Unified Prompt-Visual Interactive Segmentation of Clinical Target Volume in CT for Nasopharyngeal Carcinoma with Prior Anatomical Information

Hee Guan Khor[1], Xin Yang[2], Yihua Sun[1], Jie Wang[1], Sijuan Huang[2], Shaobin Wang[1,3], Bai Lu[3], Longfei Ma[1] and Hongen Liao[1](✉)

[1] School of Biomedical Engineering, Tsinghua University, Beijing, China.
liao@tsinghua.edu.cn
[2] State Key Laboratory of Oncology in South China, Guangdong Key Laboratory of Nasopharyngeal Carcinoma Diagnosis and Therapy, Guangdong Provincial Clinical Research Center for Cancer, Sun Yat-sen University Cancer Center, Guangzhou, Guangdong, China.
[3] MedMind Technology Co., Ltd., Beijing, China.

Abstract. The delineation of the Clinical Target Volume (CTV) is a crucial step in the radiotherapy (RT) planning process for patients with nasopharyngeal carcinoma (NPC). However, manual delineation is labor-intensive, and automatic CTV contouring for NPC is difficult due to the nasopharyngeal complexity, tumor variability, and judgement-based criteria. To address the above-mentioned problems, we introduce SAM-RT, the first large vision model (LVM) designed for CTV contouring in NPC. Given the anatomical dependency required for CTV contouring—which encapsulates the Gross Tumor Volume (GTV) while minimizing exposure to Organs-at-Risk (OAR)—our approach begins with the fine-tuning of the Segment Anything Model (SAM), using a Low-Rank Adaptation (LoRA) strategy for segmenting GTV and OAR across multi-center and multi-modality datasets. This step ensures SAM-RT initially integrates with anatomical prior knowledge for CTV contouring. To optimize the use of previously acquired knowledge, we introduce Sequential LoRA (SeqLoRA) to improve knowledge retention in SAM-RT during the fine-tuning for CTV contouring. We further introduce the Prompt-Visual Cross Merging Attention (ProViCMA) for enhanced image and prompt interaction, and the Gate-Regulated Prompt Adjustment (GaRPA) strategy, utilizing learnable gates to direct prompts for effective CTV task adaptation. Efficient utilization of knowledge across relevant datasets is essential due to sparse labeling of medical images for specific tasks. To achieve this, SAM-RT is trained using an information-querying approach. SAM-RT incorporates various prior knowledge: 1) Reliance of CTV on GTV and OAR, and 2) Eliciting expert knowledge in CTV contouring. Extensive quantitative and qualitative experiments validate our designs.

H. G. Khor and X. Yang are the co-first authors.

Keywords: Clinical target volume, Data efficient learning, Interactive image segmentation, Large vision model, Nasopharyngeal carcinoma, Prior anatomical information.

1 Introduction

Nasopharyngeal carcinoma (NPC) arises in the nasopharynx, a region where the nasal passages converge with the auditory tubes in the upper throat [1]. External beam radiation therapy (RT) is a key treatment for NPC, requiring accurate delineation of clinical target volumes (CTV) for effective RT planning [2]. This involves identifying the gross tumor volumes (GTV) and organs-at-risk (OAR) to minimize radiation damage [2].

Manual segmentation in clinical practice is time-consuming and labor-intensive, with quality heavily dependent on the oncologist's experience [2–7]. Automatic delineation of CTV for NPC is challenging due to the complex anatomy of the nasopharyngeal area and unclear boundaries with adjacent OAR. Recent advancements [8–10] enable direct CTV segmentation on computed tomography (CT) scans. However, achieving precise CTV delineation necessitates considering both the visual characteristics and spatial arrangement of the GTV and OAR. Kihara *et al.* [11] proposed using CT scans and GTV label for CTV contouring to provide explicit guidance, but their approach lacks understanding of OAR positioning. Jin *et al.* [12] improved esophageal CTV contouring by incorporating GTV and OAR information with Positron Emission Tomography (PET) scans from one clinical dataset, however this approach requires multiple imaging modalities. The current literature lacks exploration in leveraging anatomical knowledge from multi-center and multi-modality datasets for the CTV contouring task.

Recently, the Segment Anything Model (SAM) [13] has emerged as a cutting-edge large-scale vision model (LVM) that allows users to create masks for specific areas of interest using interactive methods like clicking, bounding boxes, or natural language prompts. SAM’s remarkable zero-shot and few-shot learning abilities with natural images [14] have gained attention across various domains. While studies [15–17] have adapted SAM for medical image segmentation, its potential for precise CTV contouring in RT treatment remains unexplored.

