MRSeg3D: A Perturbation-Resilient Model for 3D Medical Reasoning Segmentation

Qin Hao[1(✉)], Long Yu[2(✉)], Shengwei Tian[3], Bonian Chen[1], Cheng Zhou[4], Lei Zhang[5], Xujiong Ye[5]

[1] School of Computer Science and Technology, Xinjiang University, Urumqi, China
haoqin_1@163.com
[2] Network Center, Xinjiang University, Urumqi, China
yul@xju.edu.cn
[3] College of Software, Xinjiang University, Urumqi, China
[4] Affiliated Tumor Hospital, Xinjiang Medical University, Urumqi, China
[5] Department of Computer Science, University of Exeter, Exeter, UK

Abstract. Reasoning-based medical image segmentation has recently gained prominence by enabling complex target identification through natural language instructions. However, current research in 3D Reasoning Segmentation often prioritizes performance on curated benchmarks, leaving models vulnerable to realistic variations in instructions. To enable a systematic analysis of this fragility, we introduce a novel progressive perturbation benchmark spanning three difficulty levels. On this benchmark, state-of-the-art models show unstable performance and semantic drift, where reasoning becomes misaligned with visual evidence under perturbed instructions. We trace this issue to a critical bottleneck: insufficient visual information flow during the reasoning and mask decoding stages, which leads to cascading errors and significant performance degradation under perturbed inputs. To address these challenges, we propose MRSeg3D, a single-encoder architecture designed to facilitate tighter visual-semantic coupling. The core innovation lies in a suite of cascaded vision injectors that enforce persistent visual grounding by anchoring the model to invariant features at key decision points, thereby mitigating reasoning drift. We further introduce a Seg-Local Projector that collaborates with the mask prompt encoder to refine text-visual cues and improve the precision of [SEG] token execution. Experimental results demonstrate significant and consistent performance improvements in the Dice coefficient and text generation quality. This work provides actionable insights for next-generation intelligent clinical systems. Our code and models are available at https://github.com/lihaoqin168/MRSeg3D.

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

1 Introduction

Foundation models (FMs) based medical image segmentation frameworks are driving the development of next-generation intelligent segmentation systems by leveraging

large-scale representation learning to improve generalization, robustness, and adaptability in complex clinical environments [1]. As illustrated in **Fig. 1**, a representative 3D reasoning-based segmentation framework seamlessly integrates Large Vision Foundation Models (LVFMs) pretrained on vast, heterogeneous multimodal datasets—ranging from PubMed literature to diverse public repositories covering text, 3D/2D imaging, and image-caption pairs [2]. Architecturally, such systems typically consolidate three foundational pillars: (i) a 3D vision encoder (e.g., Vision Transformers derived from CLIP [3-4], SAM [5-7], or DINO [8-9]) to extract high-dimensional volumetric features; (ii) a Large Language Model (LLM) or Large Vision-Language Model (LVM) [10-11] to facilitate sophisticated reasoning; and (iii) a segmentation generator, often adapted from the SAM mask decoder to produce precise anatomical masks [12].

While 3D-based reasoning architectures better capture the structural complexity of medical images, clinical utility ultimately depends on the model's capacity to correctly understand and respond to human language. In clinical, physician instructions are often diverse, complex, and ambiguous, posing non-trivial challenges for medical image Reasoning Segmentation (R-Seg). However, a critical question remains under-explored: how stable are these high-performing models when exposed to the inherent uncertainty of clinical instructions? Investigating and bolstering consistency under linguistically challenging instructions thus forms the central motivation of this study.

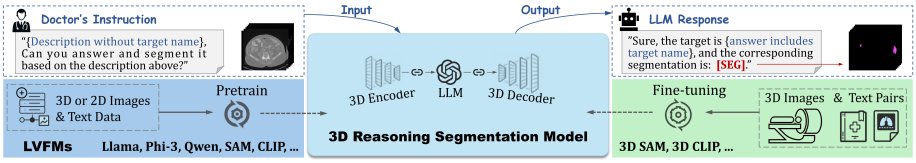

Fig. 1. The pipeline of 3D reasoning segmentation with visual-language foundation models.

