SegVol: Universal and Interactive Volumetric Medical Image Segmentation

Yuxin Du[1,2], Fan Bai[1,2,3], Tiejun Huang[2,4], Bo Zhao[1,2†]

[1]Shanghai Jiao Tong University
[2]Beijing Academy of Artificial Intelligence
[3]The Chinese University of Hong Kong
[4]Peking University

[†]Corresponding author: Bo Zhao <bo.zhao@sjtu.edu.cn>

GitHub: https://github.com/BAAI-DCAI/SegVol

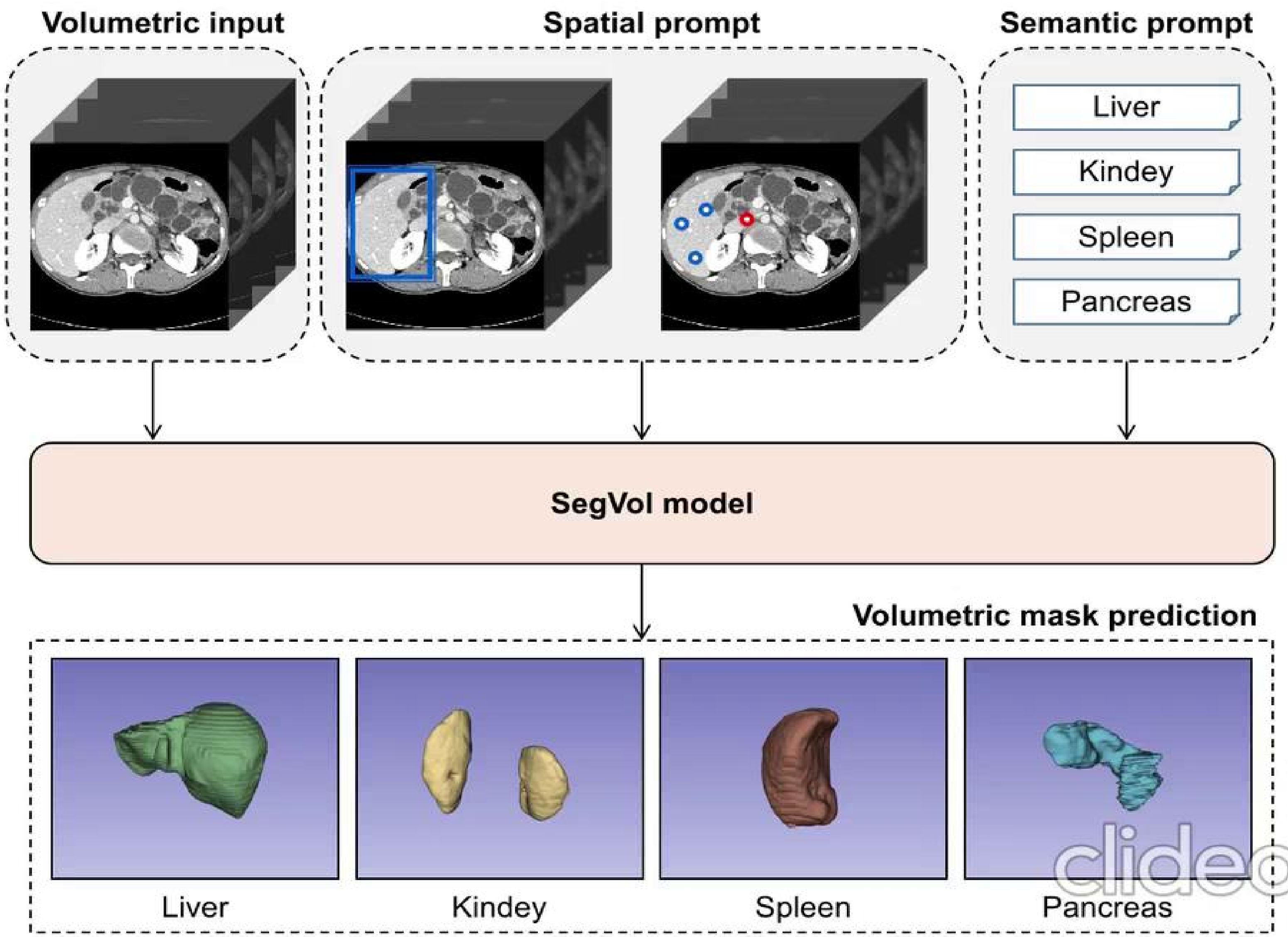
NeurIPS 2024 [Spotlight]

Content

- Start with a video demo
- Challenges
- From challenges to solutions
- Experiments

Start with a video demo of SegVol

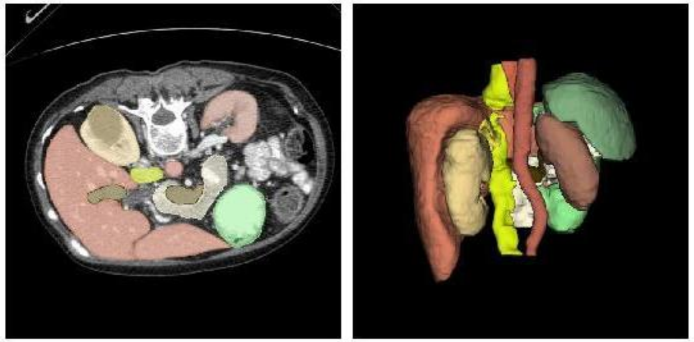
HuggingFace demo: https://huggingface.co/spaces/yuxindu/SegVol



