ViPSAM: Visual Prompting Medical Image Segmentation Using Segment Anything Model

San Lee[1], Nalee Kim[2], Jeong Il Yu[2], Hee Chul Park[2], and Boah Kim[3]*

[1] Department of Artificial Intelligence, Sungkyunkwan University, Republic of Korea
[2] Department of Radiation Oncology, Samsung Medical Center, Sungkyunkwan University School of Medicine, Republic of Korea
[3] Department of MetaBioHealth, Sungkyunkwan University, Republic of Korea

Abstract. In proton therapy planning, respiratory-gated non-contrast CT (NCCT) is commonly used for lesion segmentation; however, accurate delineation remains challenging due to low lesion-to-background contrast. Although learning-based methods have shown strong performance, they often struggle with non-contrast image segmentation. Inspired by clinical practice, where contrast-enhanced MRI is referenced to delineate lesions on NCCT, we propose ViPSAM, a visual prompting framework that leverages complementary cross-modality information. Built upon the Segment Anything Model (SAM), ViPSAM introduces a visual prompt encoder to extract guidance features from contrast-enhanced images and a visual-guided cross-attention module to integrate non-contrast and contrast-enhanced features, thereby enhancing lesion-relevant representations in low-contrast regions. The mask decoder is further adapted in a parameter-efficient manner to utilize visual prompts effectively. We evaluate the proposed method on liver lesion segmentation using NCCT acquired for proton therapy. Experimental results demonstrate that ViPSAM outperforms representative U-Net- and SAM-based methods, indicating that cross-modality visual prompting enables more robust and accurate segmentation in non-contrast images.

Keywords: Medical image segmentation · Segment anything model · Visual prompt · Multi-modality

1 Introduction

Accurate lesion delineation is important for proton therapy planning of liver tumors, as it determines target localization and proton beam delivery [5,14]. In clinical workflows, respiratory-gated non-contrast CT (NCCT) serves as the reference image for treatment planning [6]. However, precise segmentation on NCCT is challenging due to respiratory-induced anatomical variations and the inherently low contrast between lesions and surrounding liver regions [9,18,3,19]. To address this, clinicians often refer to contrast-enhanced MRI acquired in a

* Corresponding author. Email: boah.kim at skku.edu

specific respiratory phase to better identify the lesion boundaries while contouring on NCCT. Nevertheless, the manual process is time-consuming and prone to inter-observer variability [16,13]. These challenges highlight the need for robust automated segmentation methods for low-contrast treatment planning images.

With advances in deep learning, medical image segmentation has been extensively studied using U-Net-based architectures [15] and their extensions [10,4,2]. Although these models have achieved outstanding performance on various tasks, they typically operate in a single-modality setting and rely solely on intensity-based appearance cues. Accordingly, in non-contrast medical images, the low lesion-to-background contrast reduces feature discriminability, making it difficult for models to segment lesion boundaries with high performance.

Recently, foundation models trained on large-scale datasets have demonstrated strong transferability across downstream tasks, including image segmentation [11,21,17,7]. In particular, the Segment Anything Model (SAM) [11] presents a prompt-based segmentation framework and has shown remarkable generalization in medical imaging applications [12,20]. However, when applied to low-contrast images such as NCCT, SAM-based models still struggle to delineate small lesions with indistinct boundaries. Moreover, adapting foundation models to specific medical domains often requires substantial computational cost.

To address these challenges, we propose a visual prompting medical image segmentation model based on SAM, termed *ViPSAM*, which leverages complementary multi-modal information within a unified framework. Inspired by the clinical practice where radiation oncologists reference contrast-enhanced MRI to delineate lesions on NCCT, our model incorporates cross-modality visual guidance to enhance segmentation in low-contrast clinical scenarios.

Specifically, we introduce a visual prompt encoder to extract soft-tissue contrast cues from contrast-enhanced images, which serve as visual prompts and provide complementary information to non-contrast medical images. These cues are integrated with non-contrast image representations through a visual-guided cross-attention module, enabling the model to better capture lesion-relevant information. Furthermore, we adapt the SAM mask decoder by incorporating Low-Rank Adaptation (LoRA) [8] for parameter-efficient fine-tuning. Through this design, segmentation is performed on non-contrast images, while contrast-enhanced images are used solely as visual prompts.

