MedFuse-Seg: Multi-Level Visual and Semantic Context Fusion for Segmentation-Based Medical Reasoning

Keetawan Limaroon[1], Monrada Chiewhawan[2,3], Watcharapong Timklaypachara[2,3], Peerapon Vateekul[4⋆], and Titipat Achakulvisut[2⋆]

[1] Department of Computer Engineering, Faculty of Engineering, King Mongkut's University of Technology Thonburi, Bangkok, Thailand
[2] Department of Biomedical Engineering, Faculty of Engineering, Mahidol University, Nakhon Pathom, Thailand
[3] Faculty of Medicine Ramathibodi Hospital, Mahidol University, Bangkok, Thailand
[4] Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Bangkok, Thailand
titipat.ach@mahidol.ac.th, peerapon.v@chula.ac.th

Abstract. Language-driven medical image segmentation is critical for clinical workflows, yet existing models struggle with linguistic variability and complex diagnostic queries. These architectures often process semantic and spatial information in isolation, making it difficult to merge abstract clinical reasoning with precise anatomical boundaries. To address these limitations, we introduce **MedFuse-Seg**, an architecture that combines complex reasoning via Q&A generation with multi-level visual feature injection for prompt-based medical image segmentation. First, we construct **Med-ReasonSeg**, a large-scale reasoning segmentation dataset consisting of 539,383 image-mask-Q&A triplets and encompassing 9 modalities from 16 datasets. We generate referring and semantic Q&A pairs that are post-processed with a hierarchical LLM-based pipeline to ensure logical fidelity and reduce hallucinations. Second, we propose a **Multi-Level Context Fusion** mechanism that injects hierarchical features from MedSigLIP (L6-L24) directly into the MedSAM encoder, while leveraging MedGemma reasoning to guide mask decoding. MedFuse-Seg outperforms zero-shot BiomedParse and the fine-tuned LISA-7B by 13.49% and 4.89% in Dice Similarity Coefficient (DSC), and reduces 95th percentile Hausdorff Distance (HD95) by 54.04 and 15.29 pixels, respectively. These results demonstrate that combining reasoning-augmented training data with multi-level feature injection can address semantic and spatial gaps in medical segmentation. Our code and dataset are publicly available at `https://github.com/biodatlab/medfuse-seg`.

Keywords: Medical Image Segmentation · Multimodal Large Language Model · Reasoning Segmentation · Large-scale Dataset Construction.

⋆ Corresponding authors.

1 Introduction

Medical image segmentation is essential for clinical workflows, from diagnosis to treatment planning [12]. While traditional models like U-Net [23,10] established a strong baseline, they are limited to a fixed set of categories. To address this, foundation models like MedSAM [17,3,11] introduced zero-shot segmentation using spatial prompts (such as points or boxes). However, this approach remains inefficient, as it requires clinicians to provide precise location markers manually and often fails to capture irregular lesions due to fixed geometric shapes [5]. Consequently, the field is shifting toward language-driven frameworks that ground segmentation directly in diagnostic descriptions.

Despite their promise, current frameworks face distinct challenges. Generalist foundation models ranging from detectors like Grounding DINO [14] to text-promptable segmenters like SAM 3 [2] excel in natural images but struggle with medical anatomy [17]. Conversely, specialist models like BiomedParse [28] are hindered by limited data diversity. These methods rely on rigid templates, using large language models (LLMs) only for synonym generation in the training set. Coupled with the shallow reasoning of standard text encoders [25], these models are highly sensitive to how prompts are phrased. This results in poor performance when faced with the varied language used in real-world clinical queries [22].

To bridge the reasoning gap in segmentation, recent works, like LISA [13,27], adapt multimodal large language models (MLLMs) to compress visual data into a single segmentation token. However, these architectures often process understanding and localizing in isolation, losing fine-grained boundary details during compression [20,16]. We hypothesize that complex segmentation queries require models that not only reason effectively but also perceive with expert-level depth. While standard pipelines rely on final-layer outputs that prioritize abstract meaning over spatial precision, we leverage multi-level visual features to preserve the fine-grained anatomical details often lost during alignment [1].

We introduce **MedFuse-Seg**, an architecture integrating complex reasoning with multi-level visual features for medical segmentation. We construct **Med-ReasonSeg**, a large-scale reasoning dataset of 539,383 image-mask-Q&A triplets from 16 datasets, refined via LLM verification to ensure logical accuracy and minimize hallucinations. To connect semantic and spatial context, we introduce a **Multi-Level Context Fusion** mechanism that injects detailed features from MedSigLIP into the MedSAM encoder. Our model improves prompt-based segmentation by 13.49% and 4.89% in DSC compared to zero-shot BiomedParse and fine-tuned