Rethinking Interactive Image Segmentation with Low Latency, High Quality, and Diverse Prompts

Qin Liu, Jaemin Cho, Mohit Bansal, Marc Niethammer
UNC Chapel Hill

UNC
DEPARTMENT OF
COMPUTER SCIENCE

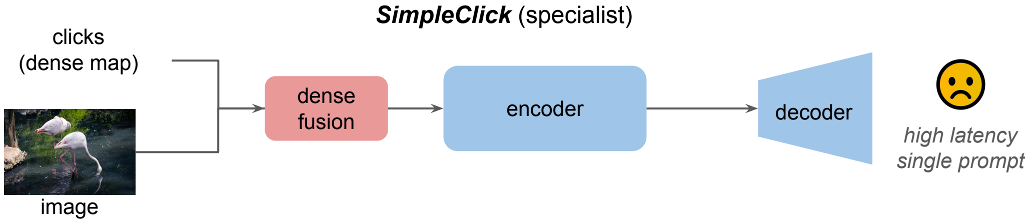
CVPR
JUNE 17-21, 2024
SEATTLE, WA

Interactive image segmentation aims to extract objects with human interactions, such as clicks and scribbles.

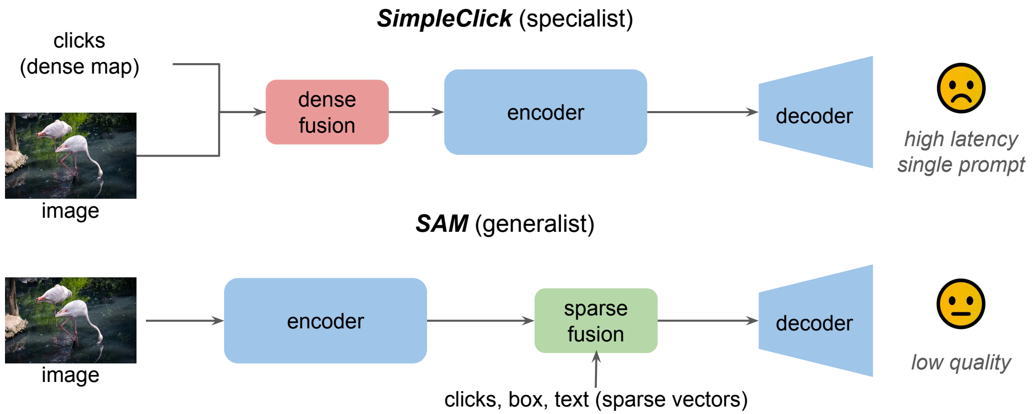
https://github.com/saic-vul/ritm_interactive_segmentation

Interactive segmentation with low latency, high quality, and diverse prompts remains challenging for existing *specialist* and *generalist* models.