Challenges

Challenges for
the universal and interactive volumetric medical image segmentation model

- Scattered and scale-limited datasets
    - Partial label problem

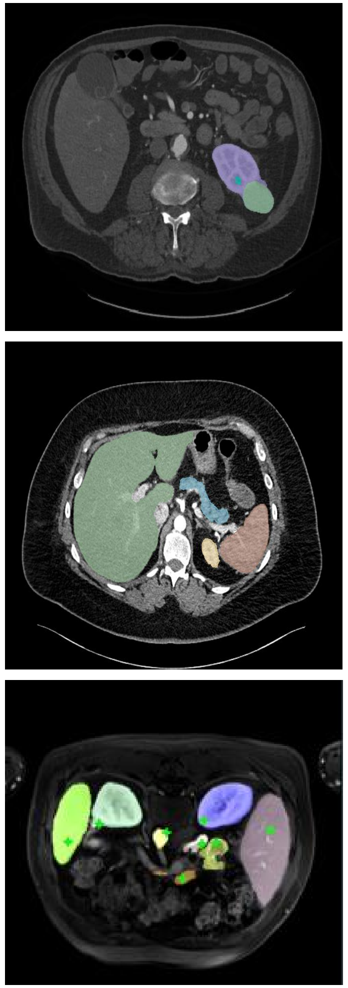


- Lack of strong 3D CT encoder
- Challenges of simple interaction in 3D volumes
- Prompt ambiguation
  one prompt that can be understood in two or more possible ways

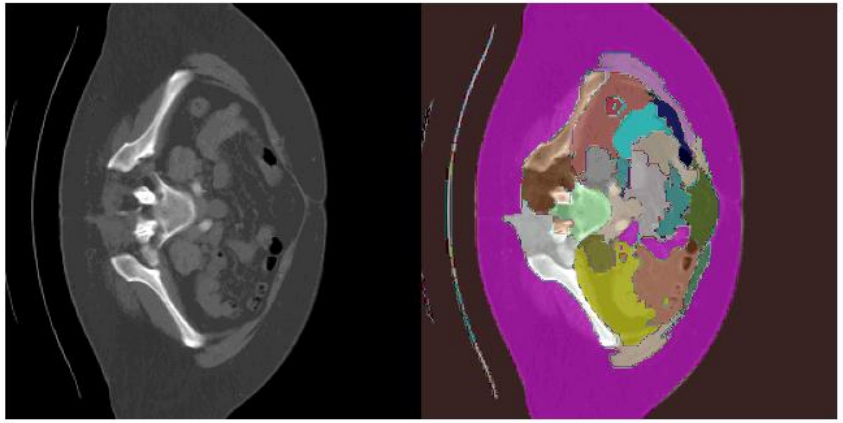
From challenges to solutions: Scattered and scale-limited datasets & Partial label problem

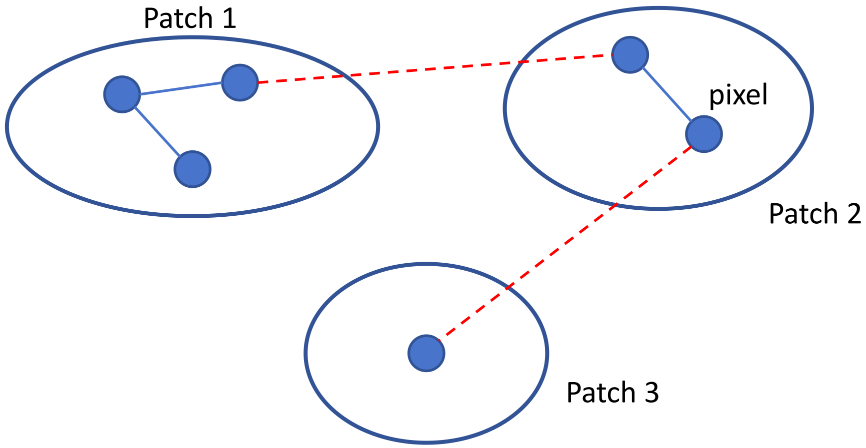
Table 4: Information of datasets involved in supervised fine-tuning and experiments.

Dataset	Anatomical Targets	Category Number	Trainset Volumes
3D-IRCADB[55]	Liver and liver tumor	47	20
AbdomenCT-1k[45]	Liver, kidney, spleen, and pancreas	4	1000
AMOS22[44]	Abdominal organs	15	240
BTCV[52]	Abdominal organs	13	30
CHAOS[40, 41, 42]	Abdominal organs	1	20
CT-ORG[33, 34, 24, 35]	Brain, lung, bones, liver, kidney, and bladder	6	140
FLARE22[56, 57]	Thoracic and abdominal organs	13	50
HaN-Seg[43]	Organs of the head and neck	30	42
KiPA22[47, 48, 49, 50]	Kidney, renal tumor, artery, and vein	4	70
KiTS19[51]	Kidney and kidney tumor	2	210
KiTS23[46]	Kidney, kidney tumor, and kidney cyst	3	489
LUNA16[36]	Left lung, right lung, and trachea	3	888
MSD-Colon[56]	Colon tumor	1	126
MSD-HepaticVessel[56]	Hepatic vessel and liver tumor	2	303
MSD-Liver[56]	Liver and liver tumor	2	131
MSD-lung[56]	Lung tumor	1	63
MSD-pancreas[56]	Pancreas and pancreas tumor	2	281
MSD-spleen[56]	Spleen	1	41
Pancreas-CT[53, 54, 35]	Pancreas	1	82
QUBIQ[63]	Kidney, pancreas, and pancreas lesion	3	82
SegTHOR[75]	Heart, trachea, aorta, and esophagus	4	40
SLIVER07[62]	Liver	1	20
TotalSegmentator[58]	Organs of the whole body	104	1203
ULS23(novel annotated set)[74]	Various lesions	-	1618
VerSe19[59, 60, 61]	Vertebrae	28	80
VerSe20[59, 60, 61]	vertebrae	28	61
WORD[64]	Thoracic and abdominal organs	16	100



raw CT → pseudo labels





Felzenszwalb-Huttenlocher(FH) algorithm

From challenges to solutions: Lack of strong 3D CT encoder

Radiopaedia:https://radiopaedia.org/

Pretrain data:

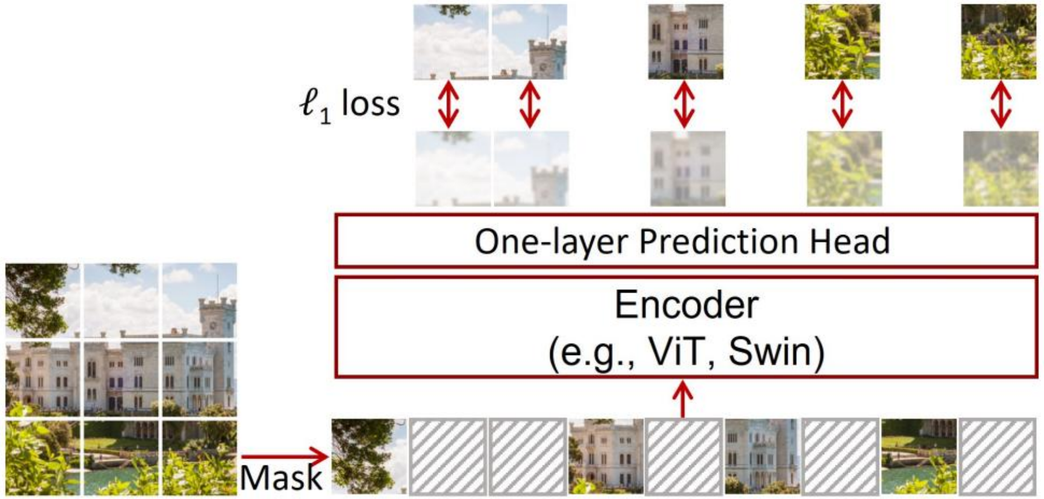
- 90K unlabeled CT scans collected from Radiopaedia + ~6K labeled CT scans

Pretrain framework:

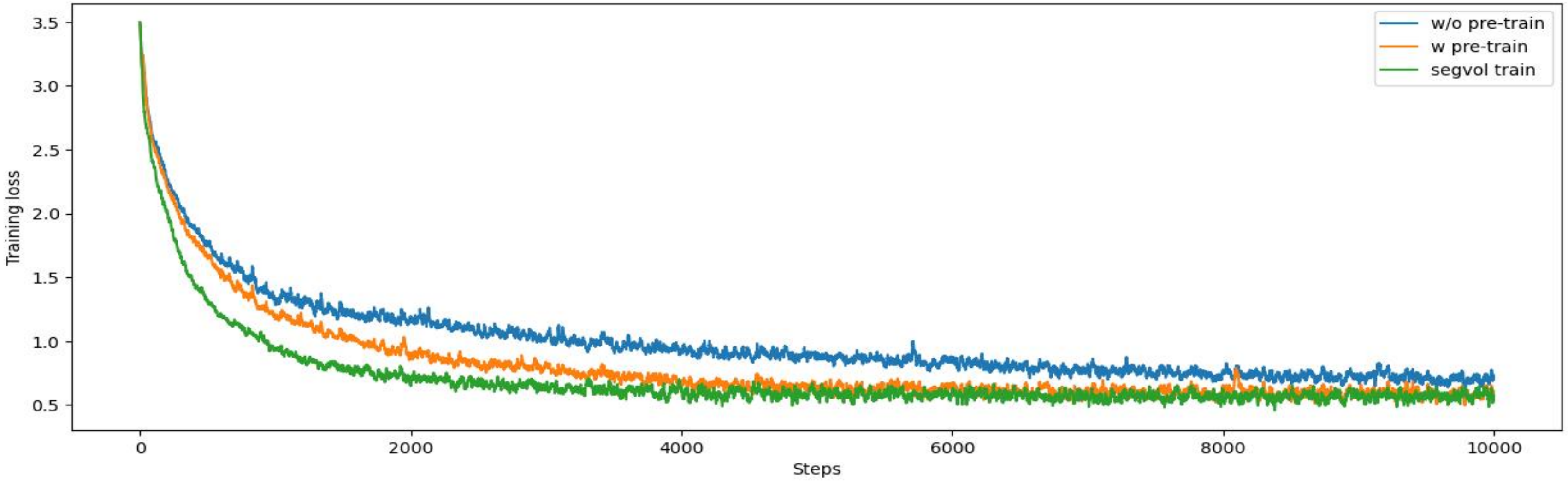
- SimMIM (MAE like)

Pretrain performance validation:

- UNETR finetuning on AMOS22 for 10K steps



Model	Encoder	Dice score(%)
UNETR	w/o pre-train	67.12
UNETR	w pretrain	79.10



From challenges to solutions: Challenges of simple interaction in 3D volumes

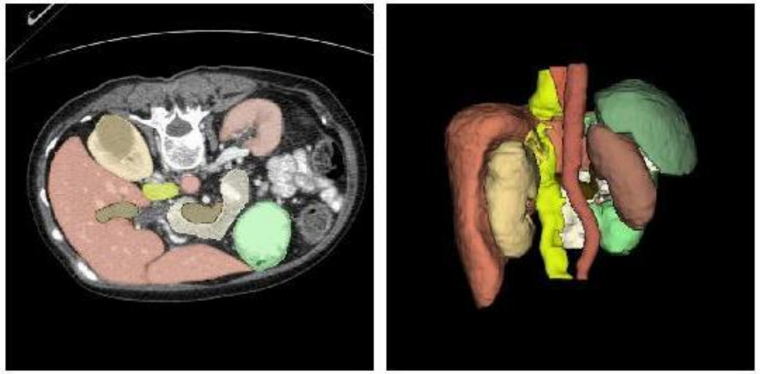
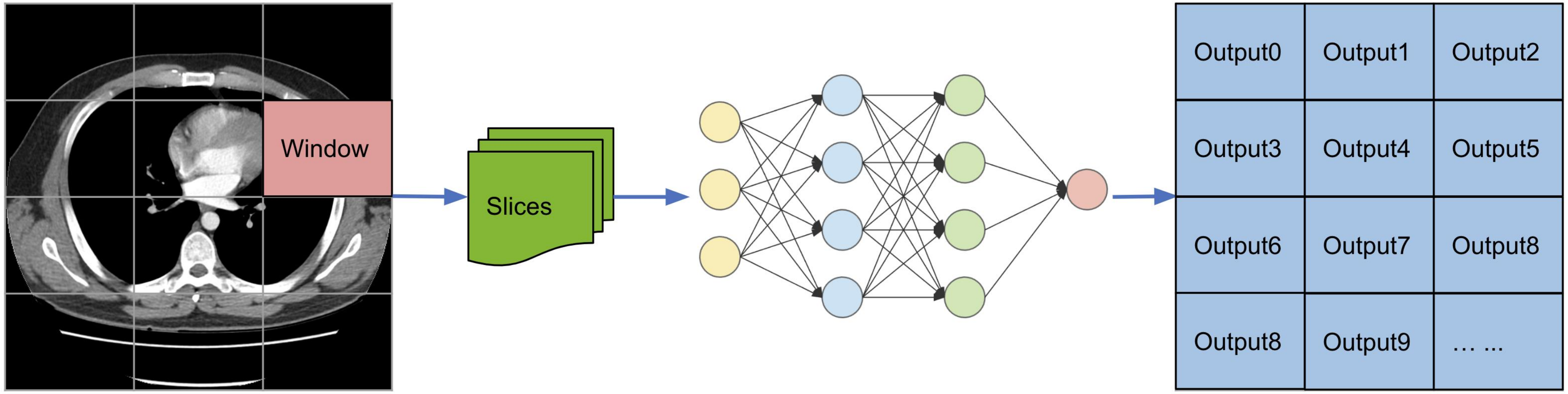
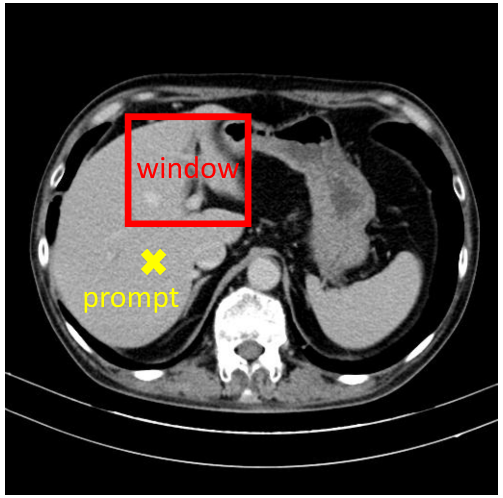


Table 1: The different settings and functions of SAM-like interactive segmentation methods.

Method	Image Domain	Dimension	Training	Prompt Type: Point	Prompt Type: Bbox	Prompt Type: Text	Inference Input
SAM[28]	Natural	2D	Full-Param	✓	✓	✓	1024×1024
MedSAM[29]	Medical	2D	Decoder	✗	✓	✗	1024×1024
SAM-Med2D[38]	Medical	2D	Adapter	✓	✓	✗	1024×1024
SAM-Med3D[39]	Medical	3D	Full-Param	✓	✗	✗	128×128×128
OURS	**Medical**	**3D**	**Full-Param**	✓	✓	✓	**Full Resolution**