In this study, we propose SAM-RT, a pioneering anatomy-aware and RT-specific LVM designed for CTV contouring in NPC. The learning of SAM-RT encompasses two critical factors: (1) **Reliance of CTV on GTV and OAR**: CTV contouring relies on the positioning of GTV and OAR. We initially employ a Low-Rank Adaptation (LoRA) strategy [18] to fine-tune SAM on segmenting GTV and OAR from multi-center and multi-modality datasets, thereby incorporating relevant anatomical knowledge into SAM-RT for CTV contouring in NPC; (2) **Eliciting expert knowledge in CTV contouring**: Human expert knowledge is crucial for RT planning to accommodate patient-specific anatomical variations. To utilize expert knowledge in CTV contouring, we propose the Prompt-Visual Cross Merging Attention (ProViCMA) to enhance interaction between CT images and user prompts, along with the Gate-Regulated Prompt Adjustment (GaRPA) strategy, which employs learnable gates to focus prompts on task-relevant blocks for effective adaptation. To facilitate knowledge retention in SAM-RT, we introduce Sequential LoRA (SeqLoRA), which updates LoRA weights for CTV contouring while preserving anatomical knowledge learned from the previous stage. Our comprehensive quantitative and qualitative evaluations affirm the effectiveness of our design.

2 Network Architecture

We aim to enhance a decision function $\mathcal{F}^C$ in target domain $\mathcal{D}_C$ for CTV contouring task $\mathcal{T}_C$ by leveraging knowledge $\mathcal{F}^S$ encapsulated in source domain $\mathcal{D}_S$ through a pre-trained SAM for natural image segmentation task $\mathcal{T}_S$ (i.e., $\mathcal{F}^S \rightarrow \mathcal{F}^C$). Considering the significant impact of GTV and OAR positioning on CTV delineation, a direct application of $\mathcal{F}^S \rightarrow \mathcal{F}^C$ without anatomical considerations appears unjustified.

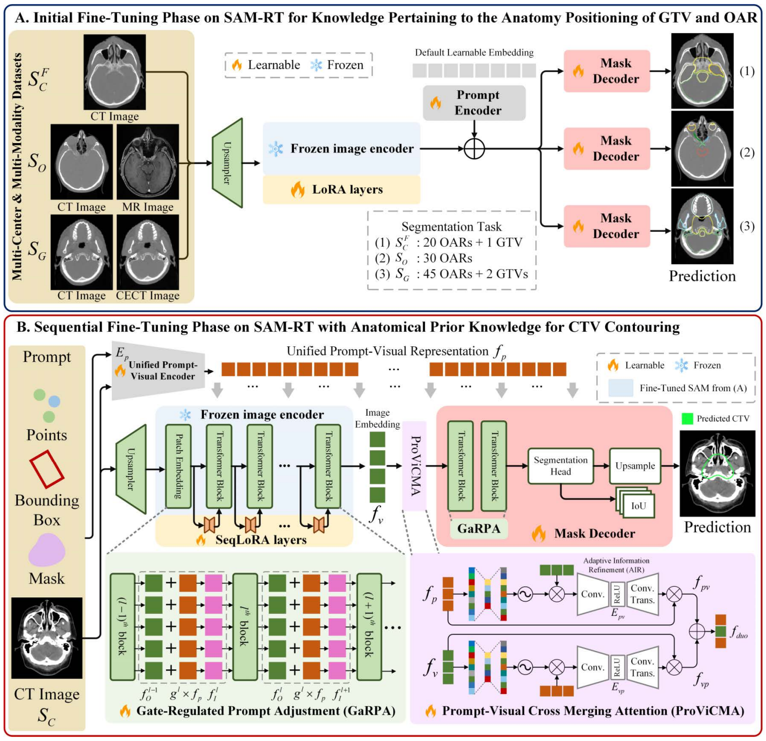

Fig. 1. Overview of the SAM-RT framework. SAM-RT learns from two key aspects: (a) Reliance of CTV on GTV and OAR: The SAM image encoder is first fine-tuned with learnable LoRA layers on the public S_G and S_O datasets and clinical S_C^F dataset, enabling the acquisition of anatomical knowledge regarding GTV and OAR; (b) Eliciting expert knowledge in CTV contouring: Sequential LoRA (SeqLoRA) is introduced to efficiently adapt transformer blocks to the CTV task using low-rank matrices, retaining anatomical knowledge learned from the initial stage. We further propose ProViCMA and GaRPA strategy to enhance interactivity between image and prompts, and to utilize learnable gates to focus prompts on task-relevant blocks effectively.