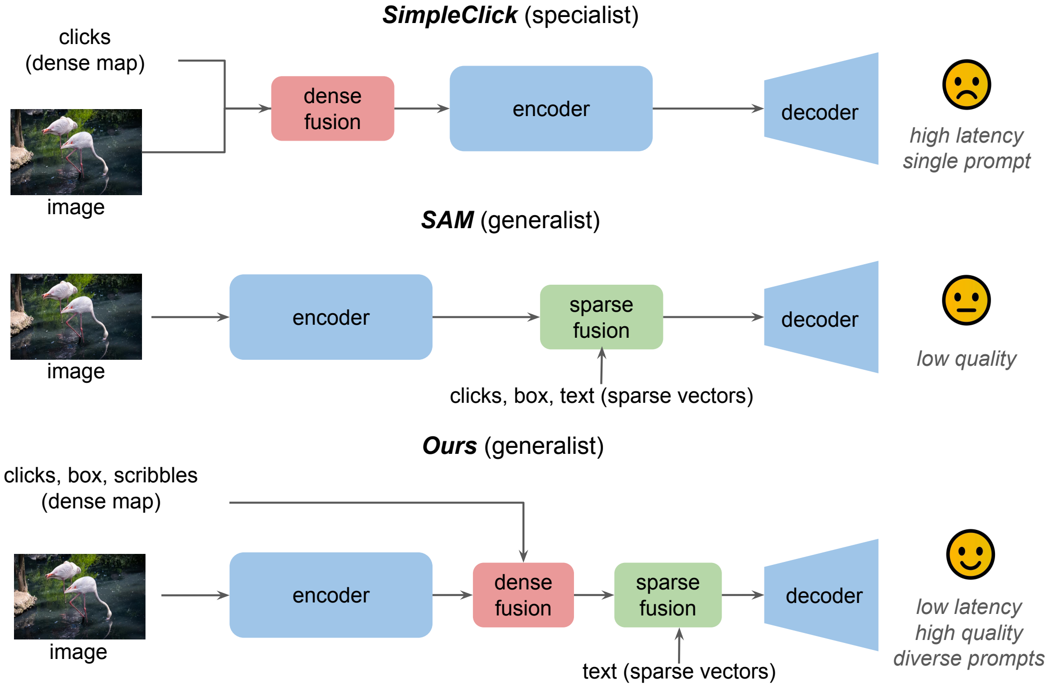
Interactive segmentation with low latency, high quality, and diverse prompts remains challenging for existing *specialist* and *generalist* models.



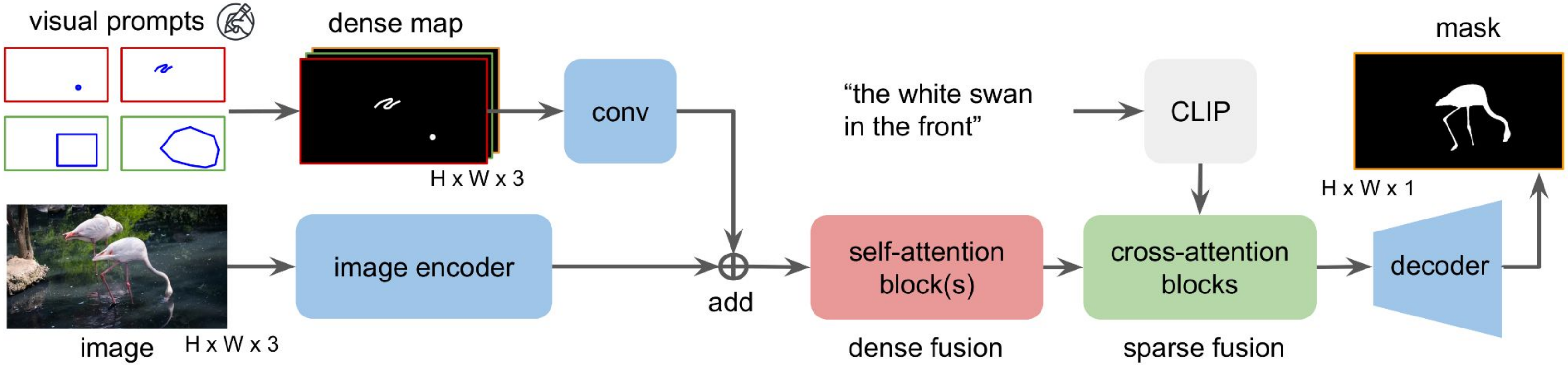
Interactive segmentation with low latency, high quality, and diverse prompts remains challenging for existing *specialist* and *generalist* models.



Interactive segmentation with low latency, high quality, and diverse prompts remains challenging for existing *specialist* and *generalist* models.