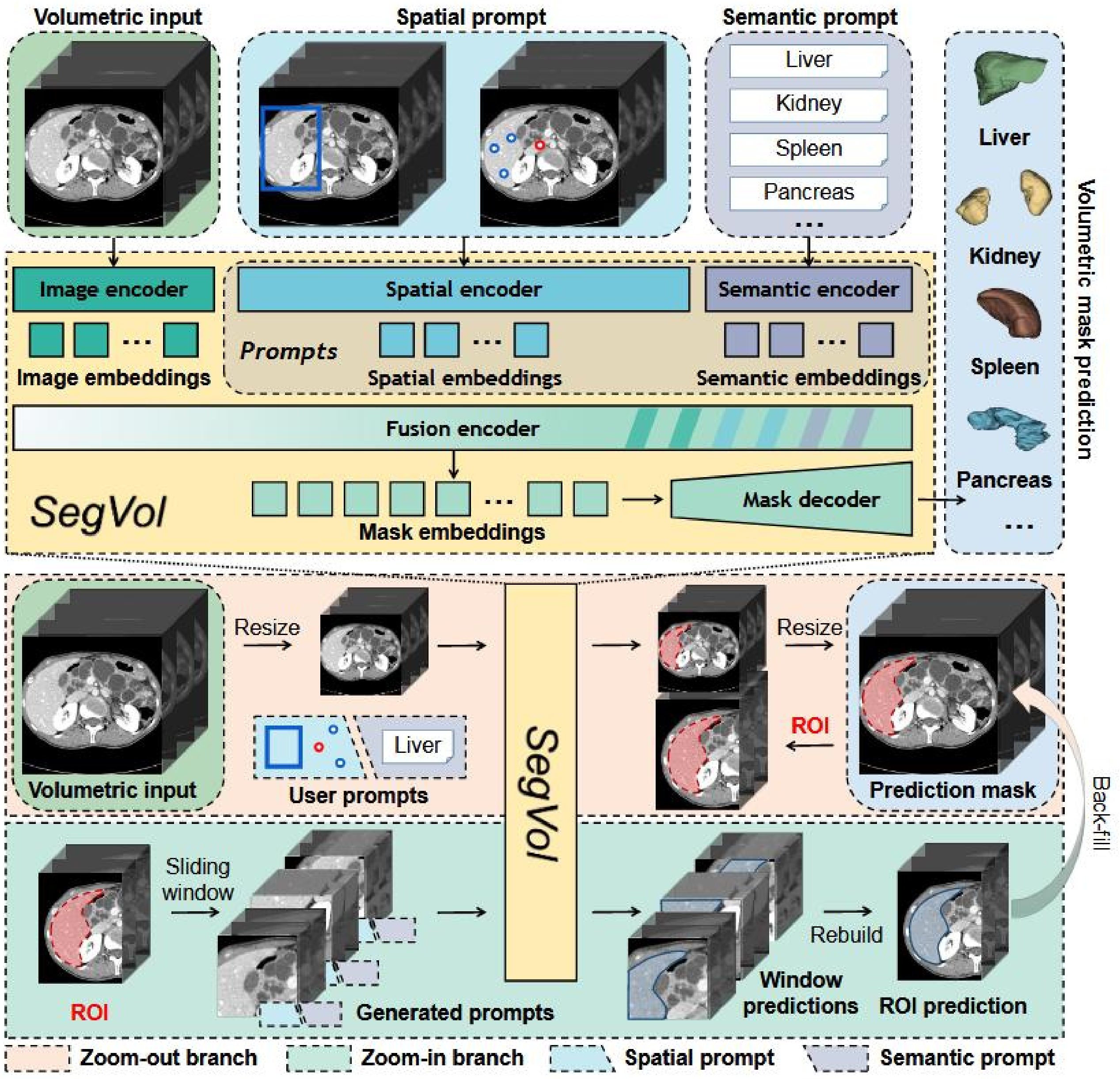




(1) Generate slices from window (2) Construct batches (3) Execute on network (4) Connect all outputs

Receptive Field of model is limited!

From challenges to solutions: Challenges of simple interaction in 3D volumes



Model architecture

- Image encoder: 3D ViT (pretrained)
- Spatial encoder: following SAM
- Semantic encoder: CLIP text encoder
- Fusion encoder: cross attention layers
- Mask decoder: deconvolution layers and interpolation

Zoom-out (global):

input: resized global volume + global prompts
output: resized global mask prediction

Zoom-in (local):

input: local volumes from sliding window
+ generated prompts
output: local mask predictions for each volume