Therefore, we propose employing transfer learning to systematically integrate prior anatomical knowledge regarding the GTV and OAR from domain $\mathcal{D}_G$ into our SAM-RT framework. This involves adapting $\mathcal{F}^S$ to a newly derived decision function $\mathcal{F}^G$ specifically for GTV and OAR segmentation task $\mathcal{T}_G$. The knowledge $\mathcal{F}^G$ within $\mathcal{D}_G$ is then used to enhance the performance of decision function $\mathcal{F}^C$ (i.e., $\mathcal{F}^G \rightarrow \mathcal{F}^C$) for the domain $\mathcal{D}_C$ of CTV contouring task $\mathcal{T}_C$. Notably, the SAM-RT framework employs a sequential transfer learning process: $\mathcal{F}^S \rightarrow \mathcal{F}^G \rightarrow \mathcal{F}^C$. Through this strategy, we can ensure the effective utilization of pertinent datasets and tasks knowledge to enrich $\mathcal{F}^G$ with anatomical insights, which is then transferred to $\mathcal{F}^C$ for CTV contouring task with limited datasets.

2.1 Reliance of CTV on GTV and OAR

To integrate SAM-RT with prior knowledge of OAR and GTV, we initially fine-tune SAM-RT using a public dataset for GTV and OAR segmentation with CT and contrast-enhanced CT (CECT) (SegRap2023 [19], S_G), a public dataset for head-and-neck (HaN) OAR segmentation with CT and magnetic resonance (MR) (HaN-Seg [20], S_O), and a clinical NPC dataset for GTV and OAR segmentation S_C^F. As shown in Fig. 1, our approach employs a wide range of multi-center and multi-modality datasets (S_G, S_O, S_C^F) to improve CTV contouring task of a clinical NPC dataset (S_C).

We utilize SAM's ViT-H as our image encoder (supplementary), freezing its parameters and employing LoRA for fine-tuning. Applying LoRA to the Query (Q) and Value (V) projection layers enhance SAM-RT's performance by affecting the attention scores via the multi-head self-attention mechanism's reliance on cosine similarity (supplementary). For the prompt encoder, we initially remove all prompts in this initial stage, allowing SAM-RT to fine-tune the learnable embedding. We fine-tune all the parameters in the mask decoder (comprising a transformer decoder and a segmentation head) directly in our initial stage. Unlike SAM's ambiguity prediction, SAM-RT's decoder predicts k semantic masks $\hat{S}_l \in \mathbb{R}^{h \times w \times k}$ with $\hat{S} = \text{argmax}_{d=-1}(\text{Softmax}(\hat{S}_l)_{d=-1})$, where $\hat{S}$ is the predicted segmentation map. The Softmax and argmax operations are performed on the channel (last) dimension. We utilize bilinear upsampling to align $\hat{S}$ with the original input size.

2.2 Eliciting expert knowledge in CTV contouring

In this section, our goal is to enhance the CTV segmentation task $\mathcal{T}_C$ performance in domain $\mathcal{D}_C$ by leveraging the prior knowledge from the pre-trained SAM-RT's image encoder $\mathcal{F}^G$ for GTVs and OARs segmentation. We apply the knowledge learned in $\mathcal{F}^G$ to enhance the decision function $\mathcal{F}^C$ for CTV contouring, effectively transferring anatomical knowledge from $\mathcal{F}^G$ to $\mathcal{F}^C$ (i.e., $\mathcal{F}^G \rightarrow \mathcal{F}^C$).

Sequential LoRA. To ensure knowledge retention during the fine-tuning of SAM-RT, we have modified LoRA into Sequential LoRA (SeqLoRA). For fine-tuning SAM-RT from $\mathcal{T}_G$ to $\mathcal{T}_C$ ($\mathcal{F}^G \rightarrow \mathcal{F}^C$), we freeze ViT-H parameters as in Section 2.1 and use SeqLoRA on the Q and V layers. SeqLoRA enables the efficient adaptation of transformer blocks for the CTV task using low-rank matrices, while preserving the anatomical knowledge acquired in the initial training phase. In the initial fine-tuning stage, we fine-tuned Q and V projections as $Q' = XW^Q + XA_1^Q B_1^Q$ and $V' = XW^V + XA_1^V B_1^V$, using encoded token sequence X, original weight matrices W^Q, W^V, and low-rank matrices $A_1^Q, B_1^Q, A_1^V, B_1^V$ with ranks r_1^Q, r_1^V for $\mathcal{T}_G$. For the CTV contouring task, SeqLoRA fine-tuning $\mathcal{F}^C$ involves updating Q'' and V'' as follows: $Q'' = XW^Q + \alpha \cdot XA_1^Q B_1^Q + \beta \cdot XA_2^Q B_2^Q$ and $V'' = XW^V + \alpha \cdot XA_1^V B_1^V + \beta \cdot XA_2^V B_2^V$, with new matrices $A_2^Q, B_2^Q, A_2^V, B_2^V$ (ranks r_2^Q, r_2^V) introduced and $A_1^Q, B_1^Q, A_1^V, B_1^V$ frozen. α and β, as learnable parameters, modulate the pre-trained $\mathcal{F}^G$ influence on $\mathcal{F}^C$.

