AoP-SAM: Automation of Prompts for Efficient Segmentation

Yi Chen[1], Muyoung Son[1], Chuanbo Hua[1], Joo-Young Kim[1]

[1] Korea Advanced Institute of Science and Technology (KAIST)

Summary

TL; DR. We propose AoP-SAM, a novel approach automatically generate essential prompts for accurate segmentation, eliminating the need for manual prompt provision.

Motivations

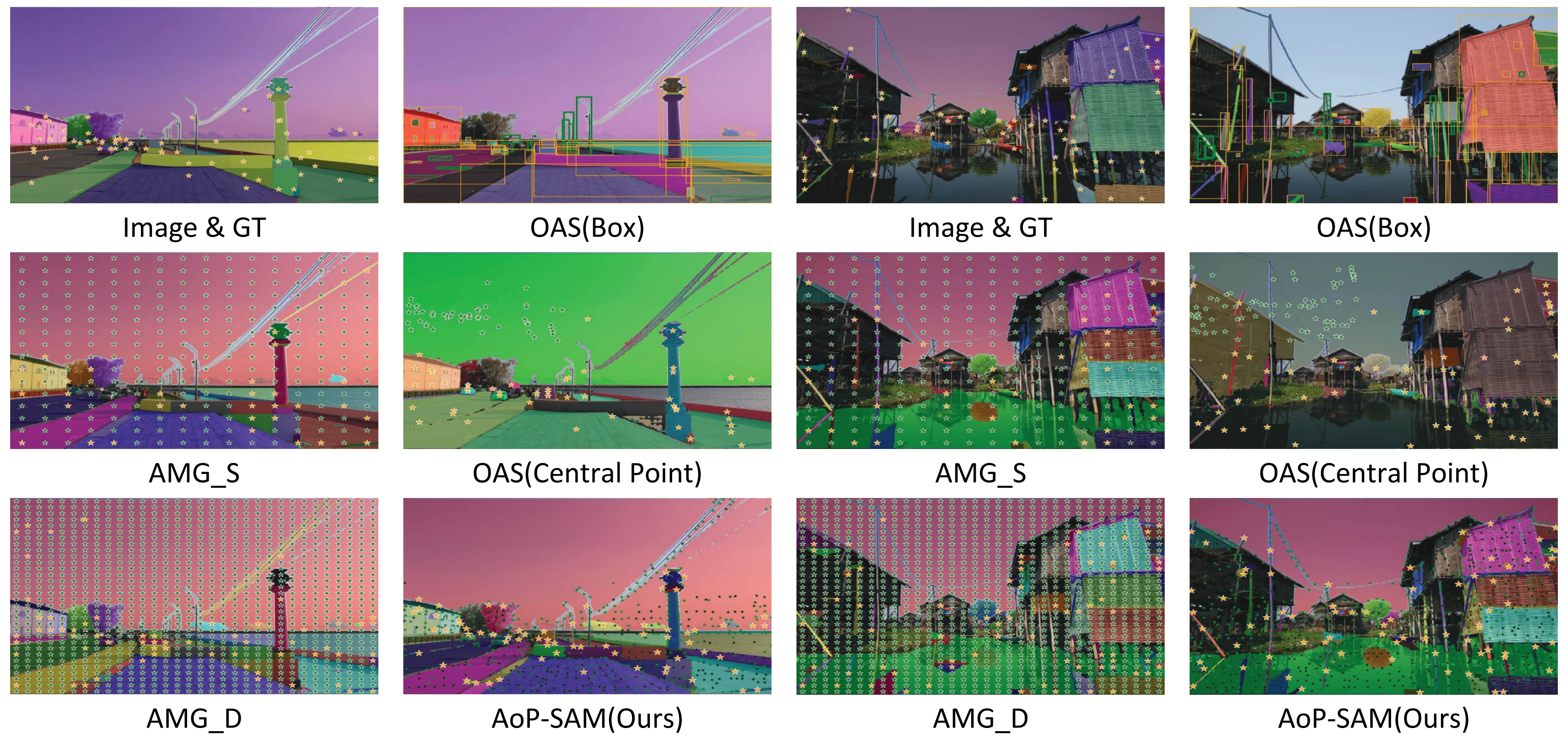


Image & GT | OAS(Box) | Image & GT | OAS(Box)

