Received: 23 August 2023 | Revised: 29 December 2023 | Accepted: 14 January 2024

DOI: 10.1002/mp.16965

TECHNICAL NOTE

MEDICAL PHYSICS

Technical note: Generalizable and promptable artificial intelligence model to augment clinical delineation in radiation oncology

Lian Zhang[1] | **Zhengliang Liu**[2] | **Lu Zhang**[3] | **Zihao Wu**[2] | **Xiaowei Yu**[3] | **Jason Holmes**[1] | **Hongying Feng**[1] | **Haixing Dai**[2] | **Xiang Li**[4] | **Quanzheng Li**[4] | **William W. Wong**[1] | **Sujay A. Vora**[1] | **Dajiang Zhu**[2] | **Tianming Liu**[2] | **Wei Liu**[1]

[1]Department of Radiation Oncology, Mayo Clinic, Phoenix, Arizona, USA

[2]School of Computing, University of Georgia, Athens, Georgia, USA

[3]Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, Texas, USA

[4]Department of Radiology, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA

Correspondence
Wei Liu, Department of Radiation Oncology, Mayo Clinic Arizona, 5777 E. Mayo Boulevard, Phoenix, AZ 85054, USA.
Email: Liu.Wei@mayo.edu

Funding information
The National Cancer Institute (NCI) Career Developmental Award, Grant/Award Number: K25CA168984; Arizona Biomedical Research Commission Investigator Award; the Lawrence W. and Marilyn W. Matteson Fund for Cancer Research; the Kemper Marley Foundation

Abstract

Background: Efficient and accurate delineation of organs at risk (OARs) is a critical procedure for treatment planning and dose evaluation. Deep learning-based auto-segmentation of OARs has shown promising results and is increasingly being used in radiation therapy. However, existing deep learning-based auto-segmentation approaches face two challenges in clinical practice: generalizability and human-AI interaction. A generalizable and promptable auto-segmentation model, which segments OARs of multiple disease sites simultaneously and supports on-the-fly human-AI interaction, can significantly enhance the efficiency of radiation therapy treatment planning.

Purpose: Meta's segment anything model (SAM) was proposed as a generalizable and promptable model for next-generation natural image segmentation. We further evaluated the performance of SAM in radiotherapy segmentation.

Methods: Computed tomography (CT) images of clinical cases from four disease sites at our institute were collected: prostate, lung, gastrointestinal, and head & neck. For each case, we selected the OARs important in radiotherapy treatment planning. We then compared both the Dice coefficients and Jaccard indices derived from three distinct methods: manual delineation (ground truth), automatic segmentation using SAM's 'segment anything' mode, and automatic segmentation using SAM's 'box prompt' mode that implements manual interaction via live prompts during segmentation.

Results: Our results indicate that SAM's segment anything mode can achieve clinically acceptable segmentation results in most OARs with Dice scores higher than 0.7. SAM's box prompt mode further improves Dice scores by 0.1∼0.5. Similar results were observed for Jaccard indices. The results show that SAM performs better for prostate and lung, but worse for gastrointestinal and head & neck. When considering the size of organs and the distinctiveness of their boundaries, SAM shows better performance for large organs with distinct boundaries, such as lung and liver, and worse for smaller organs with less distinct boundaries, like parotid and cochlea.

Conclusions: Our results demonstrate SAM's robust generalizability with consistent accuracy in automatic segmentation for radiotherapy. Furthermore, the advanced box-prompt method enables the users to augment auto-segmentation interactively and dynamically, leading to patient-specific auto-segmentation in radiation therapy. SAM's generalizability across different disease sites and

different modalities makes it feasible to develop a generic auto-segmentation model in radiotherapy.

KEYWORDS
artificial intelligence, clinical delineation, generalizable, promptable, radiation oncology, segment anything model

1 | INTRODUCTION

Recent advancements in natural language processing (NLP) have led to large language models (LLMs) that can generalize to new domains with little training data.[1,2] Models such as ChatGPT,[3] GPT-4,[3] and PaLM-2 have revolutionized NLP.[4] LLMs enable artificial intelligence (AI) systems that can perform a wide range of language tasks.[5–9] Their success has inspired interest in building similar 'foundation models' for computer vision,[10,11] as evidenced by recent developments such as Meta's Segment Anything Model (SAM).[12]

SAM presents two significant advantages over previous deep learning-based auto-segmentation models.[12] First, it's a highly efficient, generalized model capable of handling diverse natural image segmentation tasks with just a single trained model, eliminating the need for task-specific models. Second, SAM introduces a prompt functionality allowing live human interactive guidance, such as clicking or boxing an area. This differs from traditional models that aim to fully automate the task without capabilities to allow for user interaction during segmentation. The users only can rectify errors after segmentation is done.[13] SAM fosters dynamic, interactive prompts during segmentation, thereby enhancing the human-computer interaction experience. Given its robust performance on natural images, SAM has potential in medical image segmentation where physician guidance is crucial.

Radiation therapy (RT) calls for precise treatment planning to minimize damage to nearby normal tissues.[14] However, the manual process of delineating organs at risk (OARs) in simulated computed tomography (CT) images undertaken by radiation oncologists or dosimetrists is laborious and time-consuming.[15,16] This bottleneck can extend patients' waiting times and potentially have a negative impact on outcomes for patients with rapidly proliferating tumors.[17] The requirement for swift and accurate delineation becomes more acute in adaptive RT, where frequent re-delineation is needed for ongoing treatment adaptation. Inaccurate delineation can result in sub-optimal treatment plans and unintended complications, underscoring the necessity for geometric and dosimetric precision.[18–22] The advent of pencil beam scanning proton therapy has increased the need for precise OAR delineation since it is more sensitive to delineation errors compared to conventional photon-based radiation therapy.[23–29]

Numerous auto-segmentation methodologies have been proposed, including deformable image registration (DIR), atlas-based auto-segmentation, and the more recent deep learning-based segmentation (DLS).[30] Although both DIR and atlas-based auto-segmentation have seen extensive implementation in clinics, their clinical utility is compromised due to limitations in accuracy and efficiency.[31,32] The focus of auto-segmentation research has recently shifted towards artificial intelligence (AI)-based methods, with a particular emphasis on those underpinned by deep learning (DL).[33,34] Over the past few years, the number of studies and clinical applications exploring DL-based segmentation of OARs in RT has proliferated, significantly improving the efficiency of auto-segmentation in RT.[35–37] However, clinical feedback suggests a prominent issue with the current models: a lack of generalizability.[38,39] These models often necessitate the training of a unique model for each RT site of one specific institution with medical images generated by a certain imaging protocol based on one specific imaging machine. Given the variety of RT sites, imaging protocols, and machines used, a multitude of independent auto-segmentation models needs to be trained even within one institution, each requiring a considerable investment of time and a large volume of site/protocol/machine-specific training data. Some data harmonization methods have been proposed to mitigate this generalizability problem with limited success.[40] In practical clinical scenarios, the determination of the site to be delineated often requires the invocation of a site/protocol/machine-specific model. The complexity inherent in such scenarios, including mixed-site images and images from rare sites, can significantly degrade the precision of existing auto-segmentation models, and in some cases, cause errors. Thus, the development of a generalized auto-segmentation model, capable of handling multiple sites simultaneously, could significantly augment the precision and efficiency of auto-segmentation in RT processes, particularly in delineating OARs.

Another common feedback from clinical usage is that existing deep learning auto-segmentation tools lack a prompt tool for user interaction allowing for dynamic and continuous adjustments during the auto-segmentation.[13] The current workflow of existing deep learning-based auto-segmentation operates as a one-time process in that the model receives medical images and then generates contours based on them

automatically. Due to the generalizability issues of current models and the complexity of clinical scenarios, such as non-standardized and rare cases, such outputs often encounter inaccuracies.[41] Clinical staffs have to rectify these errors manually slice by slice using traditional methods once auto-segmentation is completed. This process is very tedious and time-consuming. Sometimes it even costs more time than manual delineation from scratch. This is one of the major bottlenecks preventing the wide clinical adoption of auto-segmentation tools. Furthermore, delineation by clinical staff often varies from patient to patient, that is, delineation has to be done with each patient's unique conditions considered (for example, re-treatment, prior surgery/chemo, comorbidity, patient's desire, etc.). Such segmentation customization is particularly important for high-precision radiation therapy modalities such as stereotactic body radiation therapy (SBRT) or proton therapy.[42] However, as far as we know, no effective tools to integrate patient-specific delineation needs required from clinical experience into auto-segmentation models are available. The development of a prompt-supported auto-segmentation model provided by SAM is the most successful attempt in this direction so far and it can enable clinical staff to interactively, dynamically, and continuously guide the model to rectify segmentation errors and implement patient-specific delineation needs during auto-segmentation. This is important to achieve precision radiation therapy.