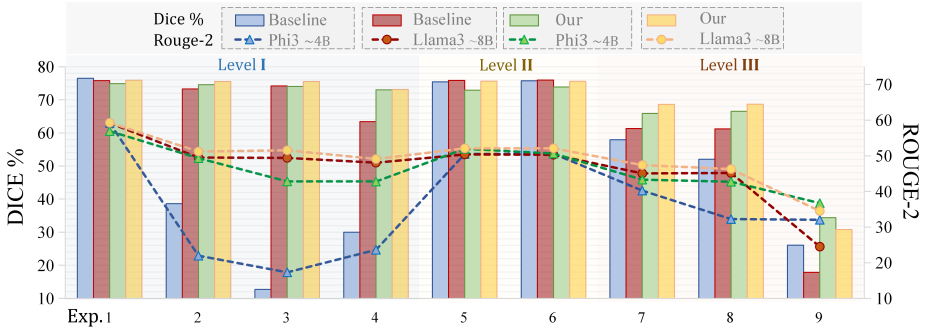

Fig. 2. Robustness evaluation under progressive perturbations (n=4,100 from 24 datasets; cf. **Table 1** and **Table 4**). The baseline refers to R1Seg-3D [13]. Note the prominent fluctuations and inconsistency in baseline models, motivating our proposed MRSeg3D.

We formalize consistency as comprising two aspects: a multi-faceted construct: (1) Instruction-level Robustness, which characterizes the stability of the reasoning process across varying linguistic complexities (ROUGE-2); and (2) Reasoning-Segmentation

Alignment, quantifying the fidelity with which accurate reasoning is reflected in precise segmentation masks (DICE). To evaluate the consistency of 3D R-Seg methods under instruction perturbation, we introduce a three-level progressive perturbation benchmark comprising nine increasingly difficult instruction templates, as detailed in **Table 1**.

Fig. 2 illustrates model performance fluctuations under this benchmark, revealing inconsistencies in both aspects. The red curve remains relatively smooth overall, showing a noticeable decline only in Exp. 9 (clinical symptoms free-text). In contrast, the blue curve exhibits significant fluctuations. This reveals that R-Seg models based on smaller LLMs (e.g., Phi-3 [14], ~4B parameters) may collapse when instructions deviate from the training distribution, demonstrating high instruction-level fragility.

Reasoning-segmentation alignment is reflected in the deviation between curves and bars. The red curves and bars reveal this gap clearly: while reasoning quality (ROUGE-2) remains stable from Exp. 2 to 4, the corresponding segmentation accuracy (Dice) exhibits a notable fluctuation, dropping at Exp. 4.

Leveraging the aforementioned observations with a comprehensive analysis of existing models (**Fig. 4**), we identify a critical limitation: as reasoning and decoding progress, the model's reliance on underlying visual evidence tends to diminish. This progressive semantic drift creates a disconnect between linguistic reasoning and visual content, rendering the entire pipeline highly susceptible to linguistic perturbations.

Motivated by these findings, we propose an optimized reasoning-segmentation framework featuring persistent visual-cue anchoring to address the insufficient integration of visual evidence at critical decision points. By maintaining a continuous link to invariant visual features, our method effectively bolsters reasoning-segmentation consistency across diverse instructions, providing actionable insights for the development of next-generation intelligent medical imaging systems [15].

2 Methodology

Table 1. Overview of proposed tiered perturbation benchmark.

