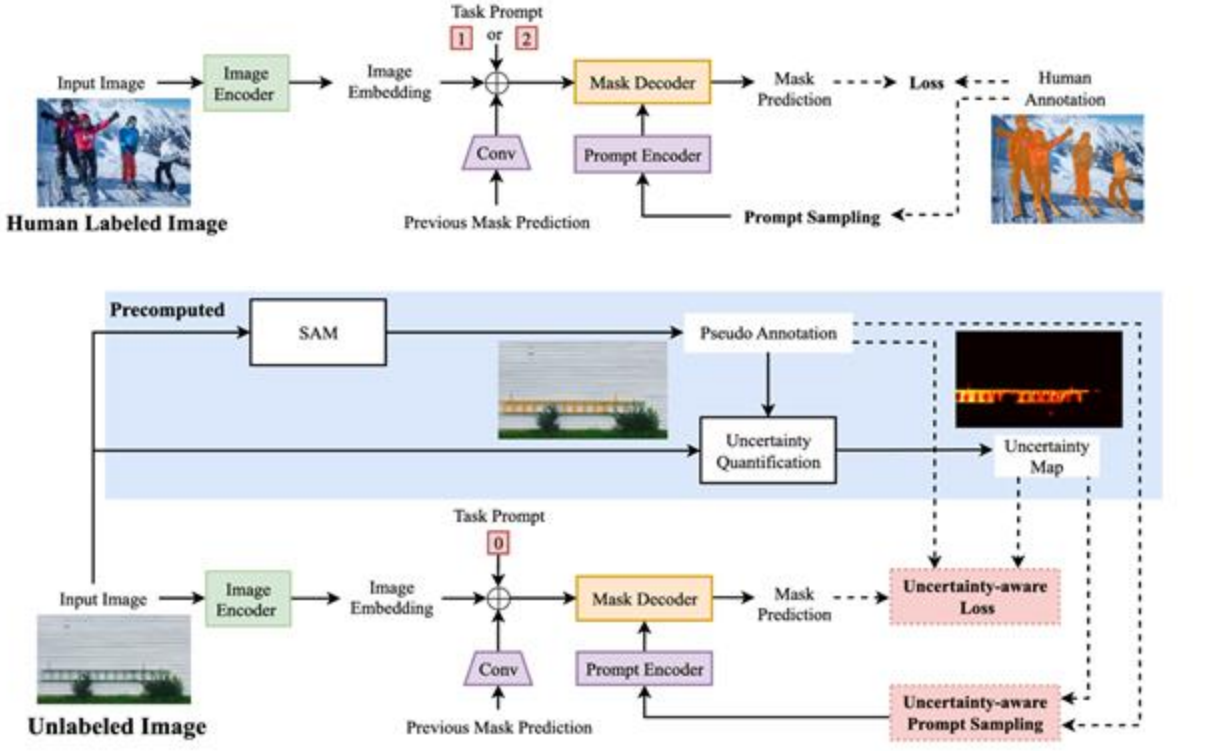
2

Segment with Uncertainty Model (SUM) framework



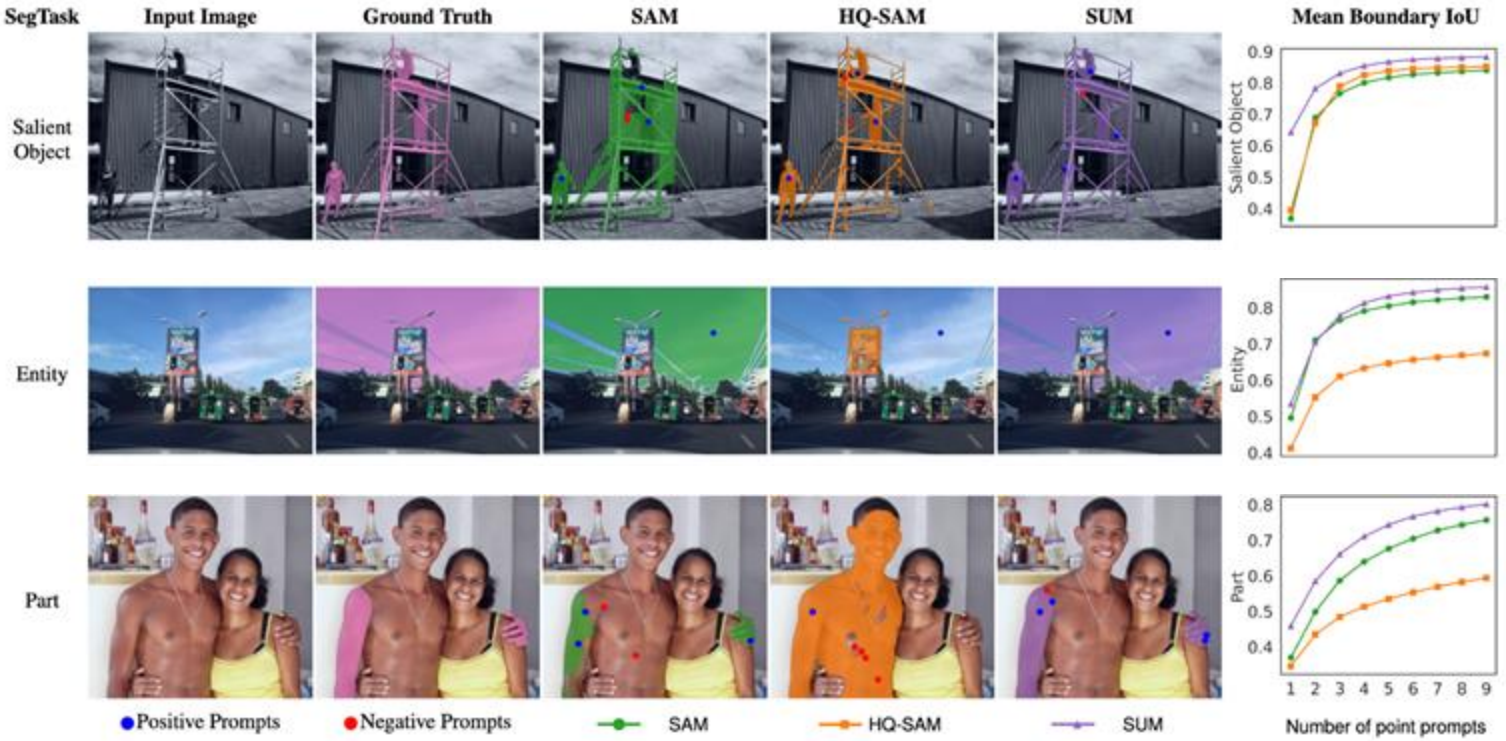
- ***Uncertainty-aware prompt sampling****:* reduces misleading prompt
- ***Uncertainty-aware loss****:* reduces the influence of regions that are expected to be inaccurate

Segment with Uncertainty Model (SUM) framework



The task prompt can also be leveraged to specify the desired segmentation task during inference.

SUM improves SAM without forgetting to “segment anything”



