R1Seg-3D: Rethinking Reasoning Segmentation for Medical 3D CTs

Qin Hao[1], Long Yu[2], Shengwei Tian[3], Xujiong Ye[4], Lei Zhang[4]

[1] School of Computer Science and Technology, Xinjiang University, Urumqi, China
haoqin@stu.xju.edu.cn
[2] Network Center, Xinjiang University, Urumqi, China
yul@xju.edu.cn
[3] College of Software, Xinjiang University, Urumqi, China
[4] Department of Computer Science, University of Exeter, Exeter, UK

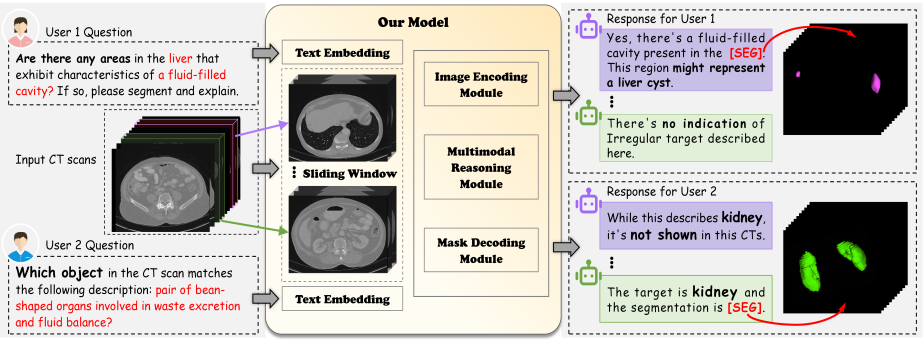
Abstract. The explosive development of large-scale model technology has provided strong support for achieving more intelligent, robust, and precise segmentation techniques. However, owing to the unique challenges posed by medical domain data, the typical 3D medical image-text alignment model, 3D CLIP, struggles to match the performance of its natural scene counterpart. This limitation hinders the application of CLIP-based text-image reasoning in medical segmentation tasks. Furthermore, CLIP has been shown to rely on high-level semantic alignment between vision and text, lacking effective support for local visual features that are crucial for dense prediction tasks. Existing reasoning segmentation methods often adopt a redundant design with two visual encoders—one from CLIP and the other from large vision models for downstream dense tasks. This adversely affects model efficiency and complicates the training process. To address these challenges, we propose a novel framework, R1Seg-3D, which unifies a visual encoder. Our approach achieves a three-way alignment of dense visual, text reasoning, and mask decoding features within a shared latent space. Compared with previous methods, R1Seg-3D implicitly incorporates more detailed spatial features into the reasoning path. Therefore, it can strengthen the reasoning ability by incorporating additional visual spatial details and directly enhances the mask decoding process. The R1Seg-3D architecture is more concise and easier to be trained. Extensive evaluations on 25 diverse datasets demonstrate that R1Seg-3D outperforms state-of-the-art methods in both performance and stability. This work advances intelligent medical imaging and lays a foundation for future research in inference-driven segmentation. Our code and models are available at https://github.com/lihaoqin168/R1Seg-3D.

Keywords: Reasoning Segmentation, Multimodal Large Language Model, Medical 3D CTs Segmentation.

1 Introduction

Accurate segmentation of anatomical structures and pathological regions from 3D CT scans is crucial for diagnosis, treatment planning, and surgical navigation [1]. In clinical practice, physicians often prefer to provide concise instructions, such as "Are there any areas in the liver that exhibit characteristics of a fluidfilled cavity?" to guide medical image analysis systems. However, existing systems typically rely on explicit predefined categories to perform segmentation, lack the ability to reason about implicit clinical intentions. Medical segmentation traditionally relies on 3D convolutional neural networks (e.g., 3D U-Net [2] and nnU-Net [3]) and transformer-based models (e.g., TransUNet [4] and UNETR [5]) for predefined categories. While effective, these methods lack adaptability to dynamic or unseen tasks, limiting their utility in diverse clinical scenarios. In recent years, remarkable advancements in large modeling techniques have been reported. Large vision models (LVMs), such as SAM [6] and Dinov2 [7], have shown strong zero-shot capabilities for diverse image distributions and extensive scenes. The SAM [6] introduces a paradigm shift by enabling flexible segmentation through text, box, and point prompts. This capability has sparked significant interest in medical imaging with adaptations such as MedSAM [8] and MedLSAM [9], highlighting its potential for visual medical image analysis.