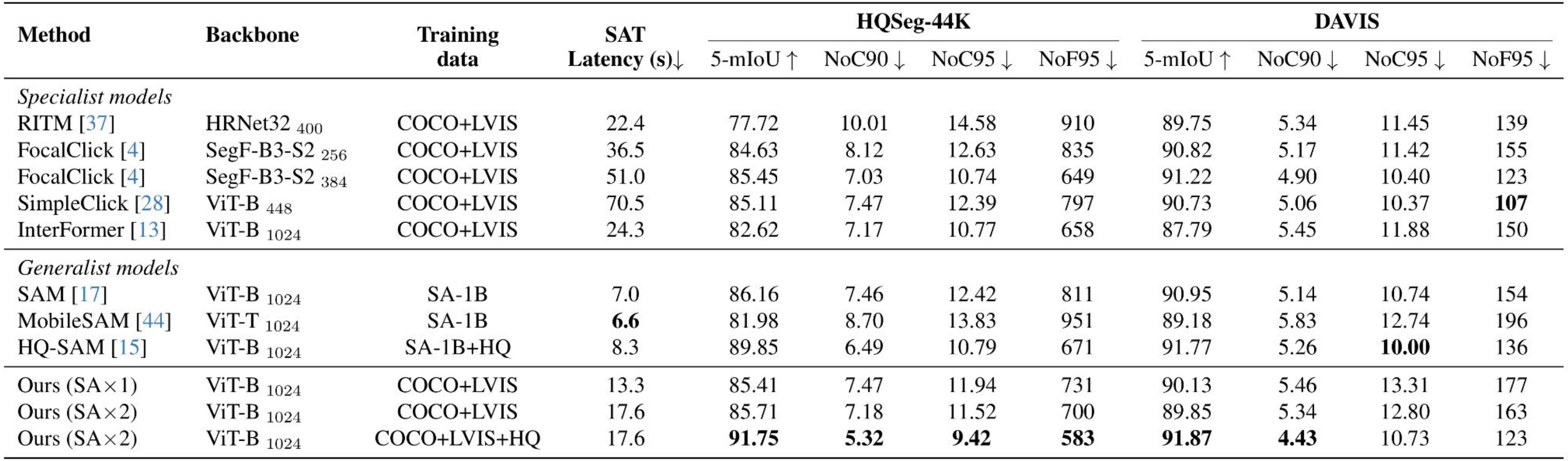
Network Architecture



Experiments

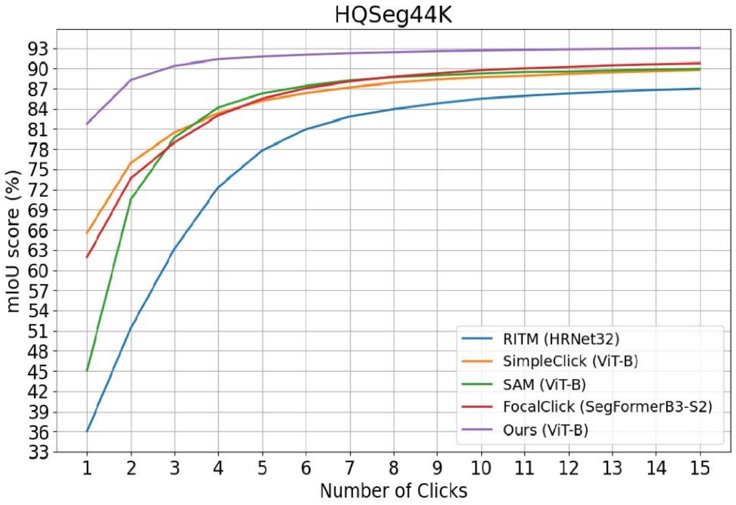
- Benchmarks
  - COCO+LVIS: 118K training images (1.2M instances)
  - HQSeg-44K: 44320 training images with extremely fine-grained image masks
  - DAVIS: 345 high-quality and high-resolution images
  - ssTEM: 20 high-resolution cell images
  - BraTS: 369 brain tumor images from 69 MRI volumes
- Baselines
  - Specialists: RITM, FocalClick, SimpleClick, InterFormer
  - Generalists: SAM, MobileSAM, HQ-SAM