Exp.	Level	Description	Label	Instruction Template
1	I	—	√	Can you segment {*Label Name*} here and output the mask?
2		Short Instruction	√	{*Randomly Select One or Two In-Distribution Samples, such as "Large organ in the upper right abdomen with various metabolic functions.", "Common site for conditions such as hepatitis and cirrhosis."*} Please segment it. Reference answer: {*Label Name*}.
3		Long Instruction	√	
4		Function & Location	√	{*Free Text*} Please segment it. Reference answer: {*Label Name*}.
5	II	Short Instruction		{*Randomly Select One or Two In-Distribution Samples*} Please answer and segment based on the above description.
6		Long Instruction		
7	III	Function & Location		{*Free Text*} Can you answer and segment it based on the above description ?
8		Function		
9		Symptoms		

While conventional segmentation paradigms primarily evaluate performance using predefined anatomical labels (**Table 1**, Exp. 1), reasoning-based methods extend this capability by deriving targets from descriptive instructions [16]. To assess model

readiness for clinical practice, we introduce a three-tier progressive benchmark comprising nine instruction variants of increasing complexity: (Level I) reference-appended queries, where a ground-truth answer accompanies the description; (Level II) target-reasoning prompts, which require the model to deduce the target from the description alone; and (Level III) free-text instructions, characterized by high linguistic variability and inherent ambiguity typical of real-world clinical discourse.

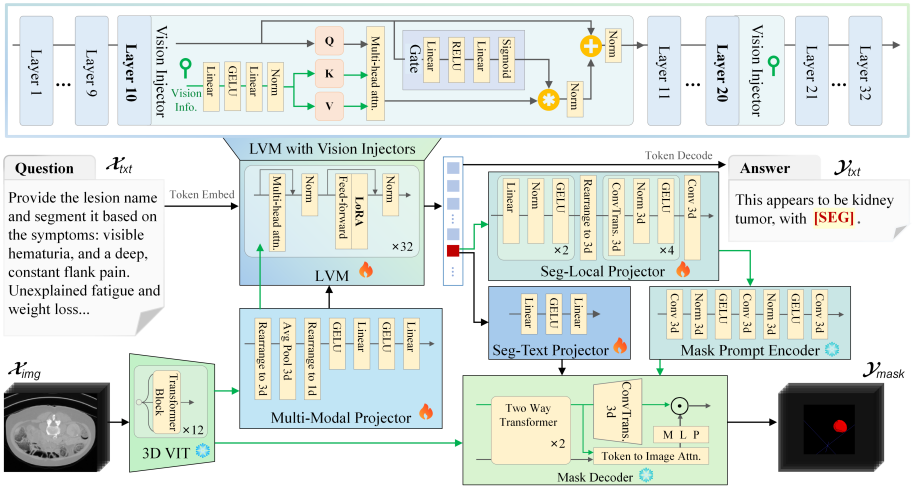

Fig. 3. Architecture of MRSeg3D. Green arrows highlight pathways carrying visual Info.

Leveraging these observations, we introduce MRSeg3D, a single-encoder architecture that enables tighter coupling of visual evidence within the reasoning-segmentation workflow. Our framework centers on the strategic integration of multi-level visual features across both [SEG] token generation and mask decoding stages to ensure consistency under instruction perturbation. To achieve this, a suite of two cascaded vision injectors provides continuous grounding, anchoring the model to visual cues during the interpretation of ambiguous instructions to prevent reasoning drift. Complementing this, a dedicated Seg-Local Projector operates in tandem with a Mask Prompt Encoder to enrich text-visual cues, substantially enhancing the precision of the final mask.

As shown in **Fig. 3**, the architecture of MRSeg3D comprises three frozen modules inherited from SAM—pre-trained during the initial stage—and four trainable components. Specifically, the frozen 3D ViT encoder, denoted as f_{enc}, extracts dense volumetric features y'_{img} from the input CT scan x_{img}. These features are subsequently mapped into a unified latent space via the multimodal projector f_{mp}. The reasoning core, f_{lvm}, is fine-tuned using Low-Rank Adaptation (LoRA) and further augmented by two injectors, f_{inj}. These injectors are positioned at the intermediate layers10 and 20, as this placement enhances visual information without interfering with the LVM's generation.