Large language models (LLMs), such as LLaMA [10] and Qwen [11], have demonstrated exceptional reasoning capabilities across a wide range of natural language processing tasks[12]. The convergence of LLMs and LVMs has given rise to vision language models (VLMs) [13], which excel with the ability to address visual question answering and complex reasoning tasks. In the medical domain, pioneering works such as LLaVA-Med[14] and BioMedGPT [15] have demonstrated the potential of VLMs for biomedical applications. To build on the visual-linguistic understanding capabilities of the aforementioned techniques, innovative reasoning segmentation methods [16, 17, 18, 19] have been proposed. These LLM-powered segmentors have the ability to interpret, process, and reason, translating abstract linguistic queries into specific pixel regions within real-world contexts. LISA [16] and its variants [20, 21, 22] have achieved remarkable success in natural scenes, representing a significant advancement toward the development of more intelligent vision systems. However, applying these techniques to 3D medical CT scenarios remains challenging.



Are there any areas in the liver that exhibit characteristics of a fluid-filled cavity? If so, please segment and explain.

Which object in the CT scan matches the following description: pair of bean-shaped organs involved in waste excretion and fluid balance?

Yes, there's a fluid-filled cavity present in the [SEG]. This region might represent a liver cyst.

There's no indication of Irregular target described here.

While this describes kidney, it's not shown in this CTs.

The target is kidney and the segmentation is [SEG].

Fig. 1. Our work proposes an open-vocabulary segmentation method with reasoning capabilities.

More specifically, the above models [16, 17, 18, 19] integrate a dual-encoder architecture: one leverages the visual-textual capabilities of the pre-trained CLIP model, and the other facilitates prompt-based segmentation via SAM [6]. However, the limitations of medical 3D CLIP diminish its design advantages, leading to computational inefficiency and redundancy while increasing training complexity in 3D medical scenarios. The 3D volumetric nature of CT scans and the scarcity of annotated medical image-text pairs pose significant challenges for training 3D medical CLIP [23]. The requirement to embed entire CT volumes at a global scale restricts the model's ability to learn from local regions of interest, which may be small or, in some cases, entirely absent. Consequently, CLIP's image-text representation in 3D medical contexts is less effective than that in natural scenes. Furthermore, CLIP visual representations often lack fine-grained detail, particularly in tasks requiring precise spatial information [24, 25]. This limits its ability to capture subtle pathological features, such as small lesions or low-contrast regions, which are critical for accurate diagnosis [26, 27, 28].

To address these challenges, we propose R1Seg-3D (Fig. 1 and Fig. 2), a novel framework that integrates the LLM and LVM techniques and is specifically designed for 3D medical imaging reasoning segmentation. Our main contributions are as follows: (1) We propose the first 3D reasoning segmentation method with a unified visual encoding framework, which improves both the efficiency and accuracy. (2) We align dense visual and textual features within shared latent spaces, enhancing reasoning with LLMs. This alignment effectively propagates to the mask decoding stage, ensuring robust reasoning and precise segmentation in complex medical scenarios. (3) R1Seg-3D's ability to interpret implicit clinical instructions and perform reasoning-driven precision segmentation represents a significant advance toward next-generation intelligent medical imaging systems.