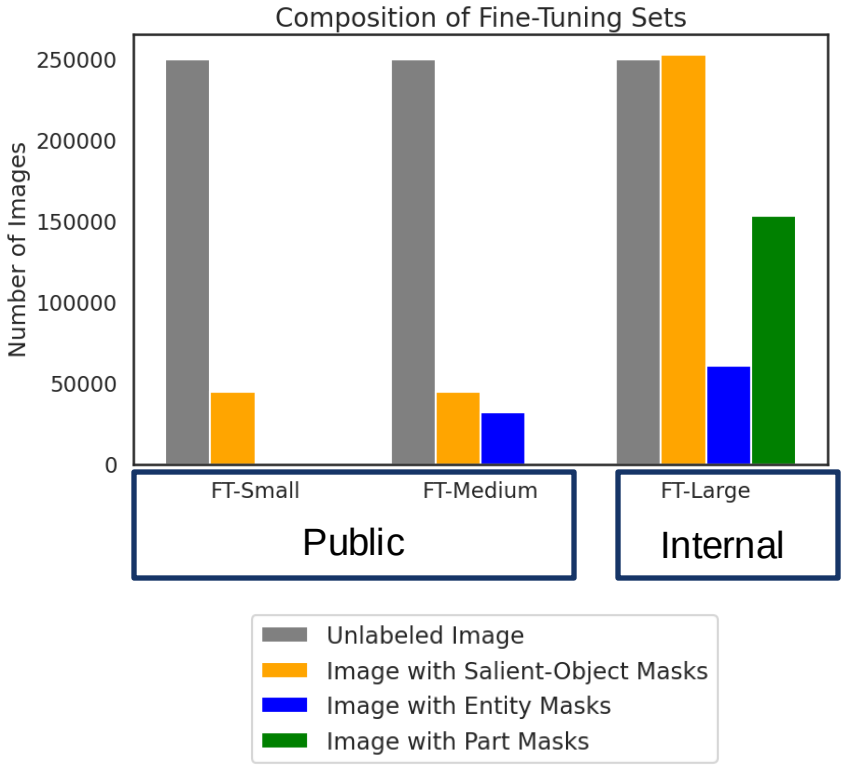
Experiments with different numbers of human annotations

Datasets

- SA-250K: Unlabeled 250,000 images.
- HQSeg-44K: human-annotated 44,320 images with high-quality salient-object masks.
- EntitySeg training set: human-annotated 31,913 images, each with an average of 20 entity masks,
- Internal dataset: A human-annotated set containing 252,798 images with salient-object masks, 60,798 with entity masks, and 153,046 focused on part segmentation for human parsing.

Fine-Tuning Sets

- FT-Small: SA-250K and HQSeg-44K
- FT-Medium: SA-250K, HQSeg-44K, and EntitySeg
- FT-Large: SA-250K and the internal dataset



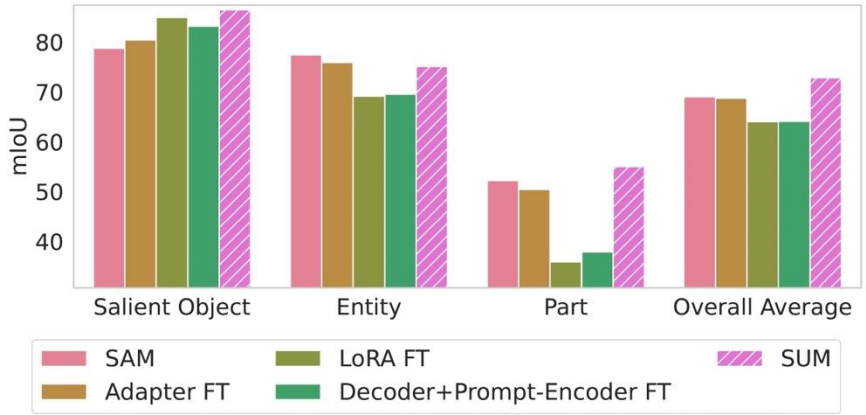
SUM outperforms SAM across different budgets