The LVM integrates visual features f_{mp} (f_{enc} (x_{img})) with textual descriptions x_{txt} to generate an answer y_{txt}, which may include a special [SEG] token h_{seg} representing the segmentation target (e.g., "kidney tumor"). The Seg-Text Projector f_{stp} and Seg-Local

Projector f_{slp} specifically extract text and visual semantics from the [SEG] token and subsequently feed them to the frozen Mask Prompt Encoder f_{mpe} and Mask Decoder f_{dec}. Finally, the mask decoder generates the 3D segmentation mask by combining dense visual features with textual and visual prompts. $L_{Dice+CE}$ denotes the sum of the cross-entropy loss L_{CE} and Dice loss L_{Dice}. This process can be formalized as:

$$[y_{txt},\ h_{seg}] = f_{lvm}\left(x_{txt}, f_{mp}\left(f_{enc}\left(x_{img}\right)\right) \mid f_{inj}\right) \tag{1}$$

$$y_{mask} = f_{dec}\left(f_{enc}\left(x_{img}\right), f_{stp}\left(h_{seg}\right), f_{mpe}\left(f_{slp}\left(h_{seg}\right)\right)\right) \tag{2}$$

$$Loss = L_{CE}\left(y_{txt},\ y_{gt}\right) + L_{Dice+CE}\left(y_{mask},\ y_{gt}\right) \tag{3}$$

3 Experiments and Results

3.1 Implementation Details.

Our experiments were conducted on 25 publicly available datasets, comprising about 6,000 3D CT volumes (150k pixel-level annotations and 120k image-text pairs [17]). All volumes were resampled to 2.0 × 1.0 × 1.0 mm³ voxel spacing, and paired text was generated using DeepSeek-V3-671B [18]. During training, we used random crops of 32 × 256 × 256 voxels and a 2:1 foreground-to-background sampling ratio to mitigate class imbalance. The instructions used for model training include the templates numbered 1 and 5 as presented in **Table 1**. The proposed framework was implemented in PyTorch [19]. The four-stage training of MSeg3D is detailed in **Table 2**, which specifies the components trained in each stage. Using DeepSpeed-accelerated mixed-precision (bfloat16) training, the process took approximately four days on 4 NVIDIA A40 GPUs.

Table 2. Four stages for training MSeg3D.

Stage	Epoch	Pretrain	Trainable Block	Supervision
1	100	3D SAM	ViT, Prompt Encoder, Decoder	Mask annotations
2	1	LLM	Multi-modal Projector	Text answers (Caption/QA)
3	6	LLM	LoRA Layers, Projectors	Text answers with [SEG] and mask annotations
4	1	LLM	LoRA Layers, Vision Injectors	

3.2 Benchmarking of 3D Reasoning Segmentation Architectures

We evaluate three mainstream 3D R-Seg models (**Fig. 4**) on 25 datasets under the standard protocol (Exp. 5). Reasoning quality is assessed using BLEU and ROUGE-2 scores, and segmentation accuracy is measured by the Dice coefficient (**Table 3**).

Method (a), M3D-LaMed, adapts LISA [21]—a 2D natural-scene R-Seg model—to the 3D medical domain. Its suboptimal performance stems from two 3D-inherent challenges: high-dimensional CT volumes and scarce region-aligned image-text pairs, which impede effective CLIP-style representation learning [13]. Moreover, encoding entire CT volumes into global embeddings limits the model's sensitivity to localized

regions, hindering its ability to capture subtle yet diagnostically critical features [17]. The "(a) + SW" variant incorporates a sliding-window for localized processing [22]; however, it leads to a decline in both Rouge and Dice scores. This suggests that in the absence of sufficient region-aligned 3D data, introducing localized modules in isolation fails to rectify—and may even exacerbate—the underlying semantic misalignment.

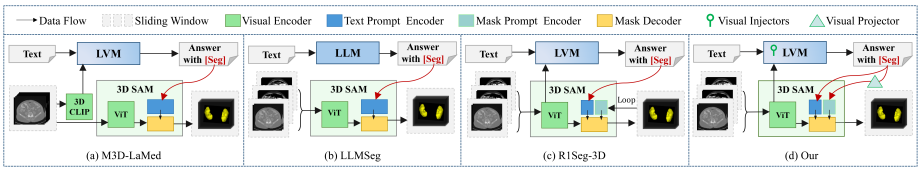

Fig. 4. Comparative illustration of three reasoning segmentation frameworks: (a) M3D-LaMed [17] uses dual vision encoders; (b) LLMSeg [20] and (c) R1Seg-3D [13] are based on a single 3D encoder architecture.