As an emerging next-generation segmentation model emphasizing generalizability and prompt functionality, SAM demonstrates such potential but its performance on medical images remains unclear, especially for radiotherapy.[43] Current research on SAM in the field of medical imaging primarily follows two directions: Firstly, evaluating the performance of SAM across various types of medical images[44–49]; and secondly, enhancing its performance through improvements to specific modules of SAM.[50–55] These studies, to some extent, demonstrated the generalizability of SAM. However, most of the data used originated from public datasets, leading to a lack of performance assessment in real-world clinical settings.[44–54] Research reports on the application of SAM in auto-segmentation for clinical radiation therapy are still limited. Florian et al. reported on the performance of SAM in the auto-segmentation of brain tumors in MRI images, but no further evaluation for the generalizability and human-AI interaction was reported, which is potentially the most important feature as the next generation auto-segmentation model for clinical radiation therapy.[43] To the best of our knowledge, our study is the first to comprehensively assess SAM's cross-disease site auto-segmentation performance in actual clinical radiation therapy, as well as its efficacy in human-AI interactions. We evaluate SAM's ability to perform zero-shot segmentation of medical images from multiple anatomical sites from clinical radiation oncology with CT images (prostate, lung, gastrointestinal, and head & neck). We assess SAM in "segment everything" mode where it generates masks for all objects and "box prompt" mode where users indicate regions of interest interactively during auto-segmentation. Dice coefficient and Jaccard index are used to measure the spatial overlap between SAM's predictions and ground truth clinical delineations.

2 | MATERIAL AND METHODS

2.1 | Datasets

This study has received approval from the Institutional Review Board (IRB) at our institute. Figure 1 illustrates the overall framework of our work to set up SAM for clinical radiotherapy segmentation. We collected case images from the four most common sites in clinical radiation therapy, namely the prostate, gastrointestinal, lungs, and head & neck. We gathered the CT images of 20 patients from each site, totaling 80 patients. From the perspective of clinical radiotherapy delineation, we divided the cases into two groups: the simple group includes the prostate and lungs, while the difficult group encompasses the gastrointestinal and head & neck. As SAM currently only supports 2D delineation, for a fair comparison, we evaluated the delineation of the reference 2D CT slice from every three slices along the *Z*-axis extracted from each selected organ. Based on the OARs that require attention during the formulation of the radiotherapy plan, we selected regions of interest for delineation comparison for different disease sites, following the recommendations of the radiation therapy oncology group (RTOG).

For the prostate, we selected the following regions of interest for evaluation: prostate, bladder, left femoral head, right femoral head, and rectum. For the lungs, we evaluated the following areas: left lung, right lung, heart, spinal cord, and esophagus. For the gastrointestinal, we assessed the liver, stomach, left kidney, right kidney, spinal cord, large bowel, and small bowel. For the head & neck, we evaluated the brain, left parotid, right parotid, spinal cord, mandible, left cochlea, and right cochlea.

All manual delineations were performed by highly experienced radiation oncologists and meet the RTOG delineation standards and are used in the formulation of clinical radiotherapy plans.

2.2 | SAM segment everything

SAM has three component parts as illustrated in Figure 1: an image encoder, a prompt encoder, and a mask decoder.[12] Algorithm 1 further illustrates the internal algorithm process of SAM. In Algorithm 1, the *checkpoint* stored all parameters of the SAM models

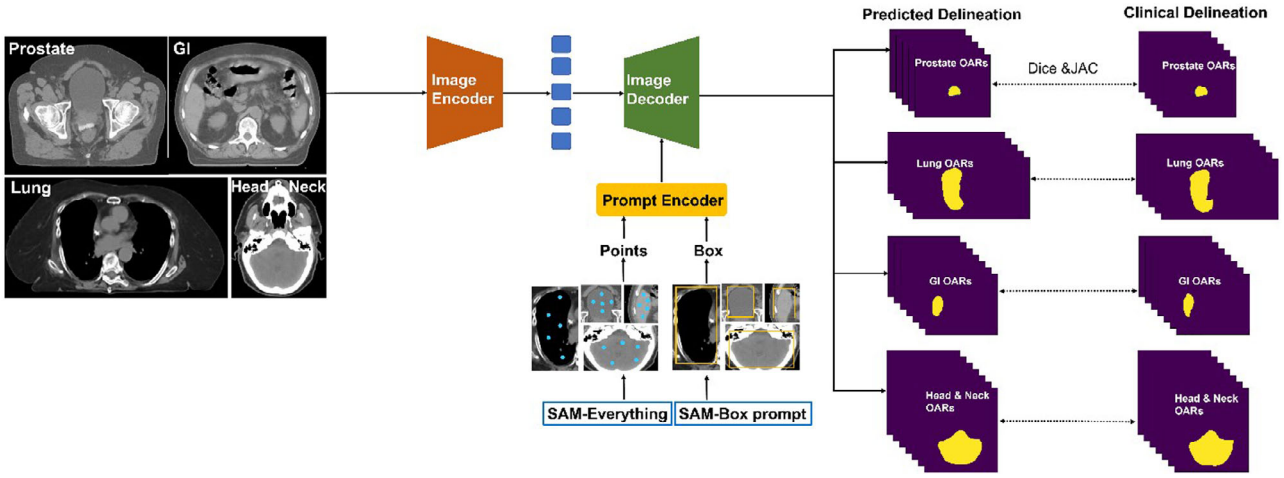

FIGURE 1 Workflow of SAM auto-segmentation in clinical radiotherapy with SAM segment everything mode and SAM box prompt mode. The prompt encoder shows the ability to take the user's clicking or boxing an area as the model's input to guide the model interactively during SAM auto-segmentation.

ALGORITHM 1 Parallel Global Convolutional Network.

Input: Checkpoint *checkpoint*, input image *input_image*, prompt *prompt*

Output: Mask *mask*

Define Function InitializeModel:

model ← SAM model built from *checkpoint*

Define Function ImagePreparation:

processed_image ← normalized and standardized version of *input_image*

Define Function InputEncoding:

image_embedding ← image encoder of *model* applied on *processed_image prompt_embedding* ← prompt encoder of *model* applied on *prompt*

Define Function MaskCreation:

mask ← mask decoder of *model* applied on *image_embedding* and *prompt_embedding*

model = InitializeModel(*checkpoint*)

processed_image = ImagePreparation(*input_image*)

image_embedding, *prompt_embedding* = InputEncoding(*model*, *processed_image*, *prompt*) *mask* = MaskCreation(*model*, *image_embedding*, *prompt_embedding*)

return *mask*

pretrained by natural images. The *model* referred to the initialized SAM model after loading these pretrained parameters. The *input_image* referred to the image inputs for segmentation, which are the CT slices from the four disease sites as shown on the left side of Figure 1. The original *input_image* will be further normalized and standardized internally, and the aligned input data will be output as *processed_image*. The image encoder as shown in Figure 1 has a backbone of Vision Transformer (ViT), which is able to convert the *processed_image* to high-level feature maps. This step is called *image_embedding* in Algorithm 1. Like the image encoder, the *prompt_embedding* in Algorithm 1 enabled by the prompt encoder takes the point or box prompt as input as shown in Figure 1. The *image decoder* will convert the high-level feature maps combined from *image_embedding* and *prompt_embedding* to the *mask* of the ROIs as described in Algorithm 1 and shown on the right side of Figure 1. In the 'segment everything' mode, SAM is designed to create segmentation masks for every possible object present within the full image, no manual priors are needed. This mode is considered the first testing approach. The commencement of this method involves producing a grid of point prompts, also known as grid sampling, which spans the entire image. To enhance the segmentation of the target regions, more random point prompts will be assigned to the target regions guiding an improved segmentation process in this study. Following that, the prompt encoder uses the sampled grid points to generate point embeddings, which are then merged with the image embeddings. The mask decoder then receives this blend as input and delivers multiple potential masks for the entire image. Afterward, a filter system is put into action to eliminate duplicate and inferior masks.