2 Methodology

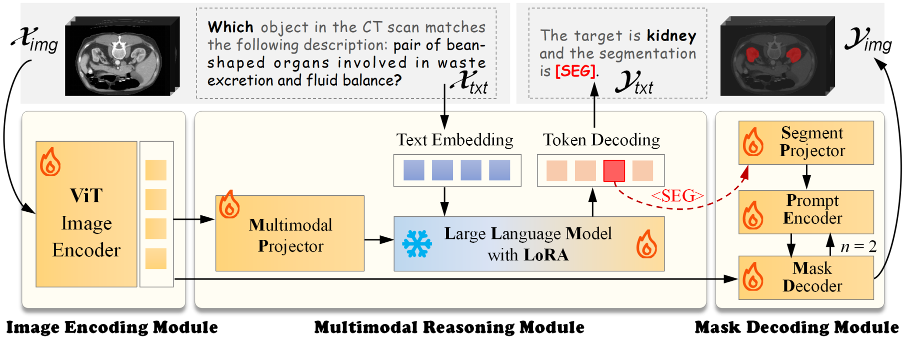

$\mathcal{X}_{img}$
Which object in the CT scan matches the following description: pair of bean-shaped organs involved in waste excretion and fluid balance?
$\mathcal{X}_{txt}$
The target is kidney and the segmentation is [SEG].
$\mathcal{Y}_{txt}$
$\mathcal{Y}_{img}$
Text Embedding
Token Decoding
ViT Image Encoder
Multimodal Projector
Large Language Model with LoRA
<SEG>
Segment Projector
Prompt Encoder
$n = 2$
Mask Decoder
Image Encoding Module
Multimodal Reasoning Module
Mask Decoding Module

Fig. 2. Model structure of our R1Seg-3D. Unlike previous methods that utilize both CLIP and SAM visual encoders, our approach employs a unified image encoder to extract dense features, which are simultaneously fed into a multimodal reasoning module and a mask decoding module. These modules collaboratively align dense image features with linguistic reasoning and mask decoding features to produce both textual responses and segmentation outputs.

Our proposed R1Seg-3D, detailed in Fig. 2, consists of three key components: a 3D image encoding module, a multimodal reasoning module, and a mask decoding module.

The 3D image encoding module, which is based on a Vision Transformer (ViT) architecture, extracts dense visual features from the input CT scan. These features are then projected into a shared latent space by the multimodal projector, aligning them with language reasoning features generated by the LLM enhanced with Low-Rank Adaptation (LoRA). The LLM integrates dense visual features with textual descriptions (e.g., "pair of bean-shaped organs involved in waste excretion and fluid balance") to generate outputs, which may include a special <SEG> token indicating the segmentation target, such as "kidney." The segment projector processes this token to produce segmentation prompts, which are further refined by the prompt encoder. Finally, the mask decoder generates 3D segmentation masks (e.g., for the kidneys) by combining visual features and segmentation prompts. This unified visual encoder architecture enhances model efficiency and improves reasoning capabilities for dense tasks by enriching the LLM's reasoning processes with spatial information.

Given an input 3D CT scan x_{img} and a corresponding question x_{txt}, the image encoding module f_{enc} processes x_{img} to extract dense visual features y'_{img}. These features are subsequently passed through a multimodal projector f_{mp}, which comprises three components: a 3D average-pooling layer, a single-layer linear transformation, and a two-layer multilayer perceptron (MLP). This projector bridges the gap between visual features and language features, facilitating multimodal integration. Simultaneously, the question x_{txt} is fed into the multimodal reasoning module, where a pre-trained large language model (LLM) f_{llm} embeds x_{txt} and generates a contextualized response y'_{txt} by integrating the visual features provided by the projector f_{mp}. This can be formulated as:

$$y'_{img} = f_{enc}\,(x_{img}) \tag{1}$$

$$y'_{txt} = f_{llm}\,(x_{txt}, f_{mp}\,(y'_{img})) \tag{2}$$

When the LLM's output y'_{txt} indicates that the image contains the segmentation target implied by the question x_{txt}, y'_{txt} includes a <SEG> token. During the forward pass, the <SEG> token is updated to encapsulate information about the segmentation target. The mask decoding module generates mask proposals guided by the <SEG> token, and the final segmentation output is generated by combining the selected proposals. We use an MLP-based segment projector f_{sp} to obtain h_{seg} from the embedded <SEG> token h'_{seg}. Finally, h_{seg} is input to the prompt encoder f_{pe} to produce the final segmentation mask y_{img} via the mask decoder f_{dec}. This can be expressed as:

$$h_{seg} = f_{sp}\,(h'_{seg}) \tag{3}$$

$$y_{img} = f_{dec}\,(f_{pe}\,(h_{seg}),\, y'_{img}) \tag{4}$$

To refine the segmentation mask iteratively, we perform a loop operation n times during mask decoding, with $n = 2$ determined experimentally to balance computational efficiency and segmentation accuracy.