We evaluate ViPSAM for liver lesion segmentation using a proton therapy dataset collected at a tertiary medical center, consisting of NCCT and corresponding contrast-enhanced MRI. Segmentation is performed on NCCT using MRI-derived visual prompts. Experimental results demonstrate that visual prompting significantly improves segmentation performance, surpassing representative U-Net- and SAM-based methods. Our contributions are as follows:

- We propose ViPSAM, a SAM-based visual prompting framework leveraging cross-modal cues for lesion segmentation in non-contrast medical images.
- We design a visual prompt encoder and a visual-guided cross-attention module that conditions non-contrast features on contrast-enhanced visual prompts, together with LoRA-based parameter-efficient adaptation within SAM.

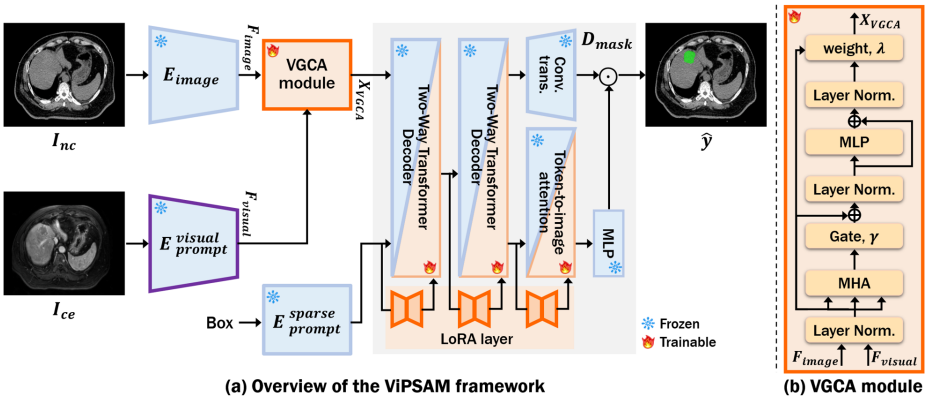

Fig. 1. (a) Overview of ViPSAM. Lesion segmentation $\hat{y}$ on non-contrast images I_{nc} is predicted by leveraging complementary information from contrast-enhanced images I_{ce} via a visual prompt encoder, a visual-guided cross-attention (VGCA) module, and LoRA-based adaptation in the mask decoder. (b) Architecture of the VGCA module.

– Experimental results on a liver lesion segmentation dataset for proton therapy with patient-matched NCCT and contrast-enhanced MRI demonstrate superior performance over representative U-Net- and SAM-based methods.

2 Proposed Method

An overview of ViPSAM, our SAM-based framework for visual-prompting image segmentation, is shown in Fig. 1. ViPSAM extends the Segment Anything Model (SAM) by incorporating (i) a visual prompt encoder that extracts soft-tissue contrast cues from cross-modality contrast-enhanced images, (ii) a visual-guided cross-attention module that conditions non-contrast image features on the extracted visual prompt representations, and (iii) Low-Rank Adaptation (LoRA) within the decoder for parameter-efficient fine-tuning. We first provide a brief overview of SAM, followed by detailed descriptions of each component.

Segment Anything Model (SAM) SAM [11] consists of an image encoder E_{image}, a prompt encoder E_{prompt}, and a mask decoder D_{mask}. The image encoder, implemented as a Vision Transformer (ViT), takes an input image I and extracts image embedding $F_{image} = E_{image}(I)$. The prompt encoder embeds a prompt P (e.g., points, boxes, or masks) into prompt tokens $E_{prompt}(P)$. The mask decoder then takes the image embedding F_{image} and the prompt tokens $E_{prompt}(P)$ to predict a segmentation mask $\hat{y} = D_{mask}(F_{image}, E_{prompt}(P))$ via Two-Way Transformer blocks that perform bidirectional cross-attention between the prompt tokens and image embedding. In our framework, we adopt SAM with a ViT-B image encoder and keep its pretrained parameters frozen during training, using box prompts as input.

2.1 Visual Prompt Encoder

To mitigate low lesion-to-background contrast in non-contrast images, we introduce a visual prompt encoder E_{prompt}^{visual}, where *visual prompt* refers to image-based guidance from contrast-enhanced images that provide clearer lesion contrast, in contrast to sparse prompts (e.g., boxes). Notably, E_{prompt}^{visual} uses the same ViT-B architecture and SAM-pretrained weights as the image encoder E_{image}, with all weights frozen during training.

