2026 JULY 19–22 VANCOUVER, BC
JOINT AAPM | COMP MEETING

Zero-Shot Segment Anything Model 2 (SAM2)–Driven Interactive Delineation of Esophageal Clinical Target Volume: A Study of Prompting Strategy

Si Wu[1], Yuting Zhao[2], Xianggao Zhu[2], Lu Bai[3], Shaobin Wang[3], Yi Du[2]

[1]Institute of Medical Technology, Peking University Health Science Center, Beijing, China,
[2]Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education/Beijing), Department of Radiation Oncology, Peking University Cancer Hospital & Institute, Beijing, China,
[3]MedMind Technology Co., Ltd., Beijing, China

ABSTRACT

Purpose:
This study evaluated the potential of Segment Anything Model 2 (SAM2) for interactive CTV delineation and investigated effective prompting strategies.

Methods:
Zero-shot SAM2.1_hiera_l was evaluated in 83 patients receiving definitive radiotherapy for esophageal cancer. Ground truth-based box and mask prompts were compared using 2–10 uniformly placed slices, including both CTV boundaries. Optimal slices for the superior-performing mask prompts were then identified by exhaustive incremental selection and analyzed by normalized z-position and target area. Performance was assessed using Dice similarity coefficient (DSC) and 95th-percentile Hausdorff distance (HD95).

Results:
Performance improved as the number of prompts increased. Mask prompts consistently outperformed bounding box prompts. Optimal slice selection achieved significantly higher DSC than uniform placement across prompt numbers ($p<0.05$), although gains diminished beyond five prompts. With five optimally selected mask prompts, SAM2 achieved a DSC of 0.85±0.05 and an HD95 of 7.52±3.04 mm. Selected slices generally had larger target areas and were more frequently located in the superior portion of the CTV.

Conclusion:
Zero-shot SAM2 showed potential for interactive esophageal CTV delineation using only five annotated slices. Prioritizing slices with larger target areas, particularly in the superior CTV region, may improve prompting efficiency and segmentation accuracy.

CONTACT

Si Wu
Peking University Health Science Center
No. 38 Xueyuan Road, Haidian District, Beijing 100191, China
wusi_med@163.com

INTRODUCTION

Esophageal clinical target volume (CTV) delineation is complicated by its extensive superior–inferior extent and pronounced inter-slice variability in morphology and cross-sectional area[1]. These anatomical characteristics pose substantial challenges to consistent and accurate target definition and limit the performance of fully automatic segmentation models.

Segment Anything Model 2 (SAM2)[2] is a promptable segmentation foundation model originally developed for natural images and videos. By accepting prompts such as points, bounding boxes, or masks, it enables flexible user-guided segmentation. Its emerging applications in medical imaging have demonstrated promising potential for interactive delineation tasks[3-4].

This study aimed to evaluate the zero-shot performance of SAM2 for esophageal CTV delineation and to investigate the effects of prompt type and prompt number on segmentation accuracy. We further explored optimal prompt-slice configurations and analyzed their spatial and morphological characteristics to inform efficient interactive prompting strategies.

METHODS AND MATERIALS

Dataset
- 83 patients with stage I–IV esophageal cancer
- received definitive radiotherapy
- Planning CT images and clinically delineated CTV contours were used for evaluation

Model
- Zero-shot SAM2.1_hiera_l

Prompt strategy
- Prompt types: bounding box and mask
- Prompt number: 2–10 slices
- Prompt strategies: Uniform-Box, Uniform-Mask, and Optimal-Mask (incremental exhaustive selection)

Optimal slice analysis
- Two factors: the relative z-position of each slice along the superior–inferior extent of the CTV and its cross-sectional CTV area

Evaluation
- DSC and HD95
- Wilcoxon signed-rank test

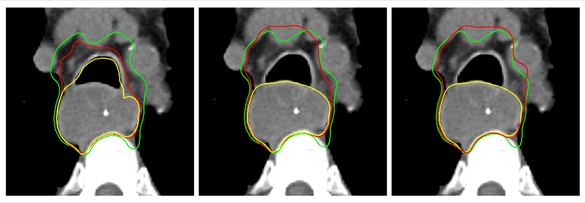

Figure 2. Representative SAM2 segmentation results with different prompting strategies. Left to right: Uniform-Box, Uniform-Mask, and Optimal-Mask prompting. Green: ground-truth CTV; Yellow: prediction with 2 prompt slices; Red: prediction with 5 prompt slices.

RESULTS

Prompt Strategy
- Segmentation performance improved with increasing prompt slices (Figure 1-2). Optimal-Mask DSC increased from 0.60 (K=2) to 0.85 (K=5) and 0.91 (K=10), with HD95 decreasing from 21.69 mm to 7.52 mm and 4.55 mm.
- Mask prompts outperformed bounding box prompts (Figure 1). At K=5, Uniform-Mask achieved higher DSC (0.82 vs. 0.67, $p<0.05$) and lower HD95 (12.05 vs. 13.90 mm) than Uniform-Box.
- Optimal-Mask prompting achieved the best performance. At K=5, Optimal-Mask achieved DSC 0.85 and HD95 7.52 mm, with limited gains beyond five prompts.