3 Experiments and Results

Implementation Details. We employed a comprehensive dataset, integrating 25 open-source datasets [29][1], which comprised 6,000 CT scan images and 150,000 pixel-level annotations. Leveraging this dataset, we utilized Qwen-72B to enrich the textual content and generate detailed CT quiz test paired data. To standardize the datasets, we applied min-max normalization uniformly to preprocess the voxels. The voxel spacing of the CT images in this study was resampled to $2.0 \times 1.0 \times 1.0$ mm^3. During training, the input images were randomly cropped to volumes of size [32, 256, 256], and the foreground and background patches were sampled at a ratio of 2:1 to ensure balanced representation. During inference, the sliding window strategy was used to process the entire volume sequentially; thus, each input matched the training volume size of [32, 256, 256]. For evaluation, pixel-level metrics, including the Dice coefficient (F1 score), precision (Pre), sensitivity (SE, or recall), specificity (SPE), and false positive (FP), were employed to assess semantic segmentation performance across models.

The training of R1Seg-3D is conducted in four stages, illustrated here using Phi3 as the LLM. In the first stage, R1Seg-3DSAM—which combines the 3D image encoding module and the mask decoding module from R1Seg-3D with a frozen text encoder—is trained on 25 labeled volumetric medical image segmentation datasets for 200 epochs, with a batch size of 16 and an input size of [32, 256, 256]. In the second stage, only the 3D multimodal projector is trained via image-text pairs for 3 epochs, with a batch size of 8×6 and a learning rate of 10^{-4}. The third stage involves fine-tuning the multimodal reasoning module while keeping the other two modules frozen, using image-question-answer-mask data for 3 epochs, with a batch size of 6×6 and a learning rate of 5×10^{-5}. This stage uses the parameter-efficient LoRA technique, with LoRA parameters set to $r = 16$, $\alpha = 32$, and a dropout rate of 0.1. In the final stage, all three modules are fine-tuned together for 3 epochs to further refine the model's performance. Mixed-precision training (bf16) is enabled by DeepSpeed, and all the experiments are implemented in PyTorch and parallelized across 6 NVIDIA A40 GPUs with 48 GB memory. Training is completed in approximately 4 days.

Comparision with M3D-LaMed. To the best of our knowledge, M3D-LaMed [23], based on LISA with a dual-encoder architecture, is the state-of-the-art reference expression segmentation model in medical 3D imaging; thus, we quantitatively compared our R1Seg-3D (Phi 3) with M3D-LaMed. The boxplot in Fig. 3 demonstrates that R1Seg-3D outperforms M3D-LaMed overall, with higher median and mean values across most datasets. M3D-LaMed has a wider interquartile range (IQR) in 14 datasets, indicating greater variability in its performance, whereas R1Seg-3D has a more consistent performance distribution. Notably, M3D-LaMed fails to accurately identify colon cancer targets in the MSD-Colon dataset, likely because of the complexity and small size of colon cancer regions.