Unified Prompt-Visual Encoder. Let $I_C \in \mathbb{R}^{H \times W \times C}$ be the original CT image from S_C, $Q_n = \{p_1, \cdots, p_n\}$ where each $p_i \in \mathbb{R}^2$ $(i = 1, \dots, n\ )$ is a point with coordinates (x_i, y_i), $B = \{p_l, p_r\}$ with $p_l, p_r \in \mathbb{R}^2$ represent the top-left and bottom-right coordinates of the bounding box, and $M \in \mathbb{R}^{H \times W \times C}$ denotes the mask prompt, we introduce a prompt encoder E_p to fuse and encode various prompts into a high-dimensional unified prompt-visual representation $f_p = E_p(I_C, Q_n, B, M)$. The ablation study and the prompt generation are detailed in the supplementary. For the mask decoder, we utilize the same training strategy mentioned in Section 2.1. In our work, the SAM-RT was fine-tuned in each phase by minimizing the 2D Dice loss function [21], $L_{DICE}(\hat{S}_l, D(S))$, which aligns the SAM-RT output with the ground-truth resolution via a downsampling operation D. During the inference phase, SAM-RT utilizes these prompts in conjunction with CT images to generate predictions based on the learned parameters.

Gate-Regulated Prompt Adjustment. To direct unified prompt-visual representations, f_p towards the relevant blocks during $\mathcal{F}^G \rightarrow \mathcal{F}^C$, we propose Gate-Regulated Prompt Adjustment (GaRPA) strategy. As shown in Fig. 1, GaRPA enables SAM-RT to determine optimal weights for inserting f_p, which can be formulated as:

$$f_I^{l+1} = g^l \cdot f_p + f_O^l \tag{1}$$

where g^l represents learnable gates at each layer, adjusting the balance between f_p and previous block outputs f_O^l for the next block input f_I^{l+1}. The GaRPA strategy enhances the adaptability of g^l in both the image encoder and mask decoder, addressing the inefficiencies and weak interactions in traditional prompt methods, allowing SAM-RT to focus on CTV areas more precisely.

Prompt-Visual Cross Merging Attention. Prompt-Visual Cross Merging Attention (ProViCMA) selects features with the highest mutual response between prompt and image features, using cross-modal attention with unified prompt-visual representation f_p and image embedding f_v as inputs to obtain prompt-to-semantic feature f_{pv} and the semantic-to-prompt feature f_{vp}:

$$f_{xy} = f_x \otimes E_{xy}\big(f_y \otimes \mathrm{Sigmoid}(\mathrm{MLP}(f_x))\big), \text{ for } (x, y) \in \{(p, v), (v, p)\} \tag{2}$$

where f_{pv} and f_{vp} represent the cross-attention features derived from the features f_p to f_v and vice versa, via MLPs. Interactive weights derived from the Sigmoid function are element-wise multiplied with their opposite features to identify those with the highest response. These features are refined through an Adaptive Information Refinement (AIR) layer, E_{pv} and E_{vp}, using convolutional techniques to enhance relevant information and filter out redundancies based on prompts. The output of E_{pv} and E_{vp} are then multiplied with f_p and f_v respectively, to obtain f_{pv} and f_{vp}. The final cross-modal attention output $f_{duo} = f_{pv} + f_{vp}$ contains effective interaction features.

3 Experiment

3.1 Experimental Setup

Datasets. The clinical S_C and S_C^F datasets are collected from Sun Yat-sen University Cancer Center (Ethics Approval Number: B2022-525-01), comprising CT scans of 121 distinct patients. The data was also submitted to the public scientific research data storage platform (www.researchdata.org.cn) with an approval number RDDA2021001956. Physicians annotated the NPC CTV for S_C and both the NPC GTV and 20 OARs for S_C^F in each CT scan. To efficiently utilize prior-knowledge from relevant datasets, we collected another two public datasets: (1) SegRap2023 [19] S_G includes CT and CECT scans for 2 GTVs and 45 OARs segmentation involving 120 NPC patients; (2) HaN-Seg [20] S_O includes CT and T1-weighted MR scans for 30 OARs segmentation involving 42 patients with HaN-related cancer. We randomly split S_C into training and test datasets. The training dataset includes S_G, S_O, S_C^F, and 100 patients from S_C (denoted as S_C^{tr}), while the test dataset comprises 21 patients from S_C (denoted as S_C^{te}).