Given a non-contrast image I_{nc} and a contrast-enhanced image I_{ce}, each image is replicated to three channels to match the SAM input format, yielding $I_{nc}, I_{ce} \in \mathbb{R}^{3\times H\times W}$, where H and W represent the height and width of the input image, respectively. When the non-contrast image is fed into E_{image} to extract:

$$F_{image} = E_{image}(I_{nc}) \in \mathbb{R}^{c\times h\times w}, \tag{1}$$

where c and $h \times w$ denote the channel dimension and spatial resolution of the feature map, respectively, the contrast-enhanced image is encoded by E_{prompt}^{visual} to obtain visual prompt features:

$$F_{visual} = E_{prompt}^{visual}(I_{ce}) \in \mathbb{R}^{c\times h\times w}. \tag{2}$$

Since both encoders have an identical architecture, the resulting feature maps can interact directly in the subsequent visual-guided cross-attention module.

2.2 Visual-Guided Cross-Attention Module

To obtain visual-prompt-guided representations with the non-contrast image serving as the spatial reference, we design a visual-guided cross-attention (VGCA) module (Fig. 1(b)) that conditions non-contrast image features F_{image} on visual prompt features F_{visual}:

$$X_{VGCA} = \mathrm{VGCA}(F_{image}, F_{visual}). \tag{3}$$

Specifically, in the VGCA module, F_{image} and F_{visual} are first flattened and normalized by layer normalization (LN) [1]:

$$X_{image} = \mathrm{LN}(\mathrm{Flat}(F_{image})), \quad X_{visual} = \mathrm{LN}(\mathrm{Flat}(F_{visual})), \tag{4}$$

where $\mathrm{Flat}(\cdot)$ reshapes $\mathbb{R}^{c\times h\times w}$ to $\mathbb{R}^{n\times c}$ with $n = h \cdot w$. Then, multi-head cross-attention (MHA) is computed as:

$$A = \mathrm{MHA}(Q = X_{image}, K = X_{visual}, V = X_{visual}), \tag{5}$$

where Q, K, and V denote the query, key, and value, respectively. Since segmentation is performed on non-contrast images, X_{image} is used as the query, while X_{visual} serves as the key and value to provide complementary lesion information from contrast-enhanced images.

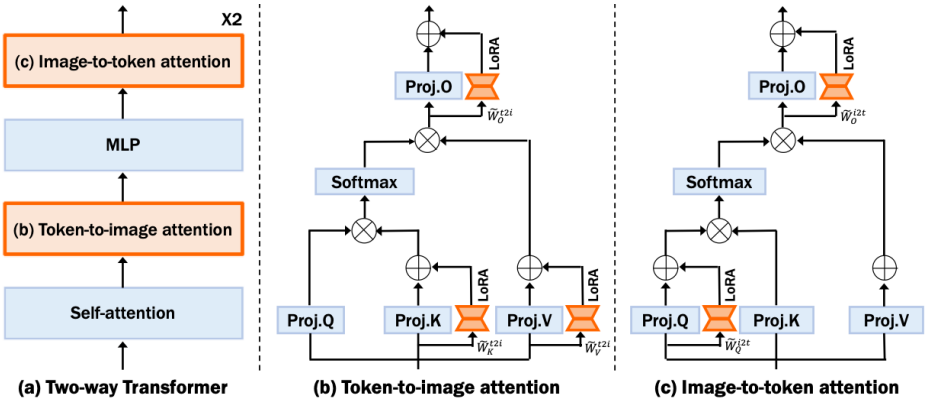

Fig. 2. LoRA placement in the SAM mask decoder. (a) Two-Way Transformer block in the mask decoder. (b,c) LoRA is applied to selected projections in cross-attention layers: token-to-image (K, V, O) and image-to-token (Q, O).