From challenges to solutions: Prompt ambiguation

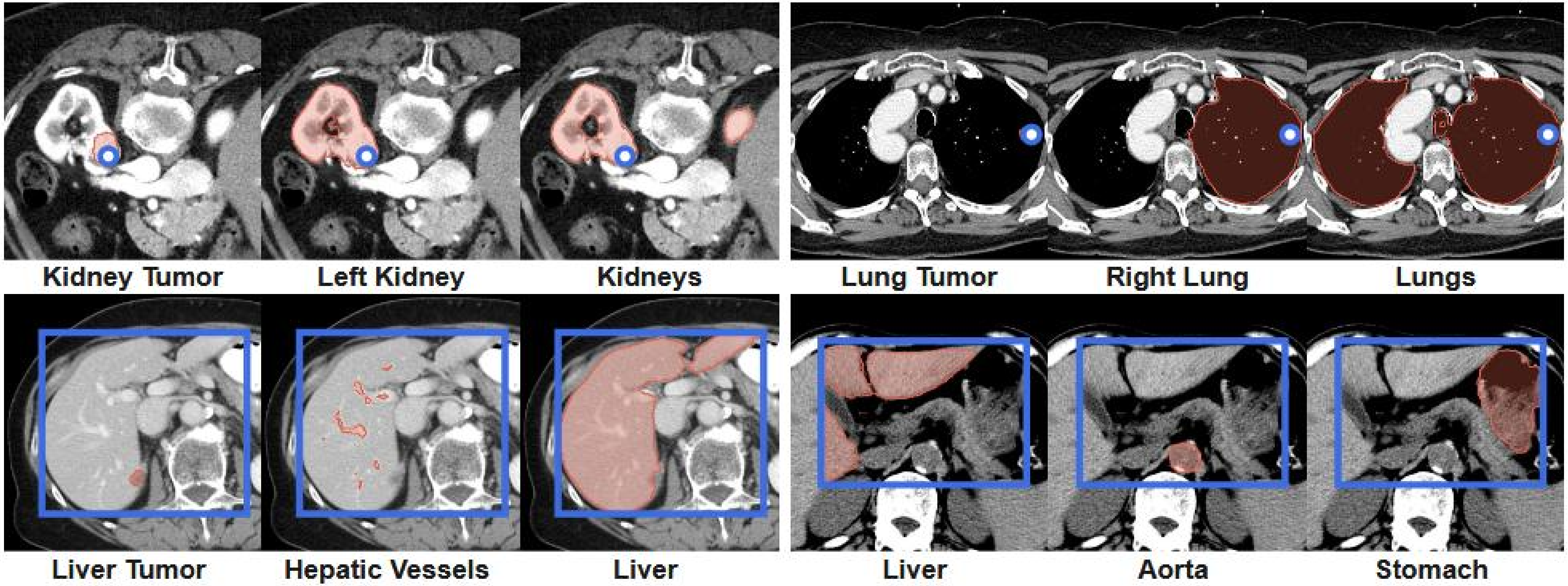


Figure 4: The four cases demonstrate that semantic-prompt can clarify the ambiguity of spatial-prompt and avoid multi-plausible outputs. Each image shows the segmentation result of SegVol using the spatial-prompt, i.e. point or bounding box, and semantic-prompt, i.e. the caption below the image.

From challenges to solutions: Prompt ambiguation

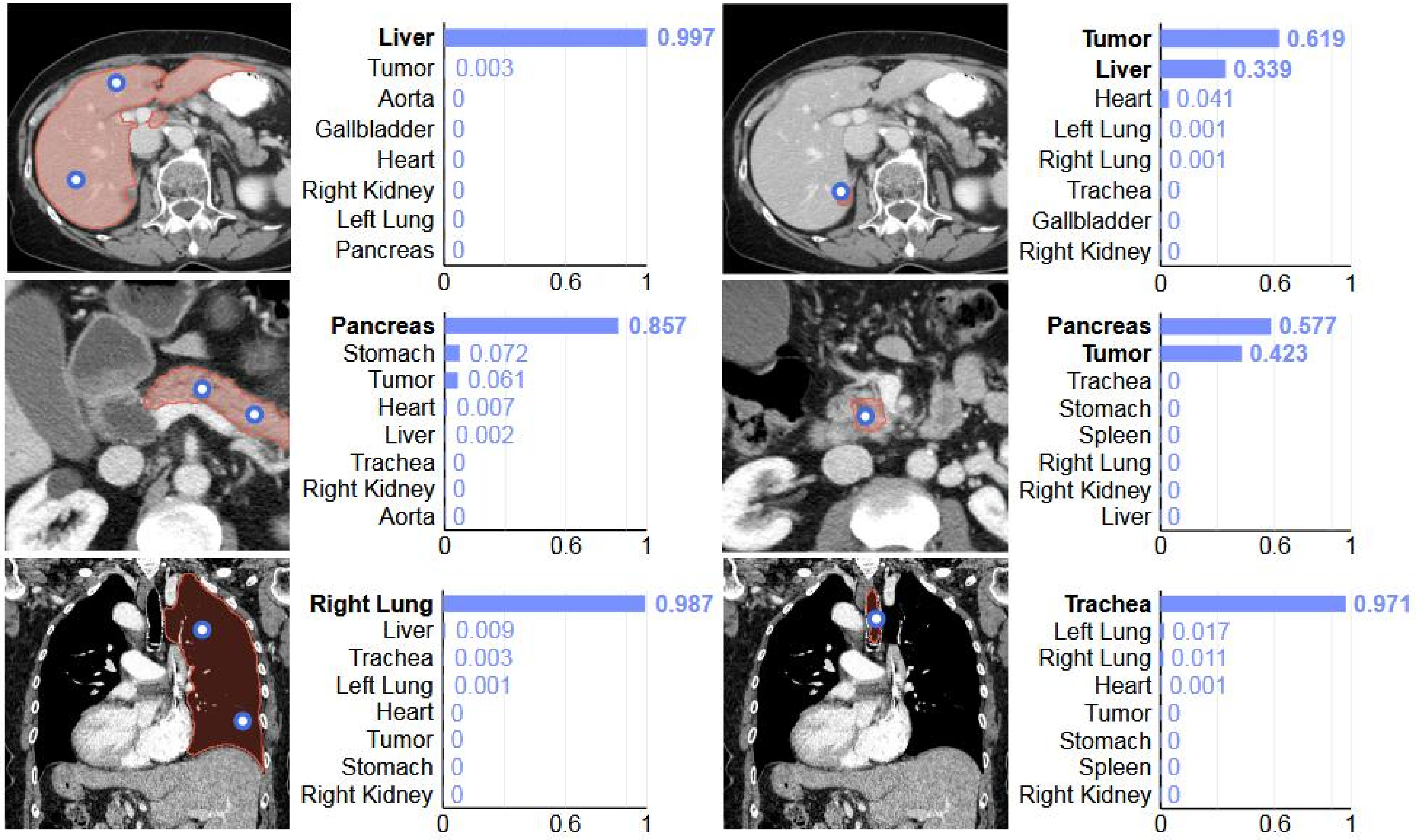


Figure 5: We identify the semantic categories of the spatial-prompt segmentation results. Each image shows the spatial-prompt and the mask prediction. The bar charts rank the top 8 semantic categories with the highest classification probabilities. The results show that SegVol is capable of identifying the anatomical category of the segmentation mask using spatial prompts.