Implementation Details. 3D medical images from S_C, S_C^F, S_G, and S_O are resampled to a resolution of 512^3 with a voxel size of 1.0×1.0×3.0 mm^3. The data is standardized by extracting axial slices, retaining those with label pixel sums > 50, and applying min–max normalization, adapting SAM for its original 2D-only capability. LoRA and SeqLoRA's rank is optimized to 4 for efficiency (supplementary). The initial learning rate I_{lr} is set to 0.005, with a warmup period W_P of 250 and a maximum of 10,000 iterations [22]. For the AdamW optimizer [23], β_1, β_2, and weight decay are set to 0.9, 0.999, and 0.1, for 500 epochs. The network is implemented using PyTorch [24] and MONAI [25], running on two NVIDIA RTX A6000 GPUs with 48GB memory.

Performance Metrics. Evaluation metrics include Dice score (DSC), averaged surface distance (ASD), and Hausdorff distance (HSD) from MONAI[1], with the Wilcoxon signed-rank test with no corrections (e.g., Bonferroni) for performance comparison.

[1] Medical Open Network for Artificial Intelligence (MONAI). https://monai.io/

3.2 Ablation Study

Reliance of CTV on GTV and OAR. We first fine-tune SAM-RT directly on S_C^{tr} (Fig. 2 (a)) without any anatomical knowledge incorporation. Then, we pre-trained SAM-RT with different combinations of S_C^F, S_G, S_O and fine-tuned it on S_C^{tr}. Table 1 shows that fine-tuning SAM-RT initially on S_C^F+S_G+S_O and sequentially on S_C^{tr} yielded the highest accuracy in CTV contouring (Fig. 2 (b)). As shown in Fig. 2 (b), the attention map is more focused within the CTV area than in Fig. 2 (a). Table 1 reveals that leveraging multi-center and multi-modality prior knowledge (S_C^F, S_G, S_O) notably enhances CTV contouring, outperforming the traditional implicit knowledge guidance strategy (SAM-RT*+LoRA) that relies solely on GTV and OAR information from the same dataset S_C.

Table 1. The results confirm that the utilization of anatomical prior knowledge from multi datasets, with knowledge-retention and effective prompt interaction strategy, enhances CTV contouring performance. † indicate SAM-RT outperforms other methods with p-value < 0.05.

Methods	Datasets				DSC (%) ↑		ASD (mm) ↓	
	S_C^{tr}	S_O	S_G	S_C^F	mean ± std.	med.	mean ± std.	med.
SAM-RT*+LoRA	✓			✓	87.09 ± 16.55†	86.81	5.97 ± 3.99†	6.19
SAM-RT+LoRA	✓				86.41 ± 10.82†	85.49	5.25 ± 1.67	5.14
	✓	✓			87.15 ± 11.23†	86.95	4.15 ± 2.36	3.95
	✓		✓		87.23 ± 12.97†	87.05	4.02 ± 1.99†	3.90
	✓			✓	86.51 ± 13.04	86.97	4.91 ± 2.07†	4.35
	✓	✓	✓	✓	88.12 ± 13.56†	87.93	3.89 ± 2.15†	3.69
SAM-RT+SeqLoRA	✓	✓	✓	✓	88.75 ± 11.70	88.53	3.78 ± 2.05	3.57
SAM-RT+SeqLoRA+GaRPA	✓	✓	✓	✓	89.12 ± 10.67†	88.71	3.04 ± 1.21†	2.91
SAM-RT+SeqLoRA+GaRPA+ProViCMA	✓	✓	✓	✓	**91.05 ± 13.54**	**90.65**	**2.19 ± 1.47**	**1.85**

* Implicit GTV and OAR Guidance for CTV Contouring: Network trained to predict GTV, CTV and OAR.

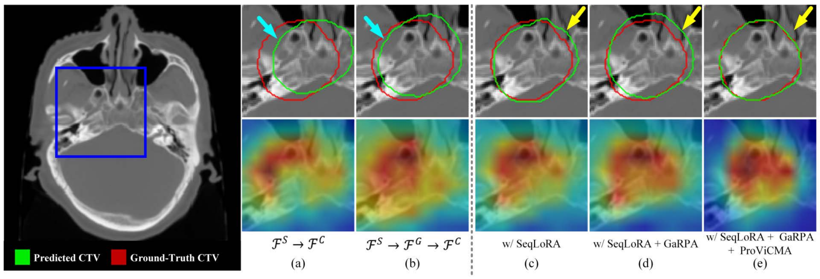
🟩 Predicted CTV
🟥 Ground-Truth CTV
$\mathcal{F}^S \rightarrow \mathcal{F}^C$
(a)
$\mathcal{F}^S \rightarrow \mathcal{F}^G \rightarrow \mathcal{F}^C$
(b)
w/ SeqLoRA
(c)
w/ SeqLoRA + GaRPA
(d)
w/ SeqLoRA + GaRPA + ProViCMA
(e)