- 5-point prompted interactive segmentation results

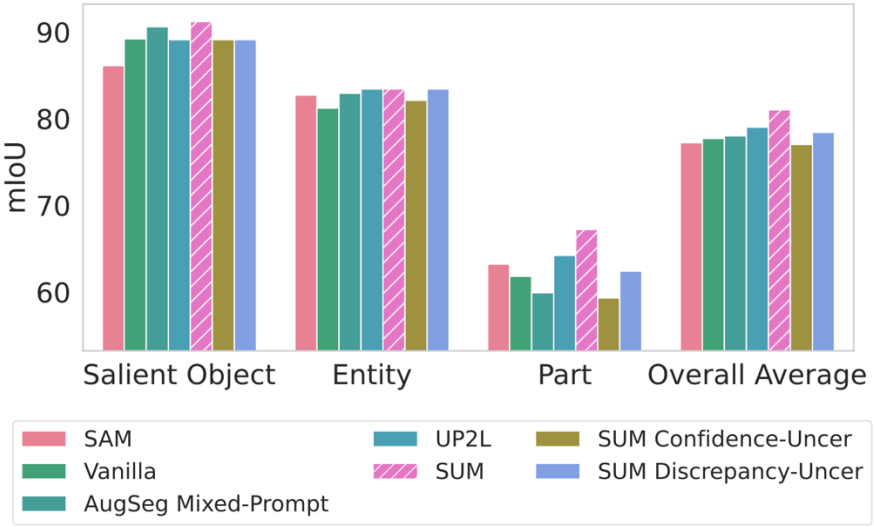
Metrics	Task	Dataset	SAM	SUM (FT-small)	SUM (FT-Medium)	SUM (FT-Large)
mBIoU	Salient object	COIFT	90.9	92.3	92.7	**93.2**
		DIS-VD	68.0	77.4	78.5	**78.6**
		HRSOD	87.1	88.4	89.8	**90.9**
		ThinObject5K	79.0	84.0	85.6	**86.2**
		Big	85.2	85.6	89.1	**90.1**
		Voceval	81.3	82.2	84.5	**84.8**
		Average	81.9	85.0	86.7	**87.3**
	Entity	Entityseg 0	78.1	78.2	**82.9**	81.6
		Entityseg 1	81.3	80.9	**84.8**	83.4
		Entityseg 2	82.5	82.1	**86.1**	84.2
		Average	80.6	80.4	**84.6**	83.1
	Part	Bodypart	73.1	76.4	76.9	**80.3**
		Easyportrait	60.7	69.5	67.2	**81.4**
		Fashionpedia	74.0	73.2	74.7	**76.6**
		Fashionpedia subpart	66.3	67.0	68.7	**69.4**
		Paco	66.7	67.2	69.0	**70.4**
		Average	68.2	70.7	71.3	**75.6**
	All	Average	76.7	78.9	80.8	**82.2**