2.3 | SAM box prompt

In the prompt mode, the box prompt signifies the spatial region that necessitates segmentation, representing the object of interest. This principal mode of evaluation employs prompts that mimic a human user's interaction during auto-segmentation, crafted while closely observing the objects. Our focus lies primarily on the box prompt, tailored to encapsulate SAM's realistic use cases for creating image masks. An experienced medical physicist typically places the box prompt interactively, guided by anatomical characteristics and clinical

experience. The placement is usually in close proximity to the region of interest margin. It's vital to remember that a single "object" of interest or a "ground truth" mask might comprise multiple disconnected segments, a situation commonly encountered in medical imaging. To ensure each distinct, contiguous region of the object of interest is accurately represented, multi-box prompts are strategically placed.

We further evaluated the impact of box size on auto-segmentation results under SAM-box prompt mode. In addition to the experienced physicist, we included another less experienced junior physicist in the box prompt mode segmentation process, comparing the results with those obtained by the experienced physicist using SAM-box prompt mode for organs in four different anatomical locations. The comparison was based on the same evaluation metrics, specifically DICE and JAC results.

2.4 | Evaluation metrics

To thoroughly assess SAM's segmentation performance, we employed two commonly used metrics, as detailed below:

1. Dice Coefficient (DICE, %): This measure of similarity is used to evaluate the degree of overlap between the prediction and the ground truth (GT) as defined as Equation (1). With a range between [0, 1], a higher value denotes a more successful performance by the model.

$$DICE\ (y, \tilde{y}) = \frac{2\,|y \cap \tilde{y}|}{|y| + |\tilde{y}|} \tag{1}$$

2. Jaccard Index (JAC, %): Also recognized as the Intersection over Union (IOU), this metric, although similar to DICE, poses more stringent demands as defined as Equation (2). It quantifies the complete overlap of label ensembles across multiple test images, accommodating fractional labels through the application of fuzzy set theory. Like the DICE coefficient, higher JAC values signify superior model performance.

$$JAC\ (y, \tilde{y}) = \frac{|y \cap \tilde{y}|}{|y\tilde{y}|} \tag{2}$$

where y denotes the volume of clinical manual delineation, and $\tilde{y}$ denotes the volume of SAM auto-segmentation.

The Wilcox signed rank test is performed to measure whether the improvement of Dice coefficients and Jaccard indices between SAM's segment anything mode and SAM's box prompt mode is statistically significant or not for each organ of each site. A p-value of smaller than 0.05 is considered to be statistically significant.

3 | RESULTS

3.1 | Example cases

Figure 2 shows the auto-segmented contours from the two experiments (SAM segment everything and SAM box prompt) and the clinical delineation in the axial plane of one typical prostate (Figure 2(a)) case, one typical lung (Figure 2(b)) case, one typical gastrointestinal (Figure 2(c)) case, and one typical head & neck (Figure 2(d)) case. More example slices are available in Figure S-1.

3.2 | SAM model accuracy

For the prostate, as shown in Figure 3, the SAM segment everything mode resulted in a Dice score of 0.715 ± 0.047 and a JAC (Jaccard Index) score of 0.556 ± 0.057 when outlining the prostate. In the bladder adjacent to the prostate, the Dice score was 0.745 ± 0.039 and the JAC score was 0.595 ± 0.049. For the femoral head_L and femoral head_R, their Dice scores were both around 0.8, and JAC scores were around 0.7. The Dice score for the rectum was relatively low, at 0.629 ± 0.043, and the JAC score was 0.463 ± 0.047.

For the lungs, as shown in Figure 3, SAM segment everything mode produced Dice scores around 0.86 for both lung_L and lung_R and JAC scores around 0.76. The heart had a Dice score of

0.675 ± 0.036 and a JAC score of 0.510 ± 0.036. The spinal cord had a relatively low Dice score of 0.461 ± 0.039, and a JAC score of 0.302 ± 0.032. For the esophagus, SAM segment everything mode was unable to outline or recognize it, resulting in Dice and JAC scores of around 0.

For the gastrointestinal, as shown in Figure 4, SAM segment everything mode produced a Dice score of 0.860 ± 0.027 and a JAC score of 0.754 ± 0.033 when outlining the liver. The results for kidney_L and kidney_R were similar, with Dice scores around 0.8 and JAC scores around 0.7. The spinal cord results were consistent with those of the lung, with Dice and JAC scores of about 0.45 and 0.3 respectively. The Dice score for the stomach was relatively low, at 0.318 ± 0.031, and the JAC score was 0.190 ± 0.026. For the large bowel and small bowel, SAM segment everything mode was unable to outline or recognize them, resulting in Dice and JAC scores of around 0.

For the head & neck, as shown in Figure 4, SAM segment everything mode produced a Dice score of 0.903 ± 0.022 and a JAC score of 0.817 ± 0.036 when outlining the brain. The mandible had a Dice score of 0.871 ± 0.021, and a JAC score of 0.772 ± 0.033. The spinal cord had similar results to other sites, with Dice

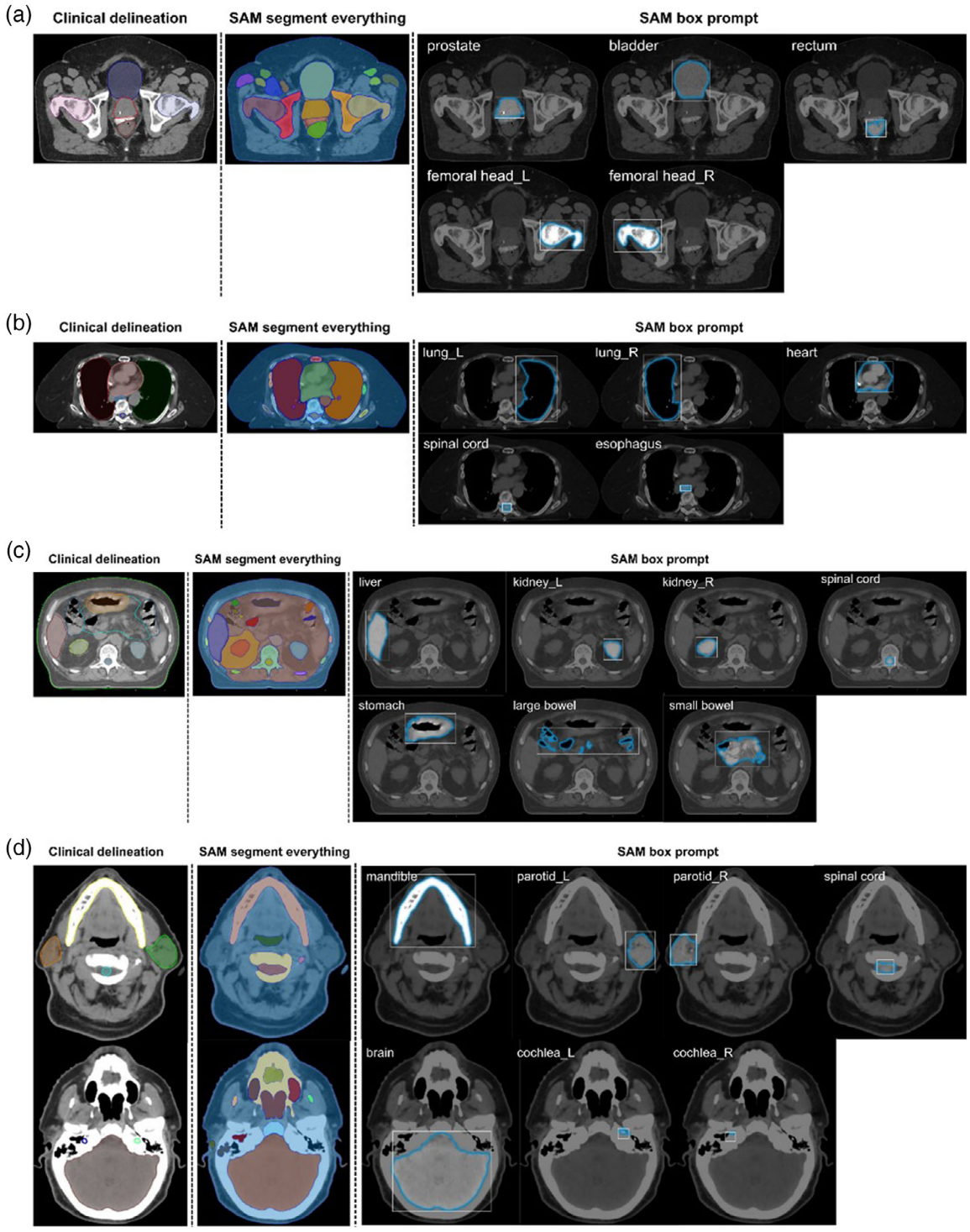

FIGURE 2 Comparison of the clinical delineation (ground truth), SAM segment everything, and SAM with box prompt of example cases. (a) Prostate, (b) lung, (c) gastrointestinal, (d) head and neck.

and JAC scores of about 0.4 and 0.3 respectively. For the parotid_L, parotid_R, cochlea_L, and cochlea_R, SAM segment everything mode was unable to outline or recognize them, resulting in Dice and JAC scores of around 0. All results were summarized in Table 1 (the third and sixth columns).