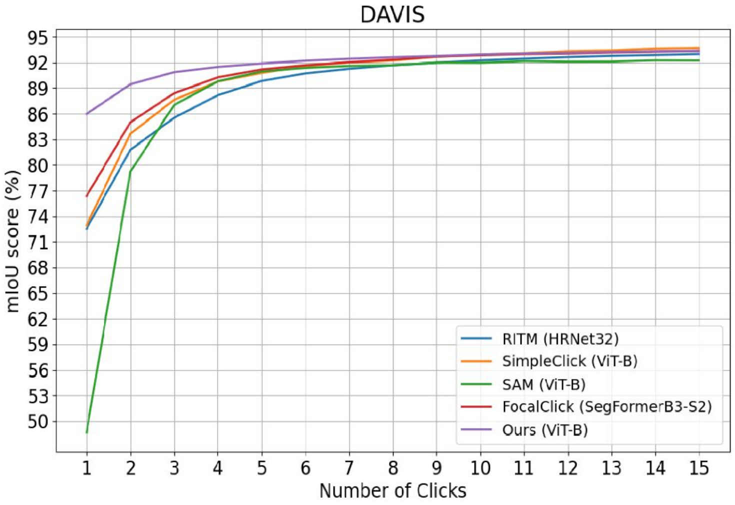
Quantitative Comparison on HQSeg-44K and DAVIS



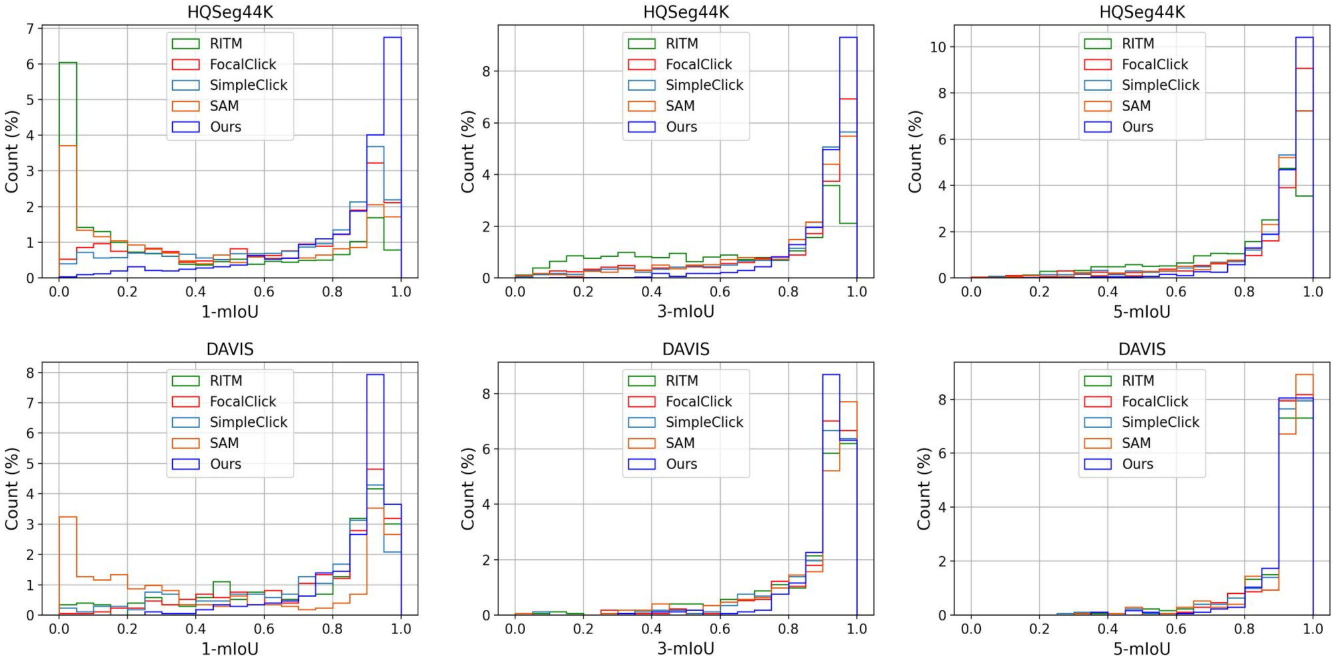
Method	**Backbone**	**Training data**	**SAT Latency (s)**↓	**HQSeg-44K**				**DAVIS**			
				5-mIoU ↑	NoC90 ↓	NoC95 ↓	NoF95 ↓	5-mIoU ↑	NoC90 ↓	NoC95 ↓	NoF95 ↓
Specialist models											
RITM [37]	HRNet32 $_{400}$	COCO+LVIS	22.4	77.72	10.01	14.58	910	89.75	5.34	11.45	139
FocalClick [4]	SegF-B3-S2 $_{256}$	COCO+LVIS	36.5	84.63	8.12	12.63	835	90.82	5.17	11.42	155
FocalClick [4]	SegF-B3-S2 $_{384}$	COCO+LVIS	51.0	85.45	7.03	10.74	649	91.22	4.90	10.40	123
SimpleClick [28]	ViT-B $_{448}$	COCO+LVIS	70.5	85.11	7.47	12.39	797	90.73	5.06	10.37	**107**
InterFormer [13]	ViT-B $_{1024}$	COCO+LVIS	24.3	82.62	7.17	10.77	658	87.79	5.45	11.88	150
Generalist models											
SAM [17]	ViT-B $_{1024}$	SA-1B	7.0	86.16	7.46	12.42	811	90.95	5.14	10.74	154
MobileSAM [44]	ViT-T $_{1024}$	SA-1B	**6.6**	81.98	8.70	13.83	951	89.18	5.83	12.74	196
HQ-SAM [15]	ViT-B $_{1024}$	SA-1B+HQ	8.3	89.85	6.49	10.79	671	91.77	5.26	**10.00**	136
Ours (SA×1)	ViT-B $_{1024}$	COCO+LVIS	13.3	85.41	7.47	11.94	731	90.13	5.46	13.31	177
Ours (SA×2)	ViT-B $_{1024}$	COCO+LVIS	17.6	85.71	7.18	11.52	700	89.85	5.34	12.80	163
Ours (SA×2)	ViT-B $_{1024}$	COCO+LVIS+HQ	17.6	**91.75**	**5.32**	**9.42**	**583**	**91.87**	**4.43**	10.73	123

Quantitative Comparison on HQSeg-44K and DAVIS

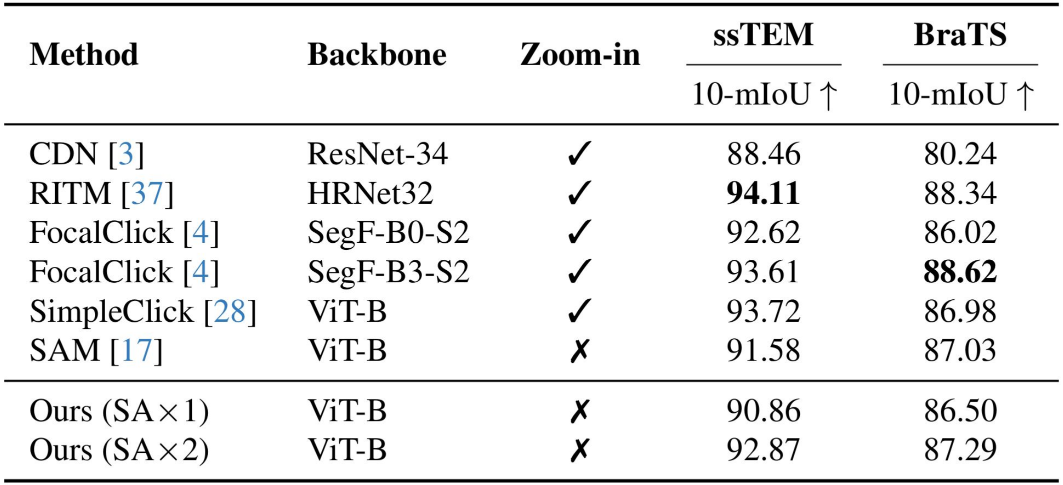




Quantitative Comparison on HQSeg-44K and DAVIS

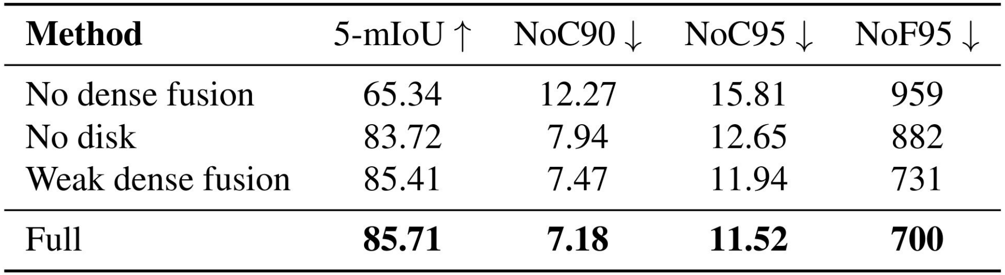


Out-of-Domain Evaluation on Medical Images



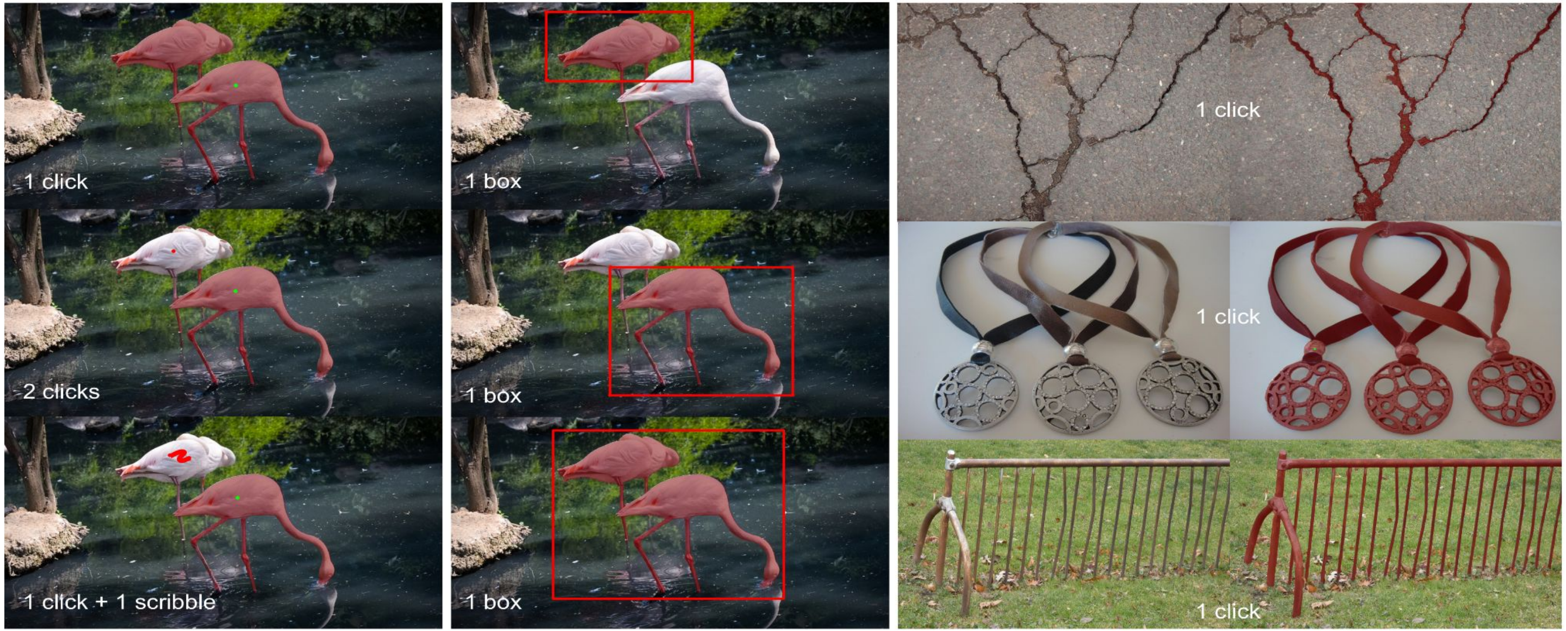
Method	**Backbone**	**Zoom-in**	**ssTEM**	**BraTS**
			10-mIoU ↑	10-mIoU ↑
CDN [3]	ResNet-34	✓	88.46	80.24
RITM [37]	HRNet32	✓	**94.11**	88.34
FocalClick [4]	SegF-B0-S2	✓	92.62	86.02
FocalClick [4]	SegF-B3-S2	✓	93.61	**88.62**
SimpleClick [28]	ViT-B	✓	93.72	86.98
SAM [17]	ViT-B	✗	91.58	87.03
Ours (SA×1)	ViT-B	✗	90.86	86.50
Ours (SA×2)	ViT-B	✗	92.87	87.29

Ablations

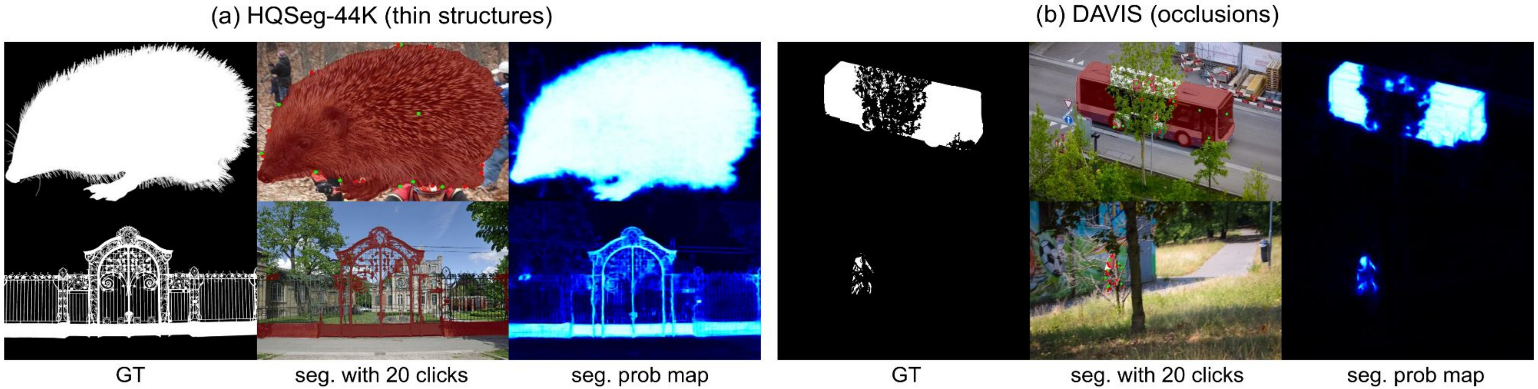


Method	5-mIoU ↑	NoC90 ↓	NoC95 ↓	NoF95 ↓
No dense fusion	65.34	12.27	15.81	959
No disk	83.72	7.94	12.65	882
Weak dense fusion	85.41	7.47	11.94	731
Full	**85.71**	**7.18**	**11.52**	**700**

Qualitative Results with Diverse Prompts



Failure Patterns and Future Work



(a) HQSeg-44K (thin structures)

(b) DAVIS (occlusions)