Comparison with previous methods

Comparison with Light-weight Fine-tuning Methods



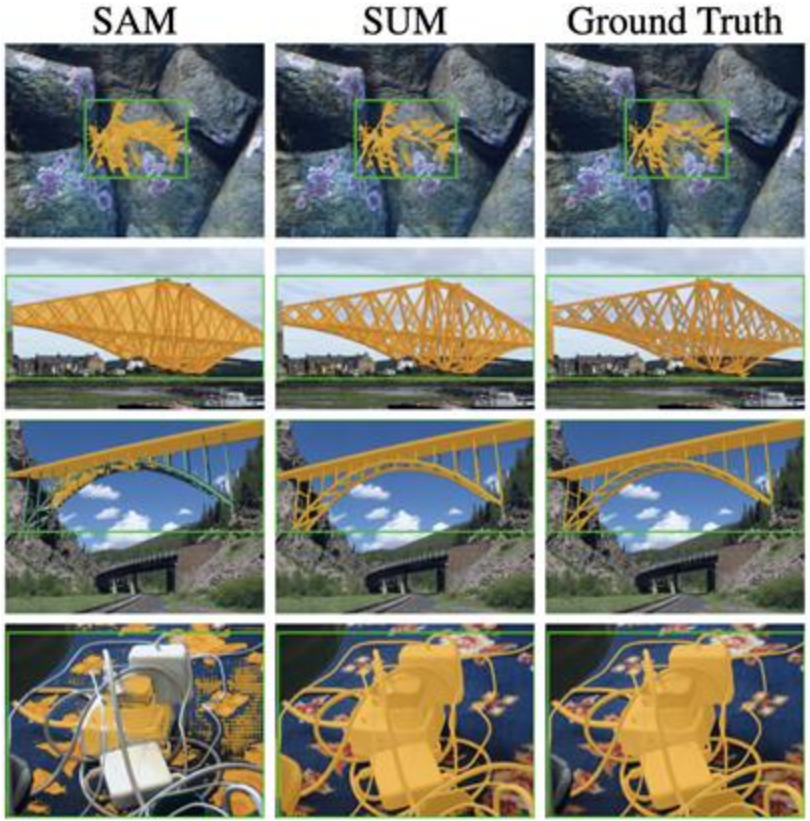
Comparison with Semi-supervised Methods



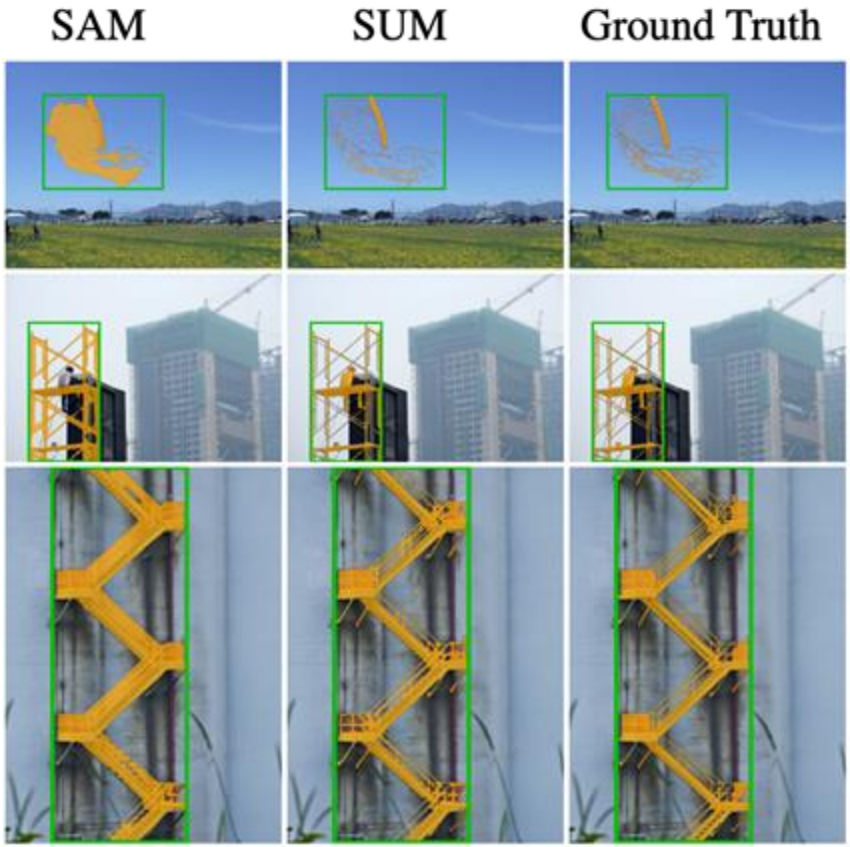
Ablation

Metrics	Point Number	Task	SAM (without fine-tuning)	Vanilla	Mask Refinement	SUM w/o Task Prompt	SUM	SUM Continuous TP
mIoU	1	Salient object	78.7	80.4	81.5	81.1	**85.2**	84.7
		Entity	77.4	78.8	78.1	78.6	**79.8**	79.2
		Part	52.2	50.2	52.3	54.4	**55.3**	54.6
		Overall average	69.0	69.3	70.3	71.0	**73.4**	72.8
	3	Salient object	86.1	90.4	91.5	91.4	**91.6**	89.6
		Entity	82.7	86.5	86.2	**86.7**	86.6	85.7
		Part	63.2	64.4	66.7	67.8	**67.9**	67.7
		Overall average	77.2	80.3	81.5	82.0	**82.1**	81.0