3.3 | Impact of box prompt

Overall, after introducing the box prompt, there was an improvement in Dice and JAC results for most organs. Some organs that could not be identified in SAM segment everything mode was now recognizable, although

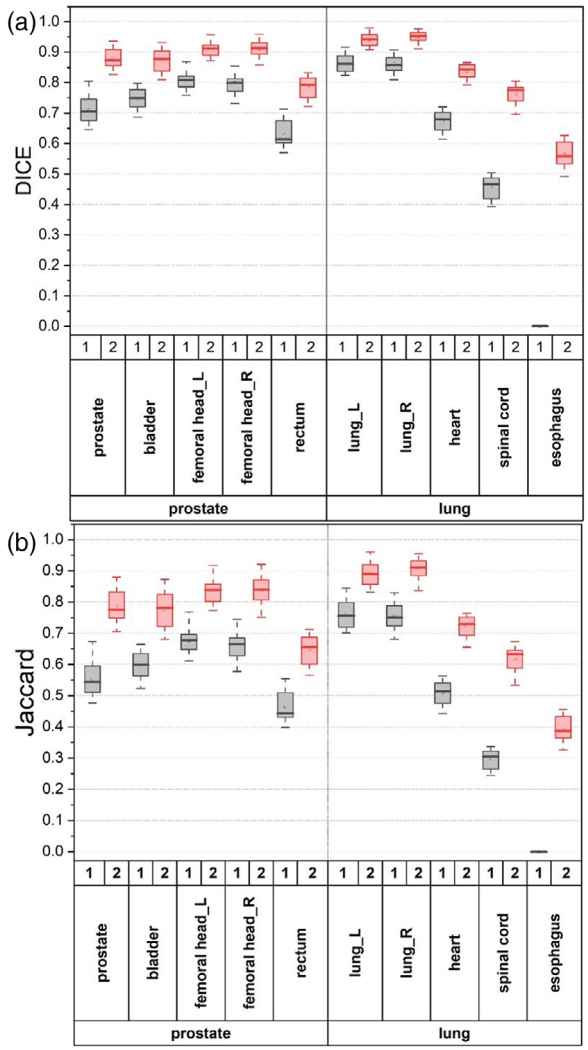

FIGURE 3 Boxplot (minimum, first quartile, median, third quartile, and maximum, respectively) of Dice coefficients and Jaccard indices of OARs between the ground truth clinical delineation and the SAM auto-segmented ones from two different experiments for the testing cases of prostate and lung. SAM's segment everything and SAM's box prompt corresponds to 1 (grey) and 2 (red) in the figure, respectively.

the resulting Dice and JAC scores were low. However, a few small organs in the head & neck remained unrecognizable even after the box prompt was employed.

For the prostate, as shown in Figure 3, the box prompt mode resulted in a Dice score of 0.878 ± 0.036 with $p = 0.0031$ and a JAC score of 0.787 ± 0.053 with $p = 0.0026$ when outlining the prostate. The bladder had a Dice score of 0.867 ± 0.045 with $p = 0.0049$, and a JAC score of 0.771 ± 0.059 with $p = 0.0057$. For femoral head_L and femoral head_R, their Dice scores each increased by around 0.1, and their JAC scores each increased by about 0.17. The rectum's Dice score increased to 0.786 ± 0.036 with $p = 0.0029$, and the JAC score to 0.651 ± 0.049 with $p = 0.0017$.

For the lungs, as depicted in Figure 3, the box prompt mode improved the Dice scores for both lung_L and lung_R by about 0.10, and the JAC scores by about 0.15. The spinal cord had a Dice score of 0.763 ± 0.033 with $p = 0.0031$, and a JAC score of 0.618 ± 0.041 with $p = 0.0021$. The heart had an improvement with Dice scores and JAC scores increasing by about 0.16 and 0.21, respectively. The esophagus was recognizable under the box prompt mode, but both Dice and JAC scores were low, at about 0.56 and 0.39, respectively.

For the gastrointestinal, as shown in Figure 4, the box prompt mode yielded a Dice score of 0.929 ± 0.015 with $p = 0.0057$ and a JAC score of 0.870 ± 0.027 with $p = 0.0063$ when outlining the liver. For kidney_L and kidney_R, the Dice and JAC scores both increased by about 0.10 and 0.13, respectively. The spinal cord saw an increase of around 0.3 for both the Dice and JAC scores. For the stomach, the Dice and JAC scores increased by about 0.26 and 0.21, respectively. The large bowel was recognizable under the box prompt mode, but both Dice and JAC scores were low, at around 0.07 and 0.04, respectively. For the small bowel, the box prompt mode recognized it, with both Dice and JAC values being around 0.15 and 0.08, respectively.

For the head & neck, as shown in Figure 4, the box prompt mode resulted in a Dice score of 0.939 ± 0.016 with $p = 0.0075$ and a JAC score of 0.883 ± 0.028 with $p = 0.0069$ when outlining the brain. The mandible had a Dice score of 0.917 ± 0.019 with $p = 0.0069$ and a JAC score of 0.842 ± 0.040 with $p = 0.0078$. The spinal cord had similar improvements as other regions, with increases of around 0.27 and 0.26 for the Dice and JAC scores, respectively. For parotid_L and parotid_R, they were recognizable under the box prompt mode, but both Dice and JAC scores were low, at around 0.57 and 0.40, respectively. For cochlea_L and cochlea_R, the box prompt mode was still unable to recognize them, with both Dice and JAC values being around 0. All results were summarized in Table 1 (the fourth and seventh columns).

As shown in Table 2, Influenced by individual experience and personal delineation styles, we observed that the box sizes used by the less experienced physicist were generally around 20% larger than those by the more experienced physicist. In organs with clear boundaries, such as the bladder and lung, the difference in box size is slightly less than 20%. However, in organs like the bowel and cochlea, the difference is greater, slightly exceeding 20%. Despite these variations in box sizes, the DICE and JAC results indicated that the final results of the delineations were closely matched, with no significant differences observed.