To regulate the influence of visual prompts, a learnable gating scalar γ is introduced through a gated residual update:

$$X_\gamma = \mathrm{LN}(X_{image} + \gamma A), \tag{6}$$

$$\tilde{X} = \mathrm{LN}(X_\gamma + \mathrm{MLP}(X_\gamma)), \tag{7}$$

where MLP denotes a feed-forward network with a Linear–GeLU–Linear structure, and $\tilde{X}$ is the cross-attended representation. The visual-prompt-conditioned image embedding X_{VGCA}, used in the mask decoder, is then computed as:

$$X_{VGCA} = \lambda X_{image} + (1 - \lambda)\tilde{X}, \tag{8}$$

where λ is a learnable weight scalar. Unlike standard cross-attention, our formulation with learnable gating and weighting preserves the non-contrast image as the primary spatial reference while enabling cross-modal visual guidance to modulate feature representations.

2.3 LoRA in Mask Decoder

In our framework, the mask decoder D_{mask} consists of stacked Two-Way Transformer blocks that alternate token-to-image and image-to-token cross-attention (Fig. 2(a)). This takes sparse prompt tokens generated by the sparse prompt encoder E^{sparse}_{prompt} from a box prompt P which provides localization cues, together with image embeddings produced by the VGCA module, and predicts the segmentation mask $\hat{y}$ by:

$$\hat{y} = D_{mask}(X_{VGCA}, E^{sparse}_{prompt}(P)). \tag{9}$$

Here, to effectively adapt the mask decoder to visual-prompt-conditioned image embeddings X_{VGCA} in a parameter-efficient manner, we incorporate Low-Rank Adaptation (LoRA) [8] into the decoder by updating frozen linear projections with learnable low-rank residuals. Details are described below.

Let $X \in \mathbb{R}^{n \times c_{\text{in}}}$ denote an image embedding and $W \in \mathbb{R}^{c_{\text{out}} \times c_{\text{in}}}$ be a frozen linear projection producing $\hat{X} = WX$. In the LoRA layer, the adapted weight is given by:

$$\tilde{W} = W + BA, \tag{10}$$

where $A \in \mathbb{R}^{r \times c_{\text{in}}}$ and $B \in \mathbb{R}^{c_{\text{out}} \times r}$ are learnable linear layer with rank $r \ll \min(c_{\text{in}}, c_{\text{out}})$. During training, W remains fixed and only (A, B) are updated, enabling parameter-efficient adaptation with minimal additional parameters.

Following this strategy, we insert LoRA into selected cross-attention projections of the mask decoder according to the role of image embeddings in each pathway. Specifically, in token-to-image (t2i) attention (Fig. 2(b)), sparse prompt tokens serve as queries (Q) to retrieve spatially relevant image cues for target localization, while image embeddings act as keys (K) and values (V). Since image embeddings are projected into K and V, LoRA is applied to the key, value, and output projection matrices, i.e., $\tilde{W}_K^{t2i}$, $\tilde{W}_V^{t2i}$, and $\tilde{W}_O^{t2i}$. In contrast, in image-to-token (i2t) attention (Fig. 2(c)), image embeddings serve as queries to attend back to sparse prompt tokens for feature refinement, while sparse prompt tokens act as keys and values. Accordingly, LoRA is applied to the query and output projection matrices, i.e., $\tilde{W}_Q^{i2t}$ and $\tilde{W}_O^{i2t}$.

Together, the t2i and i2t attention layers enable bidirectional interaction between sparse prompts and dense image features, while selective LoRA placement facilitates efficient adaptation of the mask decoder to visual-prompt-conditioned representations without full fine-tuning, resulting in accurate lesion segmentation on non-contrast images.

3 Experiments

Dataset and Metric To evaluate the proposed cross-modality visual prompting framework, we used a liver lesion dataset for proton therapy planning collected at Samsung Medical Center. The dataset comprises 73 cases, each including a mid-respiratory phase NCCT (T = 50%), a T1-weighted fat-suppressed contrast-enhanced MRI at the same respiratory phase, and expert-annotated liver and lesion masks defined on the NCCT. MRI scans were rigidly registered to the NCCT, and all scans were resampled to a voxel spacing of $0.6597 \times 0.6597 \times 5$ mm. NCCT scans were clipped to [-100, 200] HU. For both CT and MRI, each 2D axial slice was resized to 1024×1024 and normalized to [0, 1] using min-max scaling. The dataset was split into 60, 6, and 7 cases for training, validation, and testing, respectively. In total, 2,251, 213, and 245 2D axial slices containing liver lesions were used for each split. For quantitative evaluation, segmentation performance was assessed using the Dice similarity coefficient (DSC), Intersection over Union (IoU), and the 95th percentile Hausdorff distance (HD95).