Method (b) exhibits a marked drop in textual accuracy on the TS dataset, primarily due to its unimodal design: solely within the linguistic domain to produce a [SEG] token. This overlook from visual context during reasoning evidently weakened the model's ability to (1) precisely identify the described target and (2) verify its presence within the scan, underscoring the necessity of integrated visual-textual reasoning.

Method (c), R1Seg-3D, achieves superior performance by utilizing a unified single-visual-encoder architecture. This design ensures tight semantic-visual alignment in the latent space, resulting in balanced performance across both tasks.

Our MRSeg3D demonstrates exceptional textual reasoning, particularly when integrated with LLaMA [23], and achieves the highest segmentation accuracy without the need for iterative prompting ("Dice-w/o"). Our optimized visual-aware design promotes more consistent reasoning and enhances the fidelity of visual [SEG] token generation, effectively bridging the gap between linguistic intent and spatial execution.

Table 3. Comparative performance of three R-Seg models under Experiment 5. The test set comprises 241 CT volumes (25k labels) from TotalSegmentator dataset [24] and 920 CTs (4.9k labels) from 24 other datasets. The headers "Dice-w/o" and "Dice-w/" indicate scores obtained without and with the "Loop" strategy, which iteratively feeds the initial mask back to SAM.

Method	LLM	Flops	24 Datasets				TS Dataset			
			Bleu	Rouge-2	Dice-**w/o**	Dice-**w/**	Bleu	Rouge-2	Dice-**w/o**	Dice-**w/**
(a) M3D-LaMed	Phi 3 ~4B	1.441T	11.13	49.18	70.02	69.94	16.29	55.30	62.58	62.58
(a) M3D+ SW			10.88	47.32	61.32	63.97	16.14	54.61	54.47	54.47
(b) LLM+SAM		1.411T	11.54	49.53	65.55	74.41	12.13	50.48	66.85	72.19
(c) R1Seg-3D		1.437T	12.22	50.51	68.52	**75.45**	16.78	55.79	67.51	**74.28**
Our MRSeg3D		1.416T	**12.70**	**51.97**	**72.91**	73.90	**17.01**	**56.63**	**72.65**	73.05
(c) R1Seg-3D	Qwen 2 ~7B	2.392T	13.06	51.36	68.37	**74.64**	17.03	55.19	67.36	**73.10**
Our MRSeg3D		2.399T	**17.44**	**51.40**	**74.47**	74.51	**20.94**	**57.91**	**71.99**	72.17
(c) R1Seg-3D	Llama 3 ~8B	2.537T	12.82	50.39	68.21	75.86	17.39	56.12	70.39	77.35
Our MRSeg3D		2.546T	**17.73**	**52.02**	**75.64**	**76.34**	**21.66**	**56.18**	**77.08**	**77.75**

3.3 Benchmarking under Tiered Perturbations

As visualized in **Fig. 2**, MRSeg3D consistently outperforms R1Seg-3D, exhibiting a more stable performance trajectory. Notably, the Phi-based R1Seg-3D (blue) suffers a sharp collapse between Exp. 2 and Exp. 4. Its degraded Rouge scores reveal a reasoning failure triggered by the "Reference Answer" format, which was unseen during training. While intended as helpful context, these references act as linguistic noise that distracts the LLM from key diagnostic cues. MRSeg3D (green) mitigates this by injecting stable visual cues into the reasoning process, significantly enhancing model robustness.

Table 4. The R-Seg scores across perturbation levels and different LLMs.