Fig. 2. Comparing attention maps from different ablation networks, where red indicates high attention, blue indicates low attention. Blue arrows show improved CTV delineation with prior multi-center, multi-modality knowledge. Yellow arrows signify enhanced CTV delineation through expert knowledge elicitation with deep prompt interactions.

Eliciting expert knowledge in CTV contouring. Table 1 shows that our proposed SAM-RT (Fig. 2 (e)), comprising SeqLoRA, GaRPA, and ProViCMA, demonstrates best performance. As shown in Fig. 2, our SAM-RT framework (Fig. 2 (e)), through knowledge-retention and effective prompt interaction strategy, enhances CTV attention while reducing focus on OAR regions, mirroring clinical efforts to minimize radiation toxicity. All ablation networks use the same prompts (Q_{10}, B, M), isolating performance differences to learned knowledge, not variations in prompts.

3.3 Comparison to the State-of-the-Art (SOTA) Approaches

Compared with other methods, the CTV contour predicted by SAM-RT are the most accurate (as shown in Fig. 3). Table 2 shows that SAM-RT achieves the best DSC/ASD/HSD of 91.05%/2.19mm/15.57mm. Different from the traditional learning strategy that treats CTV contouring as an isolated task, SAM-RT bridges the gap of $\mathcal{F}^S \to \mathcal{F}^G \to \mathcal{F}^C$ under a unified framework, surpassing foundational models (i.e., SAM [13], SAMed [15], SAM-Med2D [16], SAM-Med3D [17]), as well as conventional methods guided by explicit and implicit GTV (i.e., UNet_{GTV} [26], DDNN [27]). The phenomenon is due to the 2D SAM model's superior feature extraction capability, pre-trained on the SA-1B dataset [13], and SAM-RT's efficient prompt interaction with strategies for acquiring anatomical-prior knowledge, enhancing SAM customization. The training details of SOTA models can refer to supplementary. The performance of SAM-RT in Tables 1 and 2 was not uniformly statistically significant, attributable to inconsistencies in data acquisition and variations in CTV delineation by different oncologists.

Table 2. Quantitative comparison of CTV contouring performance. Our proposed SAM-RT achieved the best performance, where † indicate SAM-RT outperforms other methods with *p*-value < 0.05.

Methods	DSC (%) ↑		ASD (mm) ↓		HSD (mm) ↓	
	mean ± std.	med.	mean ± std.	med.	mean ± std.	med.
nnUNet [28]	$75.63 \pm 11.45^{\dagger}$	76.98	$7.12 \pm 3.05^{\dagger}$	7.10	$32.77 \pm 17.41^{\dagger}$	31.05
UNETR [29]	$78.63 \pm 12.05^{\dagger}$	77.94	$6.05 \pm 3.64^{\dagger}$	5.91	27.11 ± 11.89	26.16
UNet_{GTV}** [26]	$79.98 \pm 13.12^{\dagger}$	78.79	5.62 ± 2.74	5.49	$26.82 \pm 15.65^{\dagger}$	24.87
DDNN*** [27]	80.21 ± 11.28	79.15	$5.59 \pm 3.55^{\dagger}$	5.68	$21.77 \pm 12.19^{\dagger}$	19.64
SI-Net [30]	$82.49 \pm 15.97^{\dagger}$	81.70	$5.06 \pm 2.41^{\dagger}$	5.20	$20.14 \pm 13.04^{\dagger}$	18.59
SAM [13]	$61.59 \pm 20.58^{\dagger}$	59.66	$9.12 \pm 5.84^{\dagger}$	9.55	58.64 ± 20.18	54.20
SAMed [15]	86.59 ± 10.05	85.97	3.02 ± 1.89	2.91	16.33 ± 11.25	15.11
SAM-Med2D [16]	$85.77 \pm 11.32^{\dagger}$	85.02	4.69 ± 2.31	4.59	$18.51 \pm 12.65^{\dagger}$	18.02
SAM-Med3D [17]	$87.59 \pm 12.61^{\dagger}$	87.05	$4.35 \pm 2.11^{\dagger}$	4.23	$16.24 \pm 15.87^{\dagger}$	16.55
SAM-RT	**91.05 ± 13.54**	**90.65**	**2.19 ± 1.47**	**1.85**	**15.57 ± 10.38**	**14.90**