4 | DISCUSSION

SAM, a model pre-trained purely based on natural images, is capable of executing delineations for radiotherapy with clinically acceptable outcomes aligned

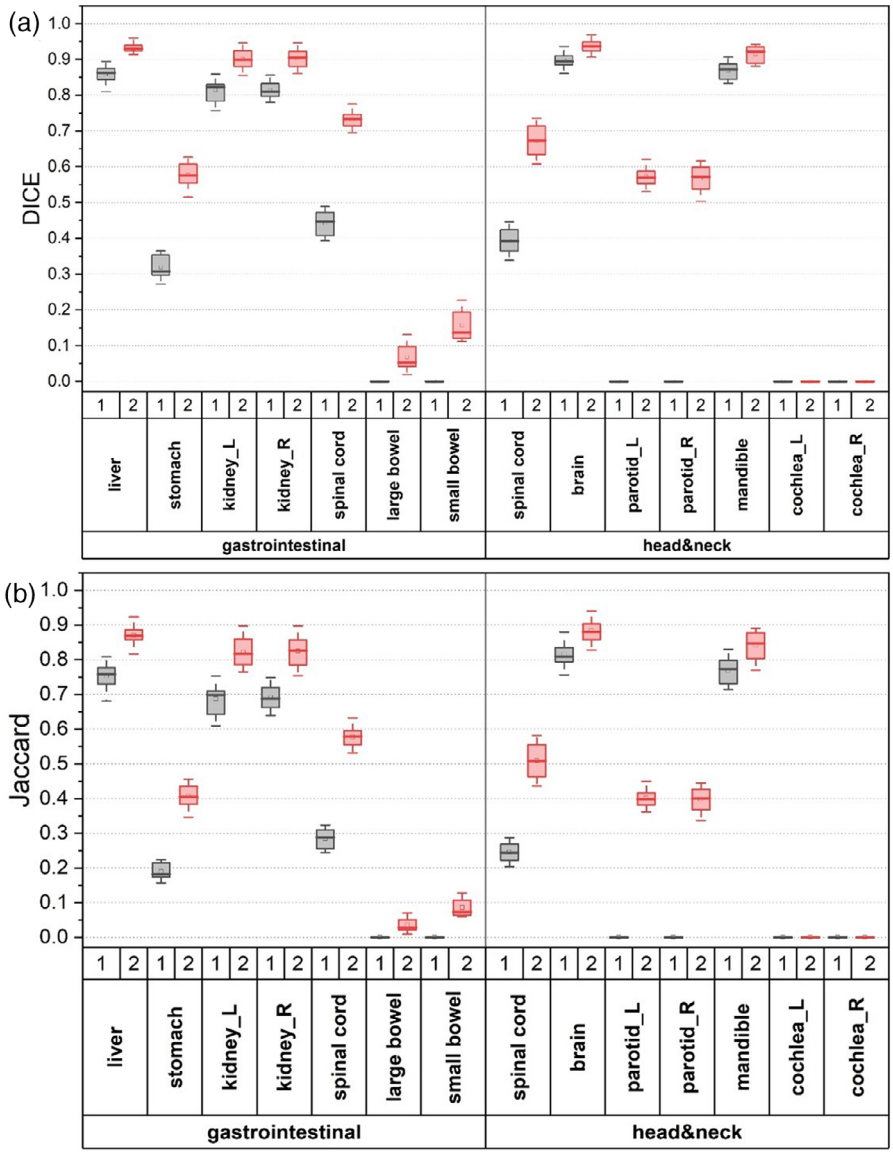

FIGURE 4 Boxplot (minimum, first quartile, median, third quartile, and maximum, respectively) of Dice coefficients and Jaccard indices of OARs between the ground truth clinical delineation and the SAM auto-segmented ones from two different experiments for the testing cases of gastrointestinal and head & neck. SAM's segment everything and SAM's box prompt corresponds to 1 (grey) and 2 (red) in the figure.

with human experience. As for the Dice and Jaccard results, under the SAM "segment everything" mode, the auto segmentation outcomes were satisfactory for the prostate's prostate, bladder, femoral head_L, and femoral head_R with (DICE: 0.7∼0.8, JAC: 0.5∼0.7), while they were less desirable for the rectum (DICE: ∼0.6, JAC: ∼0.4). For the lung, auto-segmentation for the lung_L and lung_R (DICE: 0.8∼0.9, JAC: 0.7∼0.8) were relatively better, yet less favorable for the heart (DICE: ∼0.6, JAC: ∼0.5) and spinal cord (DICE: ∼0.5, JAC: ∼0.3), with the esophagus not being recognized. The auto-segmentation outcomes for the gastrointestinal's liver, kidney_L, and kidney_R (DICE: 0.8∼0.9, JAC: 0.7∼0.8) were relatively better, whereas those for the stomach (DICE: ∼0.3, JAC: ∼0.2) were less satisfactory, with the large bowel and small bowel not recognized. For the head & neck, the brain and mandible (DICE: 0.8∼0.9, JAC: 0.7∼0.8) were better segmented, whereas the parotid_L, parotid_R, cochlea_L, and cochlea_R were not recognized. When comparing the results of segmentation across different sites, the "segment everything" mode in SAM performs better for the prostate and lung, but less satisfactory for the gastrointestinal and head & neck. If the volume and clarity of an organ are taken into account, it can be observed that the model performs better at delineating organs with distinct boundaries and larger volumes, such as the liver and brain, and less satisfactory for organs with indistinct boundaries and smaller volumes, such as the parotid and cochlea, which is in general agreement with the

TABLE 1 Dice and Jaccard scores of reference ROIs segmented by SAM in the prostate, lung, gastrointestinal (GI), and head and neck sites.

		Dice			Jaccard		
Case sites	**Reference ROIs**	**SAM-everything**	**SAM-prompt**	**p-value**	**SAM-everything**	**SAM-prompt**	**p-value**
Prostate	Prostate	0.715 ± 0.047	0.878 ± 0.036	0.0031	0.556 ± 0.057	0.787 ± 0.053	0.0026
	Bladder	0.745 ± 0.039	0.867 ± 0.045	0.0049	0.595 ± 0.049	0.771 ± 0.059	0.0057
	Femoral head_L	0.801 ± 0.033	0.912 ± 0.029	0.0027	0.673 ± 0.041	0.836 ± 0.040	0.0031
	Femoral head_R	0.803 ± 0.039	0.915 ± 0.027	0.0021	0.671 ± 0.053	0.839 ± 0.046	0.0028
	Rectum	0.629 ± 0.043	0.786 ± 0.036	0.0029	0.463 ± 0.047	0.651 ± 0.049	0.0017
Lung	Lung_L	0.867 ± 0.030	0.938 ± 0.021	0.0045	0.761 ± 0.045	0.887 ± 0.038	0.0038
	Lung_R	0.863 ± 0.028	0.945 ± 0.018	0.0039	0.759 ± 0.043	0.901 ± 0.033	0.0061
	Heart	0.675 ± 0.036	0.838 ± 0.026	0.0047	0.510 ± 0.036	0.721 ± 0.035	0.0033
	Spinal cord	0.461 ± 0.039	0.763 ± 0.033	0.0031	0.302 ± 0.032	0.618 ± 0.041	0.0021
	Esophagus	0	0.569 ± 0.038	0.0011	0	0.398 ± 0.039	0.0005
GI	Liver	0.860 ± 0.027	0.929 ± 0.015	0.0057	0.754 ± 0.033	0.870 ± 0.027	0.0063
	Kidney_L	0.809 ± 0.030	0.898 ± 0.027	0.0049	0.683 ± 0.041	0.821 ± 0.043	0.0038
	Kidney_R	0.811 ± 0.021	0.903 ± 0.026	0.0041	0.687 ± 0.029	0.827 ± 0.043	0.0047
	Spinal cord	0.445 ± 0.033	0.728 ± 0.023	0.0030	0.286 ± 0.031	0.575 ± 0.028	0.0031
	Stomach	0.318 ± 0.031	0.575 ± 0.033	0.0033	0.190 ± 0.026	0.407 ± 0.033	0.0027
	Small bowel	0	0.153 ± 0.039	0.0017	0	0.081 ± 0.029	0.0019
	Large bowel	0	0.065 ± 0.040	0.0023	0	0.035 ± 0.020	0.0021
Head & Neck	Brain	0.903 ± 0.022	0.939 ± 0.016	0.0075	0.817 ± 0.036	0.883 ± 0.028	0.0069
	Mandible	0.871 ± 0.021	0.917 ± 0.019	0.0069	0.772 ± 0.033	0.842 ± 0.040	0.0078
	Spinal cord	0.401 ± 0.033	0.673 ± 0.045	0.0039	0.249 ± 0.029	0.509 ± 0.051	0.0031
	Parotid_L	0	0.567 ± 0.021	0.0005	0	0.397 ± 0.019	0.0003
	Parotid_R	0	0.565 ± 0.029	0.0007	0	0.395 ± 0.033	0.0002
	Cochlea_L	0	0	1	0	0	1
	Cochlea_R	0	0	1	0	0	1

experiences of manual delineation. The aforementioned results demonstrate that SAM's performance in radiotherapy auto-delineation mirrors the clinical experience of human delineation, considering the variation across different sites and OARs.