Method	Loop	LLM	Level **I**			Level **II**			Level **III**		
			Bleu	Rouge-L	Dice	Bleu	Rouge-L	Dice	Bleu	Rouge-L	Dice
Baseline R1Seg-3D	w/	Phi	5.85	44.10	39.45	12.94	68.97	75.59	7.60	52.92	45.40
		Qwen	14.11	66.75	69.13	15.05	69.54	75.02	13.24	55.76	48.35
		Llama	16.10	68.12	71.66	14.75	68.82	75.91	13.61	59.35	52.68
Our MRSeg3D	w/o	Phi	8.86	61.79	74.13	12.44	68.90	73.38	10.13	60.79	59.66
		Qwen	13.60	65.72	70.64	17.32	70.48	74.60	14.37	64.98	62.62
		Llama	16.73	70.02	75.03	17.94	70.23	75.61	15.44	62.65	60.88
	w/	Llama	16.78	69.41	75.74	17.89	70.26	76.34	15.24	62.55	61.55

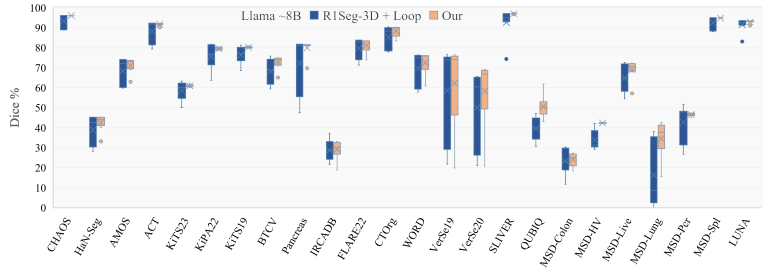

Fig. 5. Distribution of segmentation accuracy across 24 datasets in Experiments 1 to 8.

Llama-based R1Seg-3D maintains stable ROUGE scores across experiments 2~4, yet its Dice score drops at Exp. 4, indicating a fidelity gap during the inference-to-segmentation transformation. MRSeg3D addresses this bottleneck by encapsulating fused text-visual semantics to enrich the [SEG] token. Evaluations across three LLMs (**Table 4**) and 24 datasets (**Fig. 5**) demonstrate MRSeg3D's superiority without the iterative SAM loop strategy; this confirms that the Seg-Local projector effectively extracts mask-relevant information directly from the [SEG] token.

Under Level II in-distribution perturbations, all models exhibit comparable performance regardless of instruction length. However, at the more demanding Level III (free-text), our method achieves substantial gains across metrics, underscoring its superior resilience to complex linguistic perturbations. For instance, our Llama-based MRSeg3D yields margins of +1.63 BLEU, +3.20 ROUGE-L, and +8.87% in Dice.

3.4 Ablation and Visual Analysis

As demonstrate in **Table 5**, our design is well-founded: persistent visual anchoring is pivotal to mitigating reasoning drift (No. 3~4), while the dedicated mask-decoding pathway bolsters segmentation accuracy (No. 2). This complementary integration ensures robust reasoning-segmentation consistency against instruction perturbations (No. 5~6). For simple tasks, the vision injector is optional; for free-text, it is essential.

Furthermore, **Fig. 6** validates the architectural superiority of our method in generating discriminative [SEG] tokens. While R1Seg-3D suffers from feature collapse—characterized by values heavily concentrated near zero and diffused attention patterns—our network maintains a robust, high-entropy feature distribution. This enhanced representational capacity directly translates to the mask decoder's ability to produce sharper, more precisely localized activation regions on target organs.

Table 5. Ablation analysis based on Phi-based MRSeg3D. All Dice results are evaluated without the “Loop” strategy. "3 Inj.": with injection at layers 10, 20 and 26. "2 Inj.": at layers 10 and 20.