AMG_S | OAS(Central Point) | AMG_S | OAS(Central Point)

AMG_D | AoP-SAM(Ours) | AMG_D | AoP-SAM(Ours)

Automating SAM's prompt provision eliminates manual input needs, enhancing mask segmentation efficiency. However, existing approaches face limitations:

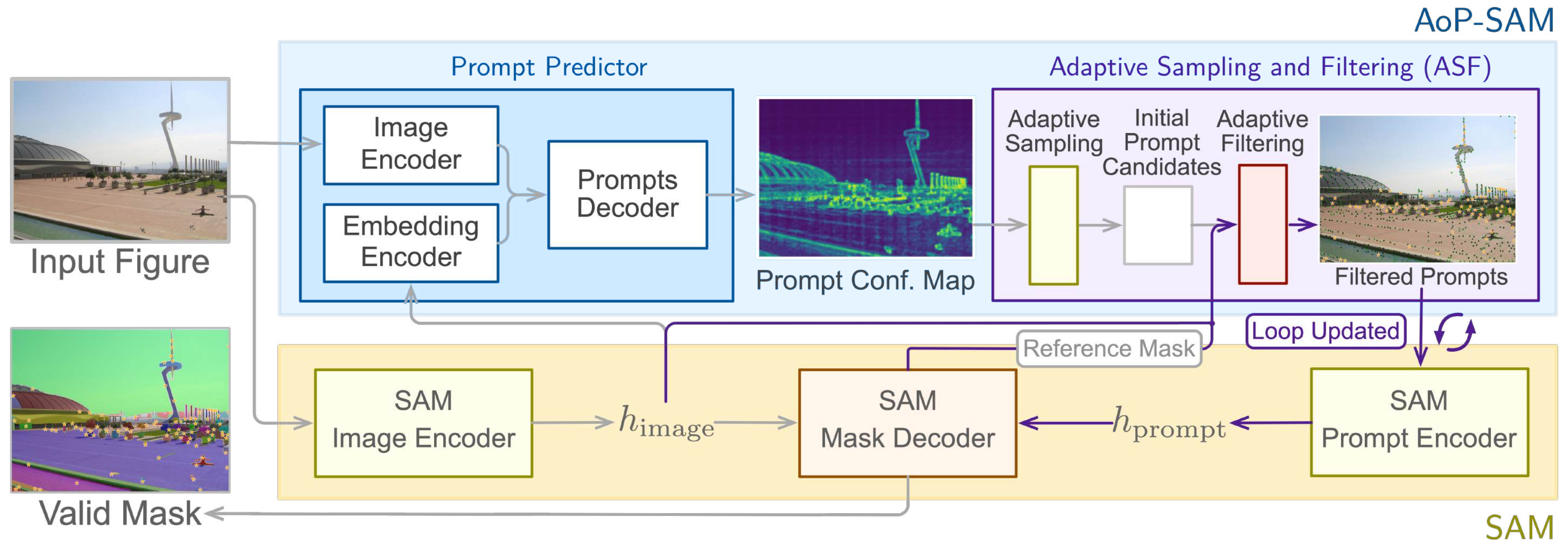
- Grid-based prompts lead to excessive mask refinements;
- Extra object detection models create computational overhead;
- Both result in increased latency and reduced efficiency.

We addresses these challenges by efficiently generating essential prompts for accurate mask generation without human intervention:

- Orange labels (stars/boxes): Valid prompts;
- Green labels: Invalid prompts;
- Black stars: Filtered prompts processed by AoP-SAM.

SAM (Segment Anything Model) is an image segmentation framework with three core components: an Image Encoder, Prompt Encoder, and Mask Decoder, which work together to generate precise segmentation masks from images and interactive inputs. Trained on over 1 billion masks, SAM achieves remarkable zero-shot generalization capabilities while optimizing efficiency by computing embeddings only once.