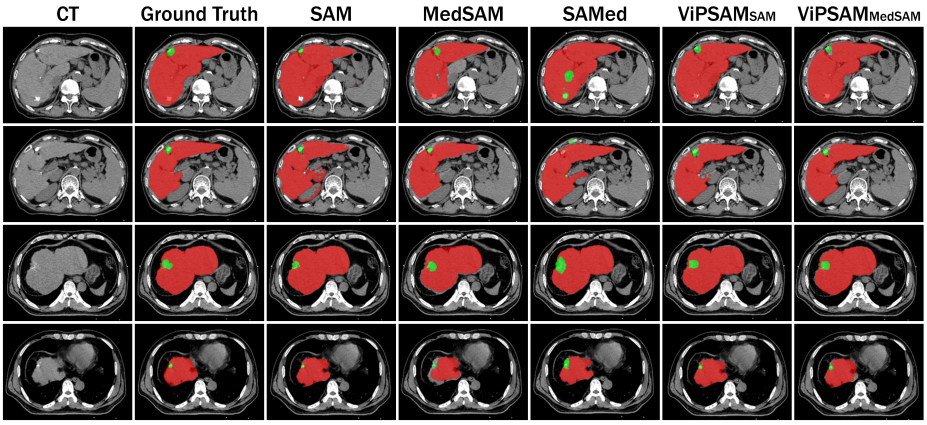

Fig. 3. Qualitative comparison results of liver lesion segmentation on non-contrast CT. Red denotes the liver and green denotes the liver lesion.

Table 1. Quantitative comparison results of the average Dice, IoU, HD95 and the number of trainable parameters. Bold indicates the best performance for each metric.

Methods	Liver			Lesion			Params
	Dice ↑	IoU ↑	HD95(mm) ↓	Dice ↑	IoU ↑	HD95(mm) ↓	(M) ↓
U-Net	0.875±0.162	0.805±0.187	27.62±31.13	0.236±0.298	0.174±0.240	60.42±31.35	4.32
nnU-Net	0.856±0.238	0.799±0.244	28.72±41.09	0.267±0.339	0.209±0.281	43.23±28.12	59.18
TransUNet	0.861±0.198	0.793±0.219	22.29±26.83	0.169±0.189	0.106±0.132	77.92±46.17	105.28
Swin-Unet	0.880±0.155	0.809±0.178	24.38±26.85	0.162±0.189	0.101±0.128	56.25±25.63	27.17
SAM	0.890±0.111	0.815±0.138	18.80±13.27	0.655±0.208	0.520±0.215	13.11±10.02	93.74
MedSAM	0.892±0.105	0.817±0.136	13.59±12.29	0.709±0.180	0.577±0.206	8.39±4.76	93.74
SAMed	0.901±0.128	0.838±0.155	18.88±27.22	0.341±0.328	0.259±0.270	48.59±28.12	3.93
ViPSAM$_{\text{SAM}}$	0.925±0.088	0.869±0.114	14.58±17.68	**0.852±0.114**	**0.757±0.150**	**5.15±3.70**	0.83
ViPSAM$_{\text{MedSAM}}$	**0.937±0.070**	**0.887±0.094**	**8.84±9.49**	0.810±0.151	0.704±0.188	5.18±2.18	0.83

Implementation Details We implemented the proposed ViPSAM using either SAM [11] (ViPSAM$_{\text{SAM}}$) or MedSAM [12] (ViPSAM$_{\text{MedSAM}}$) as the backbone, each initialized with pretrained weights. For the VGCA module, the feature channel dimension was set to $C = 256$ and the number of attention heads was set to 8. We initialize the gating scalar γ to 0.3, the weight λ to 1.0, and the LoRA rank r to 8. Box prompts were generated from manual masks with random perturbations ($\leq$20 px) during training and fixed during inference. The model was trained for 12 epochs using an equally weighted sum of Dice and binary cross-entropy losses and optimized with AdamW with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-4}. The checkpoint from the final epoch was used for testing. All experiments were conducted in PyTorch 2.5.1 on an NVIDIA RTX A6000 GPU. Memory usage was approximately 14.3GB during training and 4.8GB during inference. The source code is publicly available at https://github.com/torchViPSAM/ViPSAM.