No.	Seg-Local Proj.	Vision Inj.	Mean	Level **I**				Level **II**		Level **III**		
				Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5	Exp. 6	Exp. 7	Exp. 8	Exp. 9
1	-	-	46.20	68.84	10.61	5.25	14.83	68.52	68.61	63.59	64.44	27.50
2	√	-	44.69	**75.69**	6.80	1.75	3.99	**74.92**	**75.13**	66.41	**68.07**	15.15
3	-	√ (3 Inj.)	59.77	61.33	61.66	63.11	61.61	62.07	62.70	67.13	57.26	23.60
4	-	√ (2 Inj.)	62.10	66.08	68.17	67.62	55.60	66.48	67.16	60.85	62.98	26.87
5	√	√ (3 Inj.)	69.68	74.89	**74.59**	74.03	73.02	72.91	73.86	65.87	66.51	34.37
6	√	√ (2 Inj.)	**69.98**	72.32	74.03	**74.10**	**73.83**	72.57	73.06	**67.95**	66.83	**41.23**

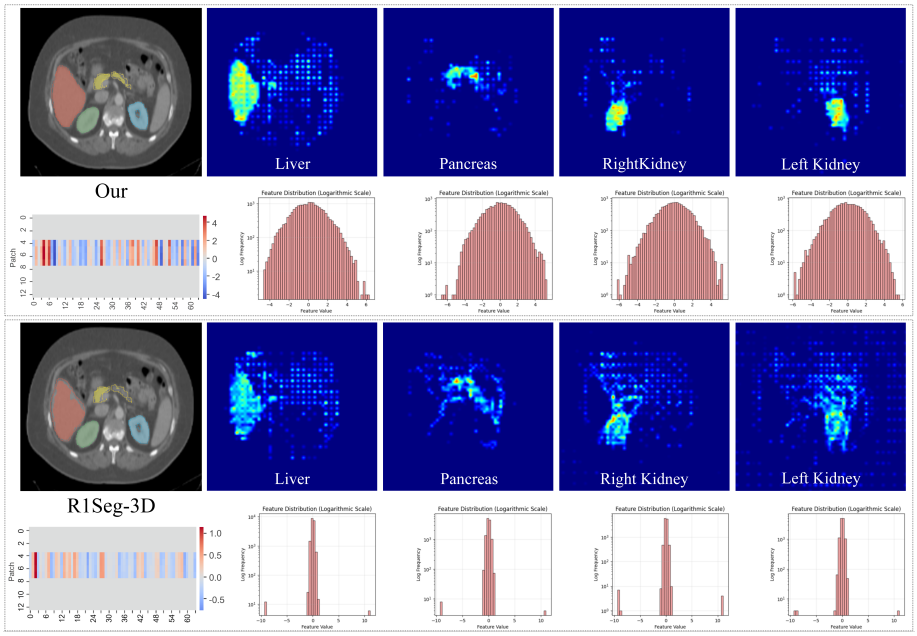

Fig. 6. Qualitative comparison and [SEG] Token feature analysis. For visualization, a 64-dimensional feature segment is randomly sampled to demonstrate inter-patch diversity. Our network (top) exhibits a broader, near-Gaussian distribution with significantly higher representational diversity compared to baseline (bottom), resulting in more precisely localized attention heatmaps.

4 Discussion and Conclusion

Our study establishes that reasoning-based medical segmentation systems are vulnerable to realistic linguistic variation, and we alleviate this by introducing both a diagnostic benchmark and a robust model, MRSeg3D. While MRSeg3D advances reasoning-segmentation consistency via persistent visual grounding, handling the full linguistic complexity of clinical discourse remains an open challenge. Future work will pivot towards pathology-specific robustness, leveraging fine-grained datasets with nuanced diagnostic context to pursue expert-level clinical decision-making.

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. We are grateful for resources from the Computing and Data Center of Xinjiang University.

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. Oda, M.: Generative AI and foundation models in medical image. Radiological Physics and Technology 18(4), 937-948 (2025)
2. Wang, W., Ma, Z., Ding, M., et al: Medical Reasoning in the Era of LLMs: A Systematic Review of Enhancement Techniques and Applications (2025)
3. Radford, A., Kim, J.W., Hallacy, C., et al: Learning Transferable Visual Models From Natural Language Supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML), pp. 8748-8763. PMLR, Virtual Event (2021)
4. Koleilat, T., Concordia.ca, Asgariandehkordi, H., et al: MedCLIP-SAM: Bridging Text and Image Towards Universal Medical Image Segmentation. In: Medical Image Computing and Computer Assisted Intervention (MICCAI), pp. 1-10. Springer Nature Switzerland, Marrakesh, Morocco (2025)
5. Ma, J., He, Y., Li, F., et al: Segment anything in medical images. Nature Communications 15(1), 654 (2024)
6. Bui, N., Hoang, D., Tran, M., et al: SAM3D: Segment Anything Model in Volumetric Medical Images. In: 2024 IEEE International Symposium on Biomedical Imaging (ISBI), pp. 1-9. IEEE, Athens, Greece (2024)
7. Du, Y., Bai, F., Huang, T., et al: SegVol: Universal and Interactive Volumetric Medical Image Segmentation. In: Advances in Neural Information Processing Systems 37 (NeurIPS), pp. 110746-110783. Curran Associates, Inc., Vancouver, Canada (2025)