Optimal slice analysis
- At K=5, selected prompt slices showed a non-uniform z-position distribution, with higher selection frequency around z-rel = 0.9, corresponding to the superior CTV region (Figure 3). Selected slices also tended to have larger cross-sectional CTV areas.
- Across K=6–10, the z-position distributions remained similar, with selected slices consistently showing higher frequencies near larger z-rel values (Figure 4).

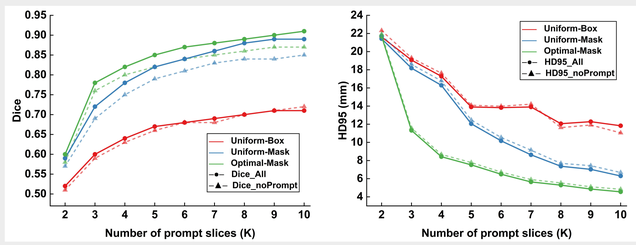

Figure 1. DSC and HD95 versus the number of prompt slices (K) for different prompt configurations. "All" includes all slices, "noPrompt" excludes prompt slices. Uniform-Box: uniformly placed GT-based bounding box prompts; Uniform-Mask: uniformly placed GT mask prompts; Optimal-Mask: performance-based selection of GT mask prompts.

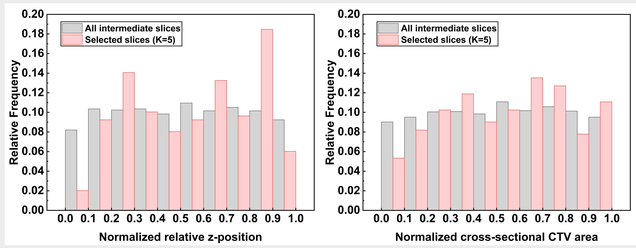

Figure 3. Relative frequency distributions of selected prompt slices (K=5) versus all intermediate slices in terms of normalized z-position (left) and cross-sectional CTV area (right). The normalized relative z-position denotes the relative longitudinal location of each slice within the target volume, smaller values indicate slices closer to the inferior target region, while larger values indicate slices closer to the superior region. The normalized cross-sectional CTV area represents the relative cross-sectional area of each CTV slice, with larger values indicating larger CTV area.

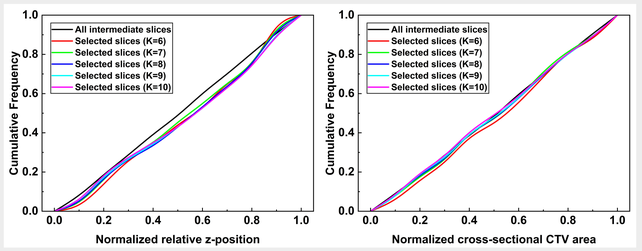

Figure 4. Cumulative frequency distributions of selected prompt slices (K=6-10) versus all intermediate slices

DISCUSSION

This study leveraged SAM2 to investigate prompting strategies for interactive CTV delineation. Different prompt types, prompt numbers, and slice selection strategies were systematically evaluated to identify efficient prompting approaches for volumetric segmentation. Overall, mask prompts improved segmentation performance, while five prompts provided a practical balance between accuracy and interaction burden.

The superior performance of mask prompts may be attributed to the richer spatial information provided to SAM2, as masks offer more accurate target shape and boundary cues than bounding boxes. In addition, increasing the number of prompt slices may improve propagation by reducing the distance between annotated anchors. SAM2 propagates information across frames, and denser prompts provide more frequent guidance with fewer propagation errors.

The observed difference between uniform and optimal prompting suggests that prompt location affects inter-slice propagation. The preference for upper CTV slices may be related to the greater anatomical variability and complex surrounding structures in this region, where informative slices may provide more effective guidance.

The findings provide a reference for prompt design in future interactive segmentation model training. Considering prompt type, quantity, and slice characteristics may facilitate the development of more efficient prompting strategies for radiotherapy target delineation.

CONCLUSIONS

SAM2 showed potential for interactive esophageal CTV delineation using limited prompts. Mask prompts provided more effective guidance than bounding box prompts, and five prompts achieved a favorable balance between accuracy and interaction burden. Optimal prompt slices were preferentially selected toward the superior CTV region.

REFERENCES

1. Wu AJ, Bosch WR, Chang DT, et al. Expert Consensus Contouring Guidelines for Intensity Modulated Radiation Therapy in Esophageal and Gastroesophageal Junction Cancer. Int J Radiat Oncol Biol Phys. 2015;92(4):911-20.
2. Ravi N, Gabeur V, Hu YT, et al. SAM 2: Segment Anything in Images and Videos. arXiv preprint arXiv:2408.00714, 2024.
3. Dong H, Gu H, Chen Y, Yang J, Chen Y, Mazurowski MA. Segment Anything Model 2: An Application to 2D and 3D Medical Images. IEEE Trans Biomed Eng. 2026;PP.
4. Yamagishi Y, Hanaoka S, Kikuchi T, et al. Using Segment Anything Model 2 for Zero-Shot 3D Segmentation of Abdominal Organs in Computed Tomography Scans to Adapt Video Tracking Capabilities for 3D Medical Imaging: Algorithm Development and Validation. JMIR AI. 2025;4:e72109.