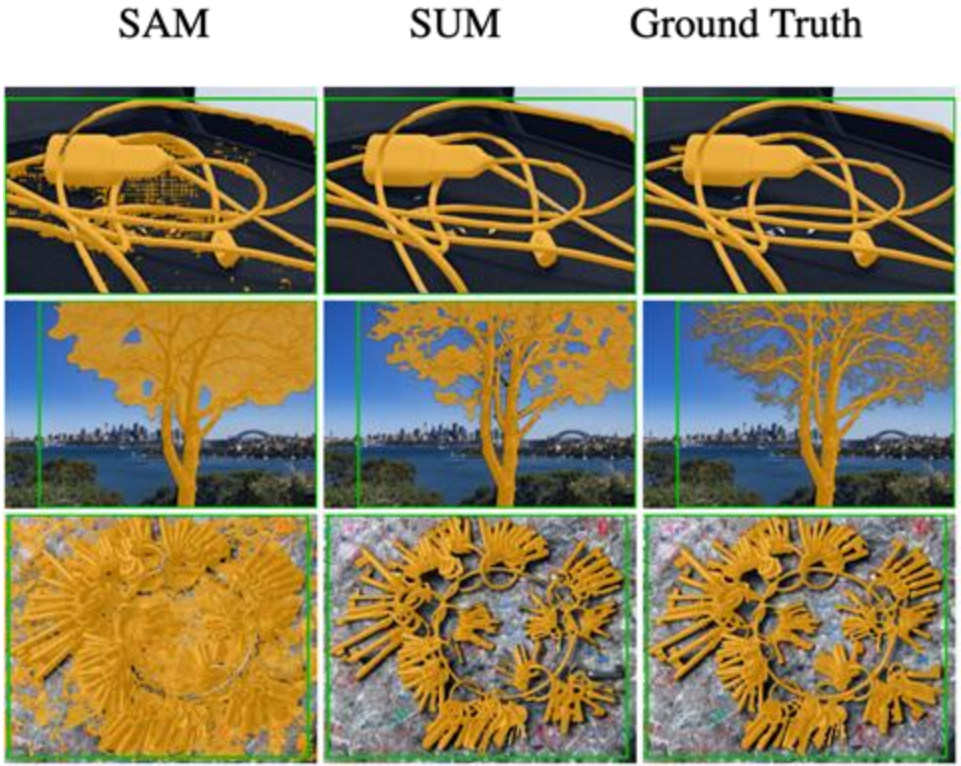
Visualization examples



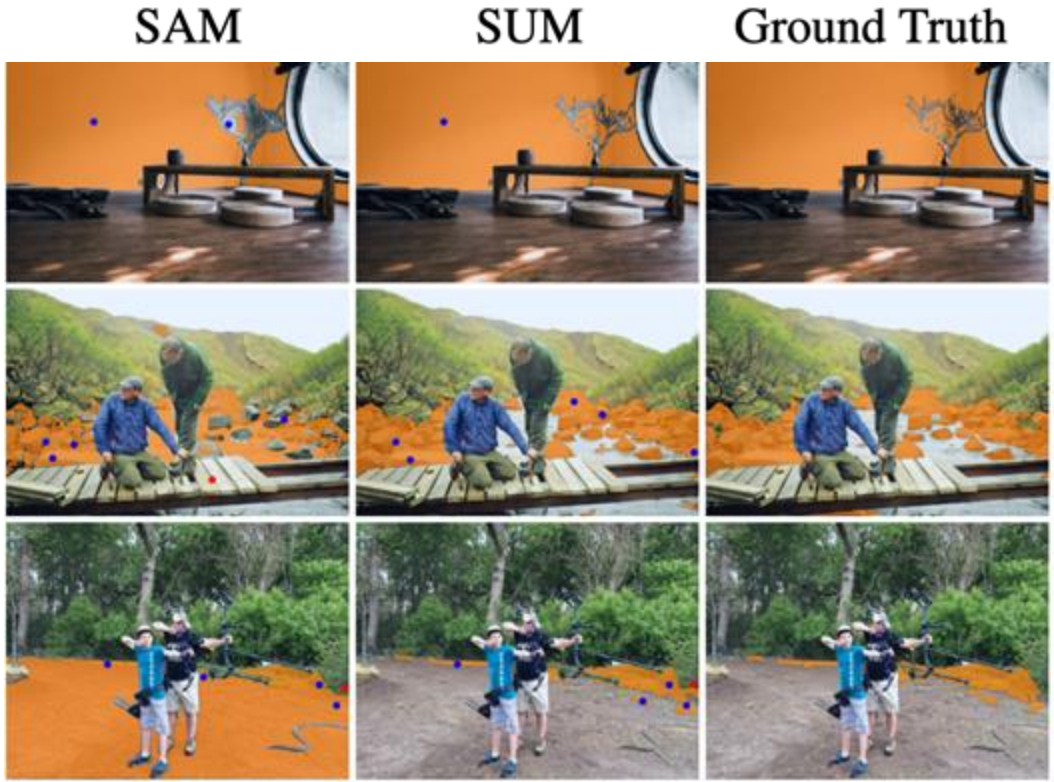
Visualization examples