8. Song, Xinrui, Xu, et al: DINO-Reg: General Purpose Image Encoder for Training-Free Multi-modal Deformable Medical Image Registration. In: Medical Image Computing and Computer Assisted Intervention (MICCAI), pp. 608-617. Springer Nature Switzerland, Marrakesh, Morocco (2025)
9. Xu, T., Hosseini, S., Anderson, C., et al: A generalizable 3D framework and model for self-supervised learning in medical imaging. npj Digital Medicine 8(1), 639 (2025)
10. Wu, J., Wang, Y., Zhong, Z., et al: Vision-language foundation model for 3D medical imaging. npj Artificial Intelligence 1(1), 17 (2025)
11. Kao, J., Kao, H.: Large Language Models in radiology: A technical and clinical perspective. European Journal of Radiology Artificial Intelligence 2, 100021 (2025)
12. Wang, H., Guo, S., Ye, J., et al: SAM-Med3D: A Vision Foundation Model for General-Purpose Segmentation on Volumetric Medical Images. IEEE Transactions on Neural Networks and Learning Systems 36(10), 17599-17612 (2025)
13. Hao, Qin, Yu, et al: R1Seg-3D: Rethinking Reasoning Segmentation for Medical 3D CTs. In: Medical Image Computing and Computer Assisted Intervention (MICCAI), pp. 415-425. Springer Nature Switzerland, Daejeon, Korea (2026)
14. Abdin, M., Aneja, J., Awadalla, H., et al: Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. arXiv preprint arXiv:2404.14219, 1-24 (2024)
15. Mansoor, I., Abdullah, M., Rizwan, M.D., et al: Reasoning with large language models in medicine: a systematic review of techniques, challenges and clinical integration. Health Information Science and Systems 14(1), 2047-2501 (2025)
16. Rokuss, M., Langenberg, M., Kirchhoff, Y., et al: VoxTell: Free-Text Promptable Universal 3D Medical Image Segmentation. arXiv preprint arXiv:2511.11450, 1-22 (2025)
17. Bai, F., Du, Y., Huang, T., et al: M3D: Advancing 3D Medical Image Analysis with Multi-Modal Large Language Models. arXiv preprint arXiv:240400578, 1-35 (2024)
18. Deepseek-ai, Liu, A., Feng, B., et al: DeepSeek-V3 Technical Report. arXiv preprint arXiv:2412.19437, 1-53 (2025)
19. Paszke, A., Gross, S., Massa, F., et al: PyTorch: An Imperative Style, High-Performance-Deep Learning Library. Advances in Neural Information Processing Systems 32 (2019)
20. Oh, Y., Park, S., Byun, H.K., et al: LLM-driven multimodal target volume contouring in radiation oncology. Nature Communications 15(1), 9186 (2024)
21. Lai, X., Tian, Z., Chen, Y., et al: 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)
22. Hatamizadeh, A., Tang, Y., Nath, V., et al: UNETR: Transformers for 3D Medical Image Segmentation. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 1748–1758. IEEE, Waikoloa, HI, USA (2022)
23. Ai@meta: Llama 3 Model Card. arXiv preprint arXiv:2407.21783, 1-92 (2024)
24. Wasserthal, J., Breit, H., Meyer, M.T., et al: TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. Radiology: Artificial Intelligence 5(5), e230024 (2023)