Upon the inclusion of the box prompt, SAM's performance in auto-segmentation for radiotherapy showed further improvement. For most organs within the four reference sites, Dice and JAC scores have risen by 0.1–0.5, and previously unrecognized OARs such as the esophagus and parotid could be identified. The improvements were statistically significant with most p-values lower than 0.05. This suggests that the box prompt from interactive user input is effective in enhancing SAM's performance in radiotherapy segmentation, and future research could consider employing diverse prompt methods for further improvement. However, cochleas were still not well recognized. This limitation exists for all auto-segmentation algorithms based on CTs. SAM is sensitive to the clarity of the OAR boundaries in medical images; for some OARs with less distinct boundaries, multi-modality images such as MRI could be considered to assist in the segmentation of some OARs. In addition, 3D segmentation is required for clinical radiotherapy, which warrants further research.

We further examined the variations in the box size in the SAM-box prompt mode by different users and their impact on the final results. Due to differences in operator experience or delineation preferences, this study found that the box sizes drawn by the less experienced physicist were approximately 20% larger than those by the experienced physicist. However, the final delineation results as indicated by DICE and JAC results were similar, showing no significant differences. This indicated that user operations of the SAM-box prompt mode had a minimal impact on SAM's automatic segmentation results across four anatomical sites, which is crucial for ensuring consistency in clinical radiotherapy automatic segmentation under real-world clinical scenarios. It's important to note that in this study, both physicists had some clinical experience, and the boxes they designed, regardless of the box size, were clinically acceptable. In actual clinical practice, it's also essential to minimize the risk of inexperienced personnels to use human-AI interaction to guide AI towards erroneous results. This aspect warrants further investigation in future research.

TABLE 2 Dice and Jaccard scores of OARs segmented under SAM-box prompt mode with different box sizes by the experienced physicist (SAM-Prompt-S) and the less experienced physicist (SAM- Prompt-J) at the prostate, lung, gastrointestinal (GI), and head and neck sites.

Case sites	Reference ROIs	Prompt box size comparison %	Dice			Jaccard		
			SAM-Prompt-S	SAM-Prompt-J	p-value	SAM-Prompt-S	SAM-Prompt-J	p-value
Prostate	Prostate	118.1%	0.878 ± 0.036	0.880 ± 0.033	>0.05	0.787 ± 0.053	0.788 ± 0.050	>0.05
	Bladder	115.5%	0.867 ± 0.045	0.869 ± 0.041	>0.05	0.771 ± 0.059	0.771 ± 0.054	>0.05
	Femoral_L	117.7%	0.912 ± 0.029	0.908 ± 0.031	>0.05	0.836 ± 0.040	0.833 ± 0.042	>0.05
	Femoral_R	117.3%	0.915 ± 0.027	0.909 ± 0.027	>0.05	0.839 ± 0.046	0.836 ± 0.050	>0.05
	Rectum	123.1%	0.786 ± 0.036	0.789 ± 0.039	>0.05	0.651 ± 0.049	0.653 ± 0.050	>0.05
Lung	Lung_L	119.7%	0.938 ± 0.021	0.939 ± 0.021	>0.05	0.887 ± 0.038	0.887 ± 0.039	>0.05
	Lung_R	117.5%	0.945 ± 0.018	0.950 ± 0.019	>0.05	0.901 ± 0.033	0.902 ± 0.035	>0.05
	Heart	123.9%	0.838 ± 0.026	0.837 ± 0.028	>0.05	0.721 ± 0.035	0.721 ± 0.034	>0.05
	Spinal cord	117.3%	0.763 ± 0.033	0.761 ± 0.033	>0.05	0.618 ± 0.041	0.617 ± 0.040	>0.05
	Esophagus	123.5%	0.569 ± 0.038	0.567 ± 0.041	>0.05	0.398 ± 0.039	0.396 ± 0.037	>0.05
GI	Liver	115.9%	0.929 ± 0.015	0.929 ± 0.015	>0.05	0.870 ± 0.027	0.871 ± 0.025	>0.05
	Kidney_L	116.6%	0.898 ± 0.027	0.902 ± 0.027	>0.05	0.821 ± 0.043	0.823 ± 0.044	>0.05
	Kidney_R	116.2%	0.903 ± 0.026	0.898 ± 0.026	>0.05	0.827 ± 0.043	0.825 ± 0.045	>0.05
	Spinal cord	119.2%	0.728 ± 0.023	0.730 ± 0.022	>0.05	0.575 ± 0.028	0.577 ± 0.029	>0.05
	Stomach	121.9%	0.575 ± 0.033	0.573 ± 0.034	>0.05	0.407 ± 0.033	0.405 ± 0.035	>0.05
	Small bowel	125.1%	0.153 ± 0.039	0.149 ± 0.036	>0.05	0.081 ± 0.029	0.079 ± 0.027	>0.05
	Large bowel	126.7%	0.065 ± 0.040	0.061 ± 0.042	>0.05	0.035 ± 0.020	0.033 ± 0.025	>0.05
Head & Neck	Brain	117.6%	0.939 ± 0.016	0.941 ± 0.017	>0.05	0.883 ± 0.028	0.883 ± 0.026	>0.05
	Mandible	116.1%	0.917 ± 0.019	0.921 ± 0.018	>0.05	0.842 ± 0.040	0.845 ± 0.037	>0.05
	Spinal cord	117.9%	0.673 ± 0.045	0.678 ± 0.045	>0.05	0.509 ± 0.051	0.513 ± 0.047	>0.05
	Parotid_L	119.5%	0.567 ± 0.021	0.565 ± 0.020	>0.05	0.397 ± 0.019	0.396 ± 0.018	>0.05
	Parotid_R	119.1%	0.565 ± 0.029	0.563 ± 0.032	>0.05	0.395 ± 0.033	0.393 ± 0.032	>0.05
	Cochlea_L	125.7%	0	0	>0.05	0	0	>0.05
	Cochlea_R	126.9%	0	0	>0.05	0	0	>0.05

Our results indicate that the SAM model demonstrates superior generalizability consistent with manual radiotherapy delineation. Judging from the Dice scores, SAM can meet the accuracy in the majority of OARs segmentation required for clinical radiotherapy (higher than 0.7).[56,57] The Dice coefficients of some OARs from SAM are lower than 0.7. This is understandable since the SAM model was trained on daily computer vision images, rather than medical imaging data.[12] This also suggests that the SAM model should be fine-tuned in the future for radiotherapy auto-segmentation using medical imaging data. Actually, in clinical radiotherapy, there is a strong demand for generalized models for automatic segmentation. Generalized models can significantly enhance the efficiency of model deployment in a complicated clinical setting, eliminating the need to train separate models for each site. Additionally, generalized models also lead to better consistency of results, minimizing inhomogeneities caused by inherent differences between models for different sites.[41] Our results highlight SAM's strong generalizability across different imaging modalities. That is, even when trained with different modality data, SAM can perform satisfactory segmentation on the new image modalities. The data paucity problem is a long-standing problem in healthcare, which prevents the wide utilization of many deep-learning-based tools. SAM's cross-modality generalizability significantly mitigates the difficulties to have sufficient training data needed for deep-learning tools. For instance, if the data are insufficient for training a segmentation tool in magnetic resonance imaging (MRI), one could utilize the data from other modalities such as CT for model training. This is particularly meaningful for small institutions with limited data or institutions that suffer from data scarcity due to patient privacy regulations. While some research efforts have aimed to develop a generalized auto-segmentation model in recent years, to the best of our knowledge, none has achieved the robust cross-site and cross-modality generalized segmentation capabilities of SAM.[58,59] SAM's generalizability is of significant importance in clinical radiotherapy, greatly simplifying and reducing the complexity of model pre-training and clinical deployment while ensuring consistency of segmentation results

across different sites, modalities, machines, and institutions.

Leveraging the box prompt feature of SAM, results indicate that users can guide auto-segmentation by box prompt interactively during auto-segmentation, thereby improving the accuracy of most OAR segmentation across all tested sites. This improvement by box prompt is evidenced by a Dice score increase ranging from 0.1 to 0.5. For organs that were not recognized by the SAM's segment anything mode, such as esophagus and parotid, box prompts enabled SAM to recognize them well. The existing deep learning-based auto-segmentation workflow is more akin to two separate processes, where the deep learning model maximizes its auto-segmentation output at stage one, followed by tedious and time-consuming manual adjustments just like the traditional manual delineation method at stage two.[60] These two stages are separated so that user-specific contour corrections or delineation preferences based on clinical experience cannot be interactively integrated into the deep learning model during auto-segmentation. The prompt-supported SAM segmentation is thus revolutionary and can convert user interactive clicking or boxing areas into input information for the model and achieves better and patient-specific precision auto-segmentation.[12] The prompt feature enables physicians and dosimetrists to interactively, dynamically, and continuously guide auto-segmentation and rectify observed errors or generate patient-specific auto-segmentation results based on clinical experience on the fly. By enhancing clinical workflow efficiency, SAM's prompt feature enables user-preferred and patient-specific interactive AI-based auto-segmentation in radiation therapy. SAM's prompt feature, therefore, holds substantial clinical value in auto-segmentation for radiotherapy.