Table 2. Ablation study of ViPSAM on prompt design and trainable components.

Config.	Prompt		Layer		Liver			Lesion		
	Visual	Sparse	CA	LoRA	Dice ↑	IoU ↑	HD95(mm) ↓	Dice ↑	IoU ↑	HD95(mm) ↓
(a)		✓		✓	0.909±0.091	0.842±0.119	12.04±11.02	0.801±0.173	0.695±0.187	5.84±3.14
(b)	✓		✓	✓	0.809±0.236	0.728±0.247	51.85±44.99	0.155±0.176	0.095±0.116	105.91±43.41
(c)	✓	✓	✓		0.932±0.081	0.881±0.108	9.62±10.68	0.808±0.136	0.669±0.181	5.04±1.87
Ours	✓	✓	✓	✓	0.937±0.070	0.887±0.094	8.84±9.49	0.810±0.151	0.704±0.188	5.18±2.18

Experimental Results We compared ViPSAM with representative segmentation models, including U-Net-based methods (U-Net [15], nnU-Net [10], TransUNet [4], and Swin-Unet [2]) and SAM-based methods using box prompts (SAM [11], MedSAM [12], and SAMed [20]). Baseline models were trained using only NCCT blue with the default settings provided in the original papers, whereas ViPSAM additionally leveraged contrast-enhanced MRI as visual prompts. All methods were evaluated on NCCT for liver and liver lesion segmentation.

Fig. 3 visualizes qualitative comparisons on NCCT. ViPSAM produces more accurate segmentation masks than the existing SAM-based models, even for small lesions with indistinct boundaries. As reported in Table 1, the proposed ViPSAM consistently outperformed both U-Net- and SAM-based methods across all evaluation metrics. Specifically, $\text{ViPSAM}_{\text{MedSAM}}$ achieved the best liver segmentation performance with a Dice score of 0.937 and an HD95 of 8.84 mm, while $\text{ViPSAM}_{\text{SAM}}$ achieved the best lesion segmentation performance with a Dice score of 0.852 and an HD95 of 5.15 mm. In contrast, U-Net-based models showed poor lesion segmentation performance under low-contrast NCCT conditions (Dice < 0.267), and SAM-based methods exhibited imprecise boundary delineation for lesions despite box-prompt guidance. Also, our method required only 0.83M trainable parameters, highlighting superior parameter efficiency.

Moreover, we assessed statistical significance using paired t-tests and Wilcoxon signed-rank tests. Both $\text{ViPSAM}_{\text{SAM}}$ and $\text{ViPSAM}_{\text{MedSAM}}$ significantly outperformed their respective baselines (SAM and MedSAM) across all metrics ($p < 0.05$) under both tests. These results support the effectiveness of our model leveraging MRI-derived visual prompting for liver lesion segmentation on NCCT.

Ablation Study To evaluate each component in ViPSAM, we conducted an ablation study with three configurations: (a) ViPSAM without visual prompting, (b) ViPSAM without sparse box prompts, and (c) ViPSAM without LoRA-based adaptation. All configurations were trained under the same settings for fair comparison.

Table 2 shows quantitative results for liver and lesion segmentation across the ablated models. Without visual prompting (a), performance decreased to a lesion Dice of 0.801 and an IoU of 0.695, compared to the full ViPSAM (Dice=0.810, IoU=0.704), highlighting the importance of incorporating cross-modal contrast information. Moreover, removing sparse box prompts (b) led to further performance degradation, suggesting their role in stabilizing lesion localization. In addition, without LoRA-based adaptation (c), the model showed reduced per-

formance with a Dice of 0.808 and an IoU of 0.669. These results indicate that visual prompting with contrast-enhanced images, VGCA-based cross-modal interaction, and LoRA-based decoder adaptation contribute synergistically to improved liver lesion segmentation on NCCT.

4 Conclusion

In this work, we propose ViPSAM, a SAM-based visual prompting framework for non-contrast medical image segmentation. By leveraging complementary information from contrast-enhanced images via a visual prompt encoder, visual-guided cross-attention, and LoRA-based parameter-efficient decoder adaptation, ViPSAM enables precise segmentation with minimal additional learnable parameters. Experimental results on a liver lesion proton therapy planning dataset demonstrate the effectiveness of the proposed framework in alleviating boundary ambiguity in non-contrast imaging, highlighting its potential to enhance segmentation reliability in clinical workflows.