Review of Challenges standing in the way

Challenges standing in the way to
the universal and interactive volumetric medical image segmentation model

- Scattered and scale-limited datasets √ collected large-scale dataset
  - Partial label problem √ pseudo label
- Lack of strong 3D CT encoder √ large-scale pretraining
- Challenges of simple interaction in 3D volumes √ zoom-out-zoom-in mechanism
- Prompt ambiguation
  one prompt that can be understood in two or more possible ways
  √ semantic prompt

Experiments: major results

Table 2: Quantitative comparative experiment results for SegVol and other 5 SAM-like interactive segmentation methods settings in terms of the median value of Dice score.

Dataset	Category	SAM(Point) [28]	SAM(Bbox) [28]	SAM-MED2D [38]	SAM-MED3D [39]	MedSAM [29]	OURS
AMOS22 [44]	Aorta	0.7267	0.4362	0.8704	0.8102	0.3387	**0.9273**
	Bladder	0.4162	0.6281	0.8417	0.4338	0.6799	**0.9120**
	Duodenum	0.1554	0.3192	0.5066	0.3820	0.3066	**0.7402**
	Esophagus	0.2917	0.3541	0.5500	0.5174	0.3610	**0.7460**
	Gallbladder	0.2831	0.6161	0.7999	0.5643	0.6609	**0.8763**
	Adrenal gland(L)	0.0555	0.4222	0.5068	0.4584	0.3766	**0.7295**
	Left kidney	0.8405	0.8274	0.9325	0.8723	0.7909	**0.9489**
	Liver	0.7477	0.5124	0.6904	0.8801	0.6137	**0.9641**
	Pancreas	0.2127	0.3392	0.5656	0.5391	0.3217	**0.8295**
	Postcava	0.2042	0.5251	0.4436	0.6683	0.5211	**0.8384**
	Prostate uterus	0.2344	0.6986	0.7518	0.6231	0.7739	**0.8557**
	Adrenal gland(R)	0.0452	0.3642	0.1681	0.3708	0.3855	**0.6994**
	Right kidney	0.8459	0.8215	0.9077	0.8632	0.7851	**0.9505**
	Spleen	0.5936	0.6536	0.9267	0.8591	0.7038	**0.9589**
	Stomach	0.4229	0.3883	0.5399	0.4576	0.4378	**0.9123**
	Average	0.4050	0.5271	0.6668	0.6200	0.5371	**0.8593**
ULS23 [74]	DeepLesion3D	0.3686	0.7473	0.3258	0.2386	**0.7680**	0.7065
	BoneLesion	0.4461	0.6671	0.1947	0.4447	0.6896	**0.6920**
	PancreasLesion	0.0675	0.5579	0.5548	0.5526	0.6561	**0.7265**
	Average	0.2941	0.6574	0.3584	0.4120	**0.7046**	**0.7046**
SegTHOR [75]	Aorta	0.2744	0.3894	0.8077	0.7703	0.3278	**0.8439**
	Esophagus	0.0348	0.2046	0.3578	0.6394	0.2196	**0.7201**
	Heart	0.6695	0.8876	0.6012	0.8325	**0.8924**	0.8172
	Trachea	**0.9147**	0.1611	0.8306	0.8485	0.1261	0.8807
	Average	0.4734	0.4107	0.6493	0.7727	0.3915	**0.8155**