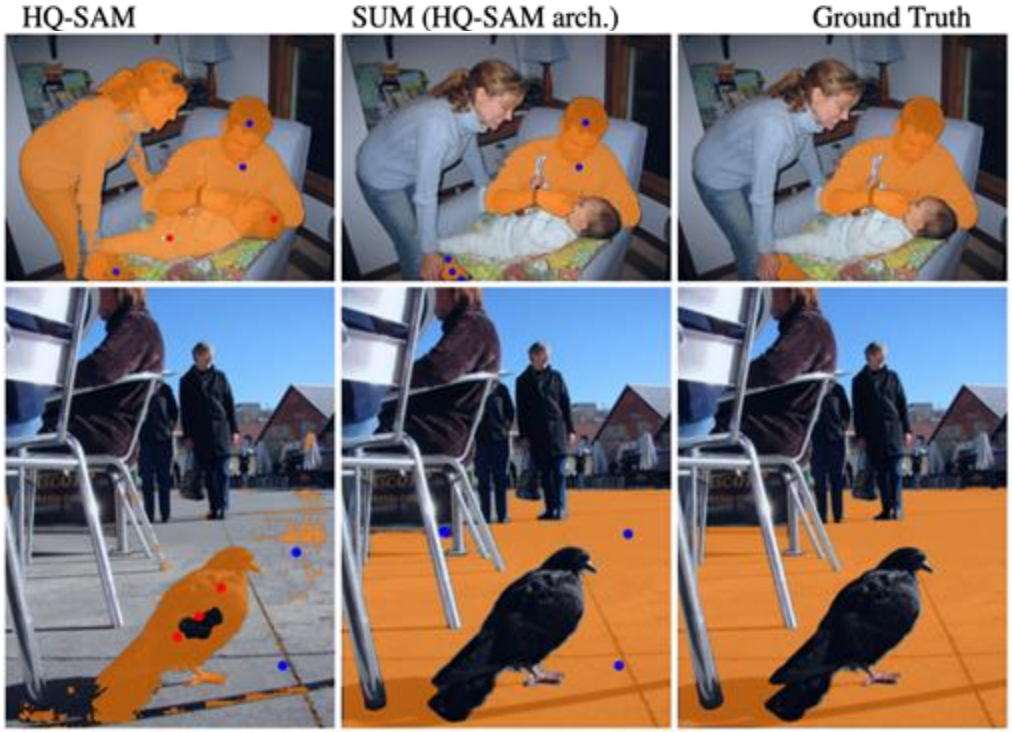
Visualization examples



Visualization examples



Visualization examples



Thanks