As a next-generation auto-segmentation platform, SAM has demonstrated superior generalization and interactive prompt capabilities in our tests for clinical radiotherapy segmentation. This potentially signifies a novel transformation for auto-segmentation in clinical radiotherapy. Despite the opportunities presented by SAM, there still exists some limitations and great potential for further improvement in its clinical application. As a new generation of general-purpose and promptable auto-segmentation model, SAM, requires more powerful computing resources. In this study, to balance the computational resources and generalizability evaluation accuracy, we selected a reference slice from every three slices for every selected organ to assess the model's auto-segmentation accuracy. Although this reference slice selection method provides a fairly accurate assessment of generalizable auto-segmentation accuracy in our study, in practical clinical applications, it's still necessary to segment every slice of the selected organ. The consumption of computing resources by large visual models like SAM will be gradually alleviated with the fast development of computing hardware, especially GPUs. In future clinical applications, the lightweighting of the current SAM models without compromising the requisite auto-segmentation accuracy remains an important research direction. Compared to the previous research on SAM in medical imaging, to the best of our knowledge, our study is the first to assess SAM's auto-segmentation capabilities in a clinical radiotherapy setting regarding its two revolutionary features: generalizability and human-AI interaction. We conducted a preliminary evaluation of its cross-disease site auto-segmentation generalizability and human-AI interaction capabilities, laying the groundwork for future research on SAM applications in clinical radiotherapy. While SAM has demonstrated strong generalizability and human-AI interaction potential, further enhancement in the context of clinical radiotherapy is still required. Some recent research on SAM offered some insights, such as using few-shot learning to further improve the network performance,[55] the addition of new adapters to the existing network structure for specific application scenarios to improve delineation precision,[50] the use of data augmentation with existing deep learning networks to enhance model performance.[51] These approaches provided avenues for further enhancing SAM's performance in clinical radiotherapy. However, given the high precision requirements and complexity of clinical radiotherapy segmentation, and especially considering the human-AI interaction capabilities supported by SAM itself, the actual performance of the SAM models still needs to be tested and evaluated in real-world clinical radiotherapy scenarios. Future work should focus on continuously fine-tuning SAM with medical images for clinical radiation therapy auto-segmentation, collaborative AI-human decision-making, and integration of diverse clinical knowledge into the SAM models to further improve SAM's performance in clinical radiation therapy auto-segmentation with balanced generalizability and accuracy. The prompt feature provided by SAM further renders segmenting and tracking moving objects possible,[61,62] which opens the door for tracking and segmenting moving OARs in real-time to account for intra- and inter-fractional anatomical changes in radiation therapy.

5 | CONCLUSIONS

In this study, we explored the clinical application of the next-generation auto-segmentation platform, SAM, for radiation therapy. It achieved clinically acceptable segmentation results with a Dice score larger than 0.7 for most OARs. SAM demonstrated superior generalization capabilities for cross-modality learning and cross-site segmentation in radiation therapy. Its revolutionary human-AI interactive prompt feature further improved the segmentation performance in radiation therapy. This

prompt feature makes user clinical experience-preferred and patient-specific auto-segmentation feasible in radiation therapy.

Our study emphasizes that foundation models like SAM should augment, not replace, human expertise, and a balanced approach recognizing both the opportunities and limitations of large foundation models is vital. We foresee a bright future where incremental progress, guided by a harmonious integration of clinical knowledge and AI, propels advancements in radiotherapy.

ACKNOWLEDGMENTS

This research was supported by the National Cancer Institute (NCI) Career Developmental Award K25CA168984, Arizona Biomedical Research Commission Investigator Award, the Lawrence W. and Marilyn W. Matteson Fund for Cancer Research, and the Kemper Marley Foundation.

CONFLICT OF INTEREST STATEMENT

The authors have no conflicts to disclose.

DATA AVAILABILITY STATEMENT

The data are available from the corresponding author upon reasonable request.

REFERENCES

1. Zhao L, Zhang L, Wu Z, et al. When brain-inspired ai meets agi. *Meta-Radiology*. 2023;1(1):100005.
2. Liu Y, Han T, Ma S, et al. Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. arXiv preprint arXiv:230401852. 2023.
3. Bubeck S, Chandrasekaran V, Eldan R, et al. Sparks of artificial general intelligence: early experiments with gpt-4. arXiv preprint arXiv:230312712. 2023.
4. Anil R, Dai AM, Firat O, et al. Palm 2 technical report. arXiv preprint arXiv:230510403. 2023.
5. Holmes J, Liu Z, Zhang L, et al. Evaluating large language models on a highly-specialized topic, radiation oncology physics. arXiv preprint arXiv:230401938. 2023.
6. Liu Z, Zhong A, Li Y, et al. Radiology-GPT: A Large Language Model for Radiology. arXiv preprint arXiv:230608666. 2023.
7. Wu Z, Zhang L, Cao C, et al. Exploring the trade-offs: Unified large language models vs local fine-tuned models for highly-specific radiology nli task. arXiv preprint arXiv:230409138. 2023.
8. Liu Z, Yu X, Zhang L, et al. Deid-gpt: Zero-shot medical text de-identification by gpt-4. arXiv preprint arXiv:230311032. 2023.
9. Dai H, Liu Z, Liao W, et al. Chataug: Leveraging chatgpt for text data augmentation. arXiv preprint arXiv:230213007. 2023.
10. Li X, Zhang L, Wu Z, et al. Artificial General Intelligence for Medical Imaging. arXiv preprint arXiv:230605480. 2023.
11. Huang Y, Yang X, Liu L, et al. Segment anything model for medical images? arXiv preprint arXiv:230414660. 2023.
12. Kirillov A, Mintun E, Ravi N, et al. Segment anything. arXiv preprint arXiv:230402643. 2023.
13. Minaee S, Boykov Y, Porikli F, Plaza A, Kehtarnavaz N, Terzopoulos D. Image segmentation using deep learning: a survey. *IEEE Trans Pattern Anal Mach Intell*. 2021;44(7):3523-3542.
14. Delaney G, Jacob S, Featherstone C, Barton M. The role of radiotherapy in cancer treatment: estimating optimal utilization from a review of evidence-based clinical guidelines. *Cancer: Interd Int J Am Cancer Soc*. 2005;104(6):1129-1137.
15. Harari PM, Song S, Tomé WA. Emphasizing conformal avoidance versus target definition for IMRT planning in head-and-neck cancer. *Int J Radiat Oncol Biol Phys*. 2010;77(3):950-958.
16. Ding Y, Feng H, Yang Y, et al. Deep-learning based fast and accurate 3D CT deformable image registration in lung cancer. *Med Phys*. 2023.
17. Chen Z, King W, Pearcey R, Kerba M, Mackillop WJ. The relationship between waiting time for radiotherapy and clinical outcomes: a systematic review of the literature. *Radiother Oncol*. 2008;87(1):3-16.
18. Liu W, Liao Z, Schild SE, et al. Impact of respiratory motion on worst-case scenario optimized intensity modulated proton therapy for lung cancers. *Pract Radiat Oncol*. 2015;5(2):e77-e86.
19. Liu W, Mohan R, Park P, et al. Dosimetric benefits of robust treatment planning for intensity modulated proton therapy for base-of-skull cancers. *Pract Radiat Oncol*. 2014;4(6):384-391.
20. Liu W, Schild SE, Chang JY, et al. Exploratory study of 4D versus 3D robust optimization in intensity modulated proton therapy for lung cancer. *Int J Radiat Oncol Biol Phys*. 2016;95(1):523-533.
21. Feng H, Sio TT, Rule WG, et al. Beam angle comparison for distal esophageal carcinoma patients treated with intensity-modulated proton therapy. *J Appl Clin Med Phys*. 2020;21(11):141-152.
22. Shan J, An Y, Bues M, Schild SE, Liu W. Robust optimization in IMPT using quadratic objective functions to account for the minimum MU constraint. *Med Phys*. 2018;45(1):460-469.
23. Schild SE, Rule WG, Ashman JB, et al. Proton beam therapy for locally advanced lung cancer: a review. *World J Clin Oncol*. 2014;5(4):568.
24. Liu C, Sio TT, Deng W, et al. Small-spot intensity-modulated proton therapy and volumetric-modulated arc therapies for patients with locally advanced non-small-cell lung cancer: a dosimetric comparative study. *J Appl Clin Med Phys*. 2018;19(6):140-148.
25. Liu W, Frank SJ, Li X, et al. Effectiveness of robust optimization in intensity-modulated proton therapy planning for head and neck cancers. *Med Phys*. 2013;40(5):051711.
26. Li H, Zhang X, Park P, et al. Robust optimization in intensity-modulated proton therapy to account for anatomy changes in lung cancer patients. *Radiother Oncol*. 2015;114(3):367-372.
27. Liu C, Yu NY, Shan J, et al. Treatment planning system (TPS) approximations matter—comparing intensity-modulated proton therapy (IMPT) plan quality and robustness between a commercial and an in-house developed TPS for nonsmall cell lung cancer (NSCLC). *Med Phys*. 2019;46(11):4755-4762.
28. Liu W, Frank SJ, Li X, Li Y, Zhu RX, Mohan R. PTV-based IMPT optimization incorporating planning risk volumes vs robust optimization. *Med Phys*. 2013;40(2):021709.
29. Liu W, Li Y, Li X, Cao W, Zhang X. Influence of robust optimization in intensity-modulated proton therapy with different dose delivery techniques. *Med Phys*. 2012;39(6Part1):3089-3101.
30. Kosmin M, Ledsam J, Romera-Paredes B, et al. Rapid advances in auto-segmentation of organs at risk and target volumes in head and neck cancer. *Radiother Oncol*. 2019;135:130-140.
31. Van Dijk LV, Van den Bosch L, Aljabar P, et al. Improving automatic delineation for head and neck organs at risk by deep learning contouring. *Radiother Oncol*. 2020;142:115-123.
32. Lee H, Lee E, Kim N, et al. Clinical evaluation of commercial atlas-based auto-segmentation in the head and neck region. *Front Oncol*. 2019;9:449587.
33. Chen X, Sun S, Bai N, et al. A deep learning-based auto-segmentation system for organs-at-risk on whole-body computed tomography images for radiation therapy. *Radiother Oncol*. 2021;160:175-184.
34. Lin H, Xiao H, Dong L, et al. Deep learning for automatic target volume segmentation in radiation therapy: a review. *Quant Imag Med Surg*. 2021;11(12):4847.