** Explicit GTV Guidance for CTV Contouring: GTV mask combined with CT images as input.

*** Implicit GTV Guidance for CTV Contouring: Network trained to predict both GTV and CTV.

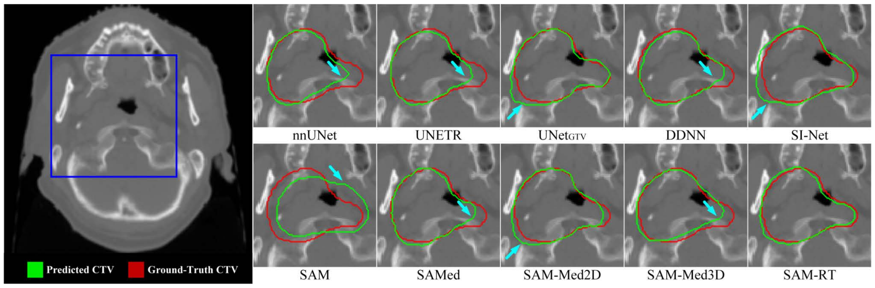

Fig. 3. Qualitative analysis of various CTV contouring methods. Limited anatomical understanding in networks can degrade CTV delineation (blue arrows). Our SAM-RT, encompassing anatomical prior knowledge with enhanced prompt interactions, demonstrates superior performance.

4 Conclusion

In this paper, we propose SAM-RT, the first LVM designed for CTV contouring. To improve the accuracy for CTV contouring, we incorporated the naturally existing prior knowledge of GTV and OAR across multi-center and multi-modality datasets. For efficient CTV task adaptation, we fine-tune SAM-RT with SeqLoRA in a knowledge-retention manner. We further proposed ProViCMA and GaRPA for dense prompt interaction and efficient CTV task adaptation. SAM-RT employs an information-querying method for its training, integrating key prior knowledge including the dependency of CTV on GTV and OAR, as well as leveraging expert insights for accurate CTV delineation. Our future research includes enhancing 3D SAM-RT development to capture structural details and integrating medical domain-specific language understanding.

Acknowledgments. This study was funded by National Key Research and Development Program of China (2022YFC2405200), National Natural Science Foundation of China (82027807, U22A2051), Institute for Intelligent Healthcare, Tsinghua University (2022ZLB001), Tsinghua-Foshan Innovation Special Fund (2021THFS0104), Guangdong Esophageal Cancer Institute Science and Technology Program (Q202221, Q202311, Q202223, Q202214).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Mohammed, M.A., Abd Ghani, M.K., Hamed, R.I., Ibrahim, D.A.: Review on Nasopharyngeal Carcinoma: Concepts, methods of analysis, segmentation, classification, prediction and impact: A review of the research literature. J Comput Sci. 21, 283–298 (2017).
2. Lee, A.W., Ng, W.T., Pan, J.J., Poh, S.S., Ahn, Y.C., AlHussain, H., Corry, J., Grau, C., Grégoire, V., Harrington, K.J.: International guideline for the delineation of the clinical target volumes (CTV) for nasopharyngeal carcinoma. Radiotherapy and Oncology. 126, 25–36 (2018).
3. Tong, Y., Gong, G., Chen, J., Lu, J., Liu, T., Cheng, P., Yin, Y.: The heterogeneous CTV-PTV margins should be given for different parts of tumors during tomotherapy. Oncotarget. 8, 89086–89094 (2017). https://doi.org/10.18632/oncotarget.21631.
4. Jeanneret-Sozzi, W., Moeckli, R., Valley, J.F., Zouhair, A., Ozsahin, E.M., Mirimanoff, R.O.: The reasons for discrepancies in target volume delineation: A SASRO study on head-and-neck and prostate cancers. Strahlentherapie und Onkologie. 182, (2006). https://doi.org/10.1007/s00066-006-1463-6.
5. Feng, M.U., Demiroz, C., Vineberg, K.A., Balter, J.M., Eisbruch, A.: Intra-observer variability of organs at risk for head and neck cancer: geometric and dosimetric consequences. Int J Radiat Oncol Biol Phys. 78, S444–S445 (2010).
6. Kim, Y., Patwardhan, K.A., Beichel, R.R., Smith, B.J., Ulrich, E.J., Mart, C., Chang, T., Plichta, K.A., Buatti, J.: Impact of contouring accuracy on expected tumor control probability for head and neck cancer: semiautomated segmentation versus manual contouring. Int J Radiat Oncol Biol Phys. 96, E545 (2016).