[1] https://github.com/BAAI-DCAI/SegVol

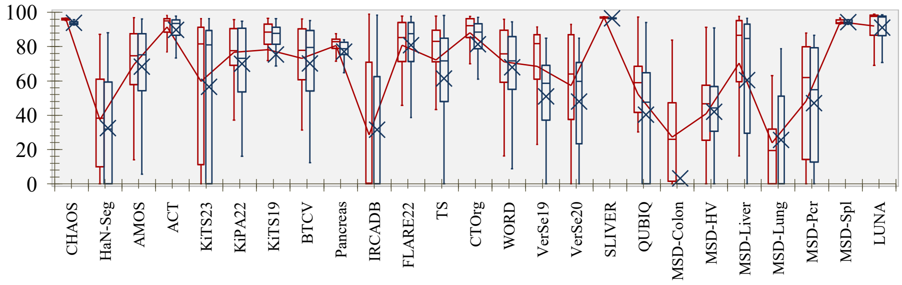


Fig. 3. Comparison of our R1Seg-3D (red) and the existing 3D medical inference segmentation method, M3D-LaMed (black), across 25 datasets in terms of F1 scores (%).

As shown in Table 1, R1Seg-3D achieves significant improvements over M3D-LaMed, with average score increases of +8.18% in F1, +6.3% in precision, and +7.64% in sensitivity. Specifically, R1Seg-3D outperforms M3D-LaMed in key datasets such as ACT (F1: 91.41% vs. 89.84%), CTOrg (F1: 88.03% vs. 81.41%), and TS (F1: 72.23% vs. 62.58%). These results underscore the robustness and generalizability of our R1Seg-3D across diverse medical image segmentation tasks.

Table 1. Segmentation performance comparison with the state-of-the-art method M3D-LaMed.

Method	F1(%)				Pre (%)				SE(%)			
	Mean	ACT	CTOrg	TS	Mean	ACT	CTOrg	TS	Mean	ACT	CTOrg	TS
M3D-LaMed	64.14	89.84	81.41	62.58	69.14	89.45	82.88	67.91	65.67	90.58	84.51	64.19
R1Seg-3D	72.32	91.41	88.03	72.23	75.44	91.89	88.40	75.01	73.31	91.35	89.49	73.33

Comparison of Different LLMs in the Reasoning Module. We evaluate the impact of integrating Phi, LLaMA, Qwen, and LLaVA-Med (pretrained on the medical domain) into the reasoning module of R1Seg-3D. We report the average performance across 25 datasets and F1 scores on six specific datasets: TS (104 labels), WORD (16 labels), AMOS (15 labels), CTOrg (6 labels), ACT (3 labels), and LUNA (3 labels).

Table 2. Performance comparison of four LLMs integrated into the reasoning module.

LLM	Size	F1 (%)	Pre (%)	SE(%)	TS	WORD	AMOS	CTOrg	ACT	LUNA
Phi 3	4B	72.32	75.44	73.31	72.23	71.19	69.51	88.03	91.41	92.03
Qwen 2.5	7B	72.37	75.20	73.36	72.12	71.66	70.89	88.77	91.34	93.00
Llama 3	8B	**76.72**	**79.40**	**77.47**	**77.02**	**76.06**	**73.92**	**89.46**	**92.00**	**93.46**
Llava-Med	7B	72.79	76.16	73.71	72.38	75.16	72.74	89.71	**92.00**	93.29

As shown in Table 2, LLaMA 3 (8B) consistently outperforms the other models across most metrics and excels in domain-specific tasks. In contrast, Phi 3 (4B) achieves competitive performance despite its smaller size, demonstrating efficiency in resource-constrained scenarios. These results highlight the importance of model size and architecture in optimizing medical image segmentation performance. Larger LLMs, such as LLaMA 3, offer superior accuracy and generalization for the reasoning module, whereas smaller models, such as Phi 3, provide a viable trade-off between performance and efficiency, making them suitable for resource-constrained environments.

Effectiveness of the Sliding window Strategy. Owing to the larger volume of 3D images than 2D images, we employ a sliding window approach to perform reasoning and segmentation sequentially, subsequently combining the results to form the final mask (as shown in Fig. 1). To assess the effectiveness of this strategy, we selected eight segmentation targets across 25 datasets, including four organs (large-volume liver, medium-volume stomach, complex-shaped pancreas, and small-volume gland) and four lesion labels (lesion, cyst, tumor, and cancer). Fig. 4 shows that the left-side violin plots (with sliding window) are more concentrated in upper regions than the right (without sliding window), indicating enhanced segmentation accuracy. The bottom row, representing lesion segmentation, exhibits high density below 20%, reflecting frequent segmentation failures. Fig. 4 highlights challenges in medical segmentation: the scarcity of datasets with region-level lesion annotations and textual descriptions, and the limitations of existing algorithms in accurately segmenting complex or small targets.