Experiments: major results

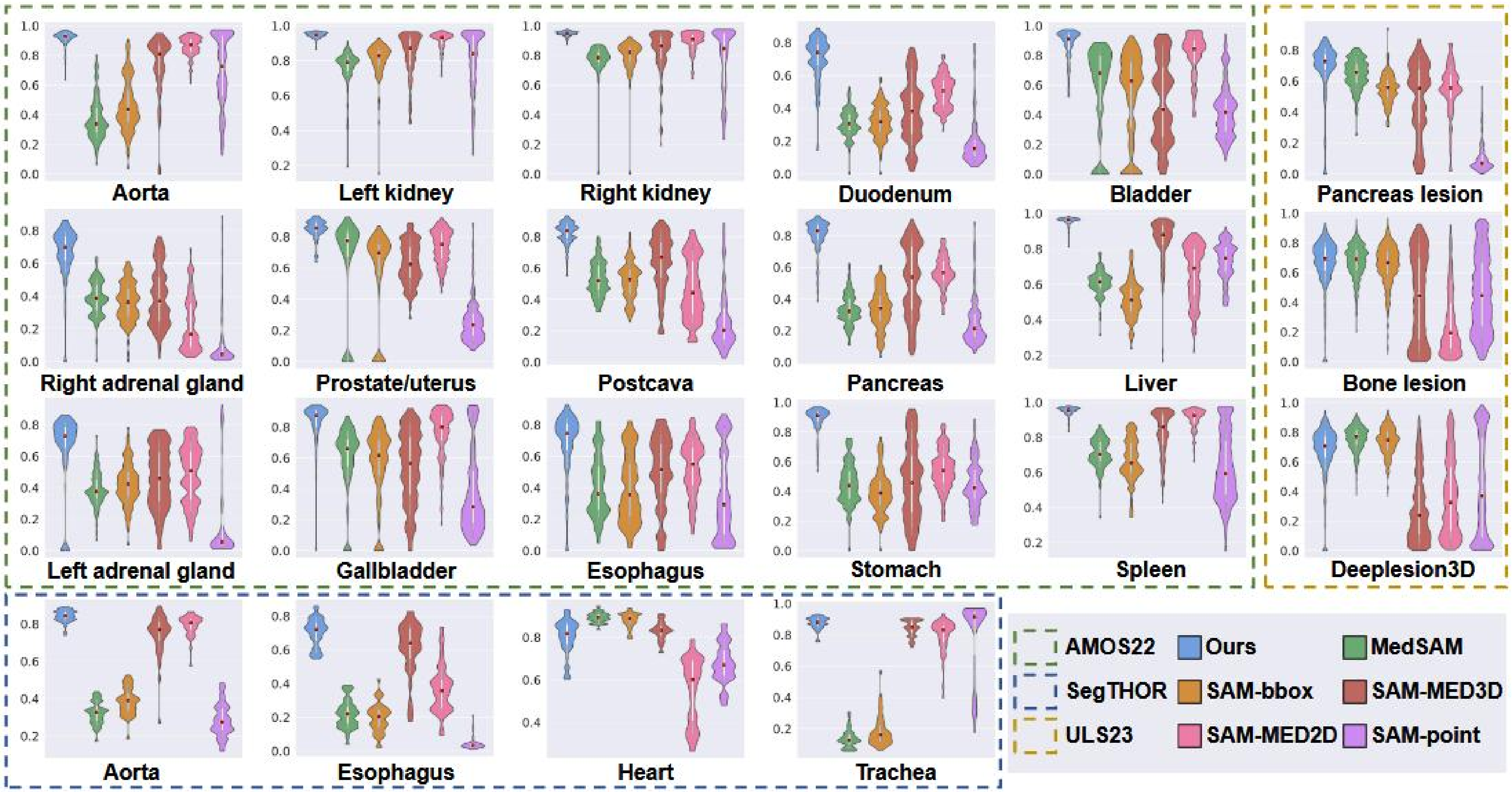


Figure 2: Violin plots for quantitative comparison experiment results of SegVol and SAM-like interactive methods[28, 38, 39, 29]. The vertical axis represents the Dice score.

Experiments: ablation results

A.

Table 3: Ablation experiment on the zoom-out-zoom-in mechanism.

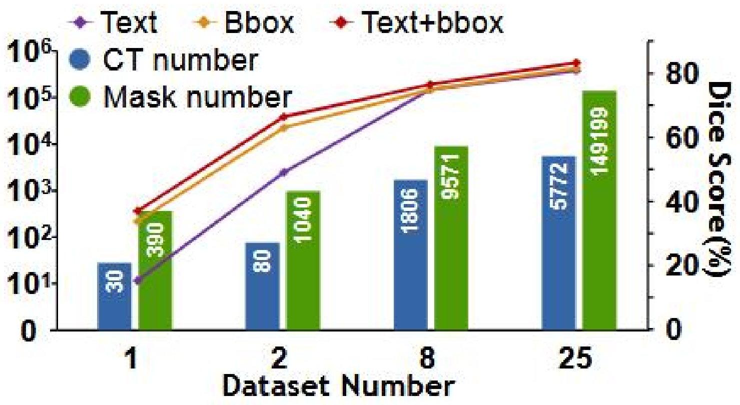
Mechanism	Dice Score Avg. ↑	Time Per Case Avg. ↓
Resize	0.4509	65 ms
Sliding window	0.6529	3331 ms
Zoom-out-zoom-in	0.7298	190 ms

Setting: splitted 20% test data of AMOS22

Conclusion:

Zoom-out-zoom-in can reduce the inference time and achieve competitive performance.

B.

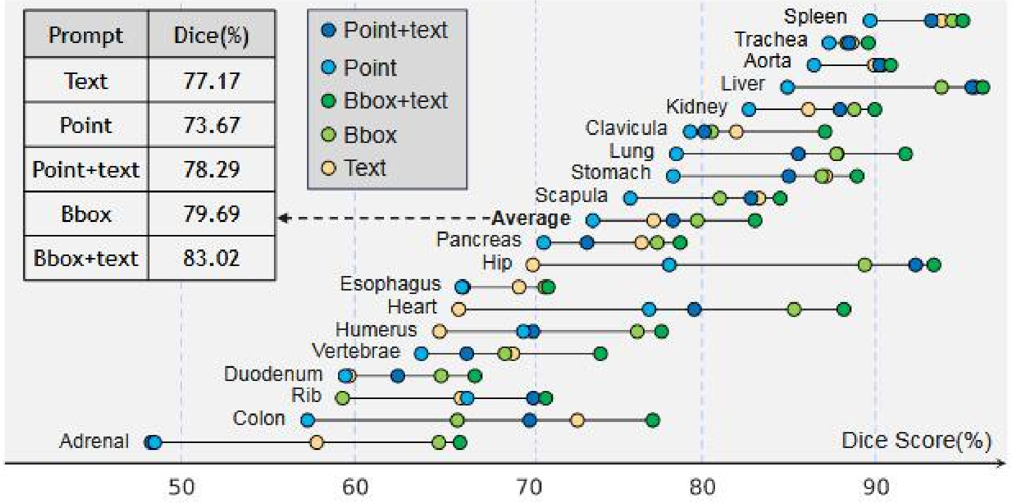


Setting: splitted 20% test data of BTCV as anchor test set

Conclusion:

Scaled dataset contributes a lot to the performance.

C.



Prompt	Dice(%)
Text	77.17
Point	73.67
Point+text	78.29
Bbox	79.69
Bbox+text	83.02

Setting: splitted 20% test data of ALL dataset

Conclusion:

semantic-prompts support spatial-prompts well.

END

Thank You!