7. Sun, Y., Khor, H.G., Huang, S., Chen, Q., Wang, S., Yang, X., Liao, H.: Second-Course Esophageal Gross Tumor Volume Segmentation in CT with Prior Anatomical and Radiotherapy Information. In: International Conference on Medical Image Compu-ting and Computer-Assisted Intervention. pp. 511–520. Springer (2023).
8. Xue, X., Qin, N., Hao, X., Shi, J., Wu, A., An, H., Zhang, H., Wu, A., Yang, Y.: Sequential and iterative auto-segmentation of high-risk clinical target volume for radiotherapy of nasopharyngeal carcinoma in planning CT images. Front Oncol. 10, 1134 (2020).
9. Men, K., Chen, X., Zhang, Y., Zhang, T., Dai, J., Yi, J.: Deep deconvolutional neural network for target segmentation of nasopharyngeal cancer in planning computed tomography images. Front Oncol 2017; 7: 315, (2017).
10. Men, K., Zhang, T., Chen, X., Chen, B., Tang, Y., Wang, S., Li, Y., Dai, J.: Fully automatic and robust segmentation of the clinical target volume for radiotherapy of breast cancer using big data and deep learning. Physica Medica. 50, 13–19 (2018).
11. Kihara, S., Koike, Y., Takegawa, H., Anetai, Y., Nakamura, S., Tanigawa, N., Koizumi, M.: Clinical target volume segmentation based on gross tumor volume using deep learning for head and neck cancer treatment. Medical Dosimetry. 48, 20–24 (2023).
12. Jin, D., Guo, D., Ho, T.-Y., Harrison, A.P., Xiao, J., Tseng, C.-K., Lu, L.: DeepTarget: Gross tumor and clinical target volume segmentation in esophageal cancer radiotherapy. Med Image Anal. 68, 101909 (2021).
13. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y.: Segment anything. arXiv preprint arXiv:2304.02643. (2023).
14. Tang, L., Xiao, H., Li, B.: Can sam segment anything? when sam meets camouflaged object detection. arXiv preprint arXiv:2304.04709. (2023).
15. Zhang, K., Liu, D.: Customized segment anything model for medical image segmentation. arXiv preprint arXiv:2304.13785. (2023).
16. Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L.: Sam-med2d. arXiv preprint arXiv:2308.16184. (2023).
17. Wang, H., Guo, S., Ye, J., Deng, Z., Cheng, J., Li, T., Chen, J., Su, Y., Huang, Z., Shen, Y.: Sam-med3d. arXiv preprint arXiv:2310.15161. (2023).
18. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685. (2021).
19. Luo, X., Fu, J., Zhong, Y., Liu, S., Han, B., Astaraki, M., Bendazzoli, S., Toma-Dasu, I., Ye, Y., Chen, Z.: SegRap2023: A Benchmark of Organs-at-Risk and Gross Tumor Volume Segmentation for Radiotherapy Planning of Nasopharyngeal Carcinoma. arXiv preprint arXiv:2312.09576. (2023).
20. Podobnik, G., Strojan, P., Peterlin, P., Ibragimov, B., Vrtovec, T.: HaN-Seg: The head and neck organ-at-risk CT and MR segmentation dataset. Med Phys. 50, 1917–1927 (2023).
21. Taha, A.A., Hanbury, A.: Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. BMC Med Imaging. 15, (2015). https://doi.org/10.1186/s12880-015-0068-x.
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016).
23. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101. (2017).
24. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L.: Pytorch: An imperative style, high-performance deep learning library. Adv Neural Inf Process Syst. 32, (2019).

25. Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D.: Monai: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701. (2022).
26. Kihara, S., Koike, Y., Takegawa, H., Anetai, Y., Nakamura, S., Tanigawa, N., Koizumi, M.: Clinical target volume segmentation based on gross tumor volume using deep learning for head and neck cancer treatment. Medical Dosimetry. 48, 20–24 (2023).
27. Men, K., Chen, X., Zhang, Y., Zhang, T., Dai, J., Yi, J., Li, Y.: Deep deconvolutional neural network for target segmentation of nasopharyngeal cancer in planning computed tomography images. Front Oncol. 7, 315 (2017).
28. Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P.F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S.: nnu-net: Self-adapting framework for u-net-based medical image segmentation. arXiv preprint arXiv:1809.10486. (2018).
29. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: Unetr: Transformers for 3d medical image segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 574–584 (2022).
30. Xue, X., Qin, N., Hao, X., Shi, J., Wu, A., An, H., Zhang, H., Wu, A., Yang, Y.: Sequential and iterative auto-segmentation of high-risk clinical target volume for radiotherapy of nasopharyngeal carcinoma in planning CT images. Front Oncol. 10, 1134 (2020).