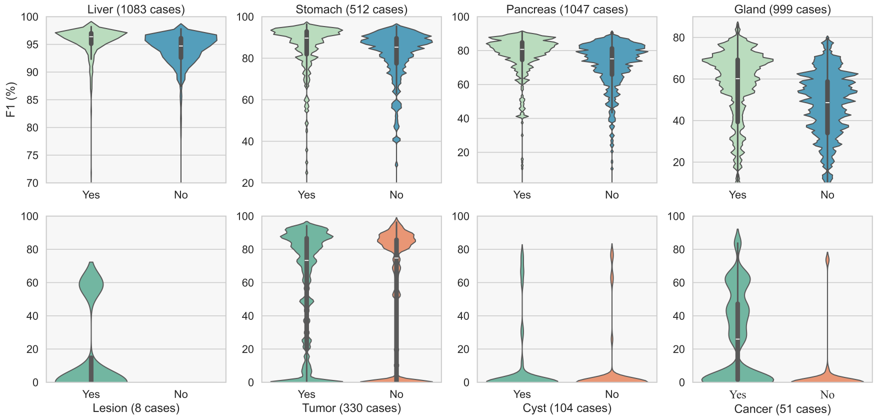


Fig. 4. Violin plots comparing segmentation performance with and without sliding window.

Effectiveness of Reasoning Modual. We perform an ablation study on the reasoning module to demonstrate that reasoning-enabled segmentation not only segments implicit targets in open-vocabulary settings but also improves target presence detection, reducing false positives in results. The specificity results in Table 3 demonstrate that the reasoning module's pre-judgment of the existence of segmentation targets reduces the number of false positive cases. R1Seg-3D with reasoning significantly reduces false positives across all datasets, decreasing from 2,146 to 592 in the total dataset, while enhancing specificity (SPE) from 82.95% to 94.63% compared to R1Seg-3DSAM (without the reasoning module). Similar notably improvements are observed on the TS, VerSe and IRCADB datasets. These results indicate that segmentation methods with reasoning capabilities, which leverage contextual reasoning and textual inputs, have significant potential for future applications in disease diagnosis.

Table 3. Evaluating the effectiveness of the reasoning module.

With/Without Reasoning	Total (10439)		TS (9678)		VerSe (478)		IRCADB (149)		Others (137)	
	FP↓	SPE(%)↑	FP↓	SPE(%)↑	FP↓	SPE(%)↑	FP↓	SPE(%)↑	FP↓	SPE(%)↑
R1Seg-3DSAM	2146	82.95	1764	84.58	191	71.45	96	60.33	95	59.05
R1Seg-3D	592	94.63	452	95.54	73	86.75	23	86.39	44	75.69

Effectiveness of the *n*-Time Loop. Our ablation study on the loop operation (*n*) in mask decoding demonstrates that iterative refinement of the segmentation mask significantly enhances performance. The loop operation, which involves feeding the embedded <SEG> token and resulting mask back into the prompt encoder, improves the model's ability to refine segmentation boundaries and capture fine-grained details. Table 4 shows that, increasing the loop count from $n = 0$ to $n = 1$ results in a substantial improvement in the mean F1 score from 67.08% to 72.12%, with the precision and sensitivity increasing from 69.58% to 76.32% and from 69.52% to 72.13%, respectively. Increasing the loop count to $n = 2$ further improved performance, achieving a mean F1 score of 72.32% (+0.1%) and sensitivity of 73.31% (+1.18%) over $n = 1$, thereby reducing the risk of missing critical anatomical or pathological regions. However, when $n = 3$, the performance gains are minimal, indicating diminishing returns. These results demonstrate that using at least one loop ($n \geqslant 1$) is highly beneficial, with $n = 2$ striking the optimal balance between accuracy and computational efficiency. These findings underscore the importance of iterative refinement for precise, reliable segmentation.