AoP-SAM



- **Prompt Predictor** Utilizes a dual-encoder architecture (CNN + ViT) to process both original image and SAM's embeddings. Processes inputs through CNN layers with ReLU activation and generates a *Prompt Confidence Map* (PCM) using sigmoid activation, highlighting optimal regions for prompt placement.
- **ASF Coarse Processing** Applies Gaussian filtering to the PCM to reduce noise and identify local maxima. These maxima serve as initial prompt candidates and are mapped back to original image coordinates for precise placement of potential prompts.
- **ASF Fine Filtering** Creates a *Prompt Elimination Map* (PEM) using cosine similarity between image features and reference masks. Applies adaptive threshold to remove redundant prompts, ensuring only essential ones remain for final mask generation.
- **Training** Leverages SA-1B dataset with over 1B masks and prompts. Uses point prompts as ground truth with MSELoss and Adam optimization over 1000 epochs. This approach maintains SAM's robust generalization capabilities while adding efficient prompt generation.

Measurement Matrix

- mIoU : Mean Intersection over Union. Accuracy by matching masks with ground truth masks using greedy algorithm;
- $\text{Inf}_{\text{Lat.}}$ (s): Inference Latency. Time taken to produce prompts;
- $\text{Peak}_{\text{Mem.}}$ (GB): Peak Memory. Measures maximum memory consumption during mask generation;
- $\#P$: Number of Essential Prompts - counts prompts needed (smaller number preferred if accuracy remains high).

Experiments

Image Encoders	Auto Prompts Methods	SA-1B				COCO				LVIS			
		mIoU ↑	$\text{Inf}_{\text{Lat.}}$ ↓	$\text{Peak}_{\text{Mem.}}$ ↓	$\#P$	mIoU ↑	$\text{Inf}_{\text{Lat.}}$ ↓	$\text{Peak}_{\text{Mem.}}$ ↓	$\#P$	mIoU ↑	$\text{Inf}_{\text{Lat.}}$ ↓	$\text{Peak}_{\text{Mem.}}$ ↓	$\#P$
MobileSAM	AMG_S	29.8	-	4.5	38.6	56.0	-	**1.9**	33.5	56.2	-	**1.9**	33.4
	AMG_D	46.9	-	9.1	71.0	60.9	-	**1.9**	55.9	61.1	-	**1.9**	55.5
	OAS(Box)	50.7	0.191	7.3	100	55.5	0.187	4.2	44	55.7	0.188	4.0	38
	OAS(Central Point)	48.7	0.188	7.7	141.0	53.9	0.167	4.3	69.0	54.5	0.164	4.3	68.1
	AoP-SAM	**51.4**	**0.101**	**4.1**	71.7	**61.5**	**0.096**	2.1	58.1	**62.3**	**0.094**	2.1	57.5
ViT_L	AMG_S	40.0	-	5.7	55.5	61.4	-	4.4	48.8	63.2	-	**4.3**	49.5
	AMG_D	65.6	-	10.3	108.9	67.7	-	**4.3**	86.0	69.2	-	**4.3**	86.5
	OAS(Box)	65.8	0.150	9.1	100	63.3	0.152	5.4	44	62.9	0.151	5.3	38
	OAS(Central Point)	67.6	0.149	9.7	199.3	64.2	0.133	5.5	98.4	63.5	0.132	5.5	98.9
	AoP-SAM	**71.1**	**0.120**	**5.4**	118.3	**68.4**	**0.116**	4.4	97.0	**69.8**	**0.117**	4.4	97.2
ViT_H	AMG_S	40.8	-	7.1	56.3	63.3	-	5.7	49.8	64.9	-	5.6	50.5
	AMG_D	66.8	-	11.8	109.6	69.5	-	5.7	87.4	71.0	-	5.6	88.0
	OAS(Box)	66.9	0.160	10.4	100	64.1	0.152	6.8	44	63.3	0.153	6.6	38
	OAS(Central Point)	68.3	0.154	11.1	207.6	65.1	0.134	6.9	102.1	63.0	0.134	6.8	102.4
	AoP-SAM	**70.6**	**0.122**	**6.6**	107.8	**70.1**	**0.120**	**5.5**	90.0	**71.9**	**0.122**	**5.5**	89.7