35. Ahn SH, Yeo AU, Kim KH, et al. Comparative clinical evaluation of atlas and deep-learning-based auto-segmentation of organ structures in liver cancer. *Radiat Oncol*. 2019;14:1-13.
36. Elguindi S, Zelefsky MJ, Jiang J, et al. Deep learning-based auto-segmentation of targets and organs-at-risk for magnetic resonance imaging only planning of prostate radiotherapy. *Phys Imag Radiat Oncol*. 2019;12:80-86.
37. Zhu W, Huang Y, Zeng L, et al. AnatomyNet: deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy. *Med Phys*. 2019;46(2):576-589.
38. Jarrett D, Stride E, Vallis K, Gooding MJ. Applications and limitations of machine learning in radiation oncology. *Br J Radiol*. 2019;92(1100):20190001.
39. Bashyam VM, Doshi J, Erus G, et al. Deep generative medical image harmonization for improving cross-site generalization in deep learning predictors. *J Magn Reson Imaging*. 2022;55(3):908-916.
40. Li W, Lam S, Li T, et al. Multi-institutional investigation of model generalizability for virtual contrast-enhanced mri synthesis. Paper presented at: International Conference on Medical Image Computing and Computer-Assisted Intervention 2022.
41. Zhang L, Wang X, Yang D, et al. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE Trans Med Imaging*. 2020;39(7):2531-2540.
42. Lee NY, Leeman JE, Cahlon O, et al. *Target volume delineation and treatment planning for particle therapy*. Springer; 2018.
43. Putz F, Grigo J, Weissmann T, et al. The Segment Anything foundation model achieves favorable brain tumor autosegmentation accuracy on MRI to support radiotherapy treatment planning. arXiv preprint arXiv:230407875. 2023.
44. Huang Y, Yang X, Liu L, et al. Segment anything model for medical images? *Med Image Anal*. 2023:103061.
45. Mazurowski MA, Dong H, Gu H, Yang J, Konz N, Zhang Y. Segment anything model for medical image analysis: an experimental study. *Med Image Anal*. 2023;89:102918.
46. Ma J, Wang B, Segment anything in medical images. arXiv preprint arXiv:230412306. 2023.
47. He S, Bao R, Li J, Grant PE, Ou Y, Accuracy of segment-anything model (sam) in medical image segmentation tasks. arXiv preprint arXiv:230409324. 2023.
48. Shi P, Qiu J, Abaxi SMD, Wei H, Lo FP-W, Yuan W. Generalist vision foundation models for medical imaging: a case study of segment anything model on zero-shot medical segmentation. *Diagnostics*. 2023;13(11):1947.
49. Zhang Y, Jiao R, How Segment Anything Model (SAM) Boost Medical Image Segmentation? arXiv preprint arXiv:230503678. 2023.
50. Wu J, Fu R, Fang H, et al. Medical sam adapter: Adapting segment anything model for medical image segmentation. arXiv preprint arXiv:230412620. 2023.
51. Zhang Y, Zhou T, Wang S, Liang P, Zhang Y, Chen DZ, Input augmentation with sam: boosting medical image segmentation with segmentation foundation model. Paper presented at: International Conference on Medical Image Computing and Computer-Assisted Intervention 2023.
52. Liu Y, Zhang J, She Z, Kheradmand A, Armand M, Samm (segment any medical model): A 3d slicer integration to sam. arXiv preprint arXiv:230405622. 2023.
53. Roy S, Wald T, Koehler G, et al. Sam. md: Zero-shot medical image segmentation capabilities of the segment anything model. arXiv preprint arXiv:230405396. 2023.
54. Gao Y, Xia W, Hu D, DeSAM GX, : Decoupling Segment Anything Model for Generalizable Medical Image Segmentation. arXiv preprint arXiv:230600499. 2023.
55. Zhang R, Jiang Z, Guo Z, et al. Personalize segment anything model with one shot. arXiv preprint arXiv:230503048. 2023.
56. Jameson MG, Holloway LC, Vial PJ, Vinod SK, Metcalfe PE. A review of methods of analysis in contouring studies for radiation oncology. *J Med Imag Radiat Oncol*. 2010;54(5):401-410.
57. Vinod SK, Jameson MG, Min M, Holloway LC. Uncertainties in volume delineation in radiation oncology: a systematic review and recommendations for future studies. *Radiother Oncol*. 2016;121(2):169-179.
58. Isensee F, Petersen J, Kohl SA, Jäger PF, Maier-Hein KH, nnu-net: Breaking the spell on successful medical image segmentation. arXiv preprint arXiv:190408128. 2019;1(1-8):2.
59. Isensee F, Jaeger PF, Kohl SA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. 2021;18(2):203-211.
60. Samarasinghe G, Jameson M, Vinod S, et al. Deep learning for segmentation in radiation therapy planning: a review. *J Med Imag Radiat Oncol*. 2021;65(5):578-595.
61. Dai H, Ma C, Liu Z, et al. Samaug: Point prompt augmentation for segment anything model. arXiv preprint arXiv:230701187. 2023.
62. Rajič F, Ke L, Tai Y-W, Tang C-K, Danelljan M, Yu F, Segment Anything Meets Point Tracking. arXiv preprint arXiv:230701197. 2023.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Zhang L, Liu Z, Zhang L, et al. Technical note: Generalizable and promptable artificial intelligence model to augment clinical delineation in radiation oncology. *Med Phys*. 2024;51:2187–2199. https://doi.org/10.1002/mp.16965