Table 4. The *n* times loop operation during mask decoding.

n-Time Loop	F1(%)				Pre (%)				SE(%)			
	Mean	ACT	CTOrg	TS	Mean	ACT	CTOrg	TS	Mean	ACT	CTOrg	TS
$n = 0$	67.08	83.40	85.41	67.51	69.58	84.25	87.03	69.59	69.52	83.38	86.18	70.05
$n = 1$	72.12	91.27	87.13	72.15	**76.32**	**92.64**	**89.11**	**75.80**	72.13	90.43	87.98	72.35
$n = 2$	72.32	**91.41**	88.03	72.23	75.44	91.89	88.40	75.01	73.31	91.35	89.49	73.33
$n = 3$	**72.37**	91.36	**88.39**	**72.38**	75.62	91.85	88.39	75.25	**73.33**	**91.37**	**90.01**	**73.52**

4 Discussion and Conclusion

To advance the development of intelligent 3D medical imaging systems capable of interpreting implicit clinical objectives, we propose R1Seg-3D, a reasoning-driven open-vocabulary segmentation method designed for increased efficiency and accuracy. Our approach integrates state-of-the-art techniques including LLMs and LVMs, introducing a novel unified visual encoding framework that addresses the current challenges in 3D medical segmentation. Furthermore, we enhance segmentation precision through sliding windows and an iterative refinement mechanism, significantly reducing misjudgments. Extensive experiments demonstrate that R1Seg-3D outperforms existing methods across multiple specific tasks, providing a robust and generalizable solution for diverse medical imaging applications. The success of R1Seg-3D stems from its ability to align dense visual features with textual features via a unified visual encoder, ensuring consistent and accurate mask generation.

In conclusion, our proposed framework addresses the limitations of existing methods, offering a more flexible and accurate approach to segmenting anatomical and pathological regions in 3D CT scans. This work represents a significant step forward in 3D medical image segmentation. Future research will focus on improving the reasoning and segmentation of challenging targets, such as combining 3D imaging with patient symptom descriptions to predict and segment more complex pathological regions.

Acknowledgments. This work was supported by the National Natural Science Foundation of China [grant number 62162058] and the project of Tianshan Talent Training Program 2023TSYCLJ0023.

Disclosure of Interests. There is no conflict of interest in this work. We declare that we have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Cerrolaza, J.J., Picazo, M.L., Humbert, L., Sato, Y., Rueckert, D., Ballester, M.Á.G., Linguraru, M.G.: Computational anatomy for multi-organ analysis in medical imaging: A review. Medical Image Analysis 56, 44–67 (2019)
2. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: Learning dense volumetric segmentation from sparse annotation. Lecture Notes in Computer Science, 424–432 (2016)
3. Isensee, F., Jaeger, P.F., Kohl, S.A.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)
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:210204306, (2021)
5. 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 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 1748–1758. IEEE, Waikoloa, HI, USA (2022)
6. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., Dollár, P., Girshick, R.: Segment anything. In: Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3992–4003. IEEE, (2023)
7. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:230407193, (2024)

8. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications 15(1), 654 (2024)
9. Lei, W., Xu, W., Li, K., Zhang, X., Zhang, S.: MedLSAM: Localize and segment anything model for 3D CT images. Medical Image Analysis 99, 103370 (2025)
10. AI@Meta: Llama 3 Model Card. (2024)
11. Qwen Team: Qwen2.5: A party of foundation models. (2024)
12. Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P.S., Yang, Q., Xie, X.: A survey on evaluation of large language models. ACM Transactions on Intelligent Systems and Technology 15(3), 1–45 (2024)
13. Zhang, J., Huang, J., Jin, S., Lu, S.: Vision-language models for vision tasks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 46(8), 5625–5644 (2024)
14. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems 36, 28541–28564 (2023)
15. Luo, Y., Zhang, J., Fan, S., Yang, K., Hong, M., Wu, Y., Qiao, M., Nie, Z.: BioMedGPT: An open multimodal large language model for bioMedicine. IEEE Journal of Biomedical and Health Informatics, 1–12 (2024)
16. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: LISA: Reasoning segmentation via large language model. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9579–9589. IEEE, Seattle, WA, USA (2024)
17. Ren, Z., Huang, Z., Wei, Y., Zhao, Y., Fu, D., Feng, J., Jin, X.: PixelLM: Pixel reasoning with large multimodal model. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 26364–26373. IEEE, Seattle, WA, USA (2024)
18. Yang, S., Qu, T., Lai, X., Tian, Z., Peng, B., Liu, S., Jia, J.: LISA++: An Improved Baseline for Reasoning Segmentation with Large Language Model. arXiv preprint arXiv:231217240, (2024)
19. Zhang, Z., Ma, Y., Zhang, E., Bai, X.: PSALM: Pixelwise segmentation with large multimodal model. In: Proceedings of the Lecture Notes in Computer Science, pp. 74–91. Springer Nature Switzerland, Cham (2024)
20. Yuan, Y., Li, W., Liu, J., Tang, D., Luo, X., Qin, C., Zhang, L., Zhu, J.: Osprey: Pixel understanding with visual instruction tuning. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 28202–28211. IEEE, Seattle, WA, USA (2024)
21. Zhang, T., Li, X., Fei, H., Yuan, H., Wu, S., Ji, S., Loy, C.C., Yan, S.: Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. Advances in Neural Information Processing Systems 37, 71737–71767 (2025)
22. Rasheed, H., Maaz, M., Shaji, S., Shaker, A., Khan, S., Cholakkal, H., Anwer, R.M., Xing, E., Yang, M.-H., Khan, F.S.: GLaMM: Pixel grounding large multimodal model. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13009–13018. IEEE, Seattle, WA, USA (2024)
23. Bai, F., Du, Y., Huang, T., Meng, M.Q.-H., Zhao, B.: M3D: Advancing 3d medical image analysis with multi-modal large language models. arXiv preprint arXiv:240400578, (2024)
24. Chen, Y.-C., Li, W.-H., Sun, C., Wang, Y.-C.F., Chen, C.-S.: SAM4MLLM: Enhance multimodal large language model for referring expression segmentation. Lecture Notes in Computer Science, 323–340 (2024)

25. Wysoczańska, M., Siméoni, O., Ramamonjisoa, M., Bursuc, A., Trzciński, T., Pérez, P.: CLIP-DINOiser: Teaching CLIP a few DINO tricks for open-vocabulary semantic segmentation. Lecture Notes in Computer Science, 320–337 (2024)
26. Dong, X., Bao, J., Zheng, Y., Zhang, T., Chen, D., Yang, H., Zeng, M., Zhang, W., Yuan, L., Chen, D., Wen, F., Yu, N.: MaskCLIP: Masked self-distillation advances contrastive language-image pretraining. In: Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10995–11005. IEEE, Vancouver, BC, Canada (2023)
27. Cui, B., Islam, M., Bai, L., Ren, H.: Surgical-DINO: Adapter learning of foundation models for depth estimation in endoscopic surgery. International Journal of Computer Assisted Radiology and Surgery 19(6), 1013–1020 (2024)
28. Pérez-García, F., Sharma, H., Bond-Taylor, S., Bouzid, K., Salvatelli, V., Ilse, M., Bannur, S., Castro, D.C., Schwaighofer, A., Lungren, M.P., Wetscherek, M.T., Codella, N., Hyland, S.L., Alvarez-Valle, J., Oktay, O.: Exploring scalable medical image encoders beyond text supervision. Nature Machine Intelligence 7(1), 119–130 (2025)
29. Du, Y., Bai, F., Huang, T., Zhao, B.: Segvol: Universal and interactive volumetric medical image segmentation. Advances in Neural Information Processing Systems 37, 110746-110783 (2025)