Method's variant			SA-1B				COCO				LVIS			
Prompt Predictor	Adaptive Sampling	Adaptive Filtering	mIoU ↑	$\text{Inf}_{\text{Lat.}}$ ↓	$\text{Peak}_{\text{Mem.}}$ ↓	$\#P$	mIoU ↑	$\text{Inf}_{\text{Lat.}}$ ↓	$\text{Peak}_{\text{Mem.}}$ ↓	$\#P$	mIoU ↑	$\text{Inf}_{\text{Lat.}}$ ↓	$\text{Peak}_{\text{Mem.}}$ ↓	$\#P$
✓			57.2	0.059	7.2	106.4	67.9	0.078	5.7	70.4	60.9	0.075	5.7	60.6
✓	✓		72.8	0.130	10.1	120.1	70.5	0.122	5.7	97.9	71.7	0.121	5.7	97.5
✓	✓	✓	71.3	0.122	6.6	107.8	70.1	0.112	5.7	91.1	71.9	0.122	5.7	89.7

(a) Sampling Smoothing Factor

Factor	mIoU ↑	$\text{Inf}_{\text{Lat.}}$ ↓	$\text{Peak}_{\text{Mem.}}$ ↓
1	**72.4**	0.124	51.5
2	70.4	0.122	42.2
3	67.3	**0.118**	32.7
4	63.4	0.122	**24.6**

(b) Confidence Intensity Threshold

Thr.	mIoU ↑	$\text{Inf}_{\text{Lat.}}$ ↓	$\text{Peak}_{\text{Mem.}}$ ↓
0.1	**70.9**	0.121	10.1
0.2	70.4	0.122	9.75
0.3	68.7	**0.116**	**9.52**
0.4	66.4	0.117	9.60

(c) Prompt Spacing Factor

Factor	mIoU ↑	$\text{Inf}_{\text{Lat.}}$ ↓	$\text{Peak}_{\text{Mem.}}$ ↓
4	**72.7**	0.123	10.0
5	71.6	0.123	9.88
6	70.4	0.122	**9.75**
7	68.9	**0.117**	9.82

(d) Prompt Elimination Threshold

Thr.	mIoU ↑	$\text{Mask}_{\text{Lat.}}$ ↓	$\text{Ratio}_{\text{elim.}}$ ↑
1.25	68.4	**0.671**	**51.5**
1.3	70.4	0.799	42.3
1.35	71.6	0.93	32.7
1.4	**72.2**	1.04	24.6

Performance Hightlights

- Achieves highest mIoU scores across all datasets and encoders, outperforming methods using bounding box prompts.
- Demonstrates excellent computational efficiency with fast inference (0.122s latency) and low memory usage (6.6MB peak).
- Surpasses both baseline methods (AMG-S, AMG-D) and advanced approaches (OAS), achieving better balance between segmentation accuracy and prompt generation efficiency.
- Successfully maintains high performance while keeping resource usage within practical limits, suitable for real-world applications.

Parameter Analysis

- *Sampling Smoothing Factor* impacts the coverage area of Gaussian filtering - larger factors provide stronger smoothing and reduce memory usage during processing.
- *Confidence Intensity Threshold* and *Prompt Spacing Factor* optimize point prompt generation from confidence maps, ensuring accurate and reliable point selection for critical areas.
- *Prompt Elimination Threshold* controls the balance between efficiency and accuracy - lower thresholds increase prompt removal ratio for faster mask generation with minimal accuracy trade-off.

Prompt Methods	Latency (Sec) ↓ (SA-1B)	Latency (Sec) ↓ (COCO)	Latency (Sec) ↓ (LVIS)	Peak Mem ↓ (GB)
OAS(Box)	1.16	1.01	0.99	0.78
OAS(Central)	1.32	1.21	1.23	0.78
AoP-SAM	**0.65**	**0.77**	**0.84**	**0.042**

Experiment on Nvidia Jetson Orin Nano Edge GPU