Acknowledgments. This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2026-25481239), AI Graduate School Support Program (Sungkyunkwan University) (RS-2019-II190421), and SKKU Academic Research Support Program, Sungkyunkwan University, 2025.

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

References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
2. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European conference on computer vision. pp. 205–218. Springer (2022)
3. Chang, J.Y., Zhang, X., Knopf, A., Li, H., Mori, S., Dong, L., Lu, H.M., Liu, W., Badiyan, S.N., Both, S., et al.: Consensus guidelines for implementing pencil-beam scanning proton therapy for thoracic malignancies on behalf of the ptcog thoracic and lymphoma subcommittee. International Journal of Radiation Oncology* Biology* Physics **99**(1), 41–50 (2017)
4. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A.L., Zhou, Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306 (2021)
5. Cheng, J.Y., Liu, C.M., Wang, Y.M., Hsu, H.C., Huang, E.Y., Huang, T.T., Lee, C.H., Hung, S.P., Huang, B.S.: Proton versus photon radiotherapy for primary hepatocellular carcinoma: a propensity-matched analysis. Radiation Oncology **15**(1), 159 (2020)
6. Cheung, A.L.Y., Zhang, L., Liu, C., Li, T., Cheung, A.H.Y., Leung, C., Leung, A.K.C., Lam, S.K., Lee, V.H.F., Cai, J.: Evaluation of multisource adaptive mri fusion for gross tumor volume delineation of hepatocellular carcinoma. Frontiers in oncology **12**, 816678 (2022)

7. Cox, J., Liu, P., Stolte, S.E., Yang, Y., Liu, K., See, K.B., Ju, H., Fang, R.: Brainsegfounder: Towards 3d foundation models for neuroimage segmentation. Medical Image Analysis **97**, 103301 (2024)
8. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
9. Iannaccone, R., Laghi, A., Catalano, C., Rossi, P., Mangiapane, F., Murakami, T., Hori, M., Piacentini, F., Nofroni, I., Passariello, R.: Hepatocellular carcinoma: role of unenhanced and delayed phase multi–detector row helical ct in patients with cirrhosis. Radiology **234**(2), 460–467 (2005)
10. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
11. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
12. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
13. Marshall, C., Thirion, P., Mihai, A., Armstrong, J.G., Cournane, S., Hickey, D., McClean, B., Quinn, J.: Interobserver variability of gross tumor volume delineation for colorectal liver metastases using computed tomography and magnetic resonance imaging. Advances in Radiation Oncology **8**(1), 101020 (2023)
14. Newhauser, W.D., Zhang, R.: The physics of proton therapy. Physics in Medicine & Biology **60**(8), R155 (2015)
15. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
16. Sherer, M.V., Lin, D., Elguindi, S., Duke, S., Tan, L.T., Cacicedo, J., Dahele, M., Gillespie, E.F.: Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: A critical review. Radiotherapy and Oncology **160**, 185–191 (2021)
17. Shi, P., Qiu, J., Abaxi, S.M.D., Wei, H., Lo, F.P.W., Yuan, W.: Generalist vision foundation models for medical imaging: A case study of segment anything model on zero-shot medical segmentation. Diagnostics **13**(11), 1947 (2023)
18. Tsai, Y.L., Wu, C.J., Shaw, S., Yu, P.C., Nien, H.H., Lui, L.T.: Quantitative analysis of respiration-induced motion of each liver segment with helical computed tomography and 4-dimensional computed tomography. Radiation Oncology **13**(1), 59 (2018)
19. Velec, M., Moseley, J.L., Eccles, C.L., Craig, T., Sharpe, M.B., Dawson, L.A., Brock, K.K.: Effect of breathing motion on radiotherapy dose accumulation in the abdomen using deformable registration. International Journal of Radiation Oncology* Biology* Physics **80**(1), 265–272 (2011)
20. Zhang, K., Liu, D.: Customized segment anything model for medical image segmentation. arXiv preprint arXiv:2304.13785 (2023)
21. Zhang, Y., Zhou, T., Wang, S., Liang, P., Zhang, Y., Chen, D.Z.: Input augmentation with sam: Boosting medical image segmentation with segmentation foundation model. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 129–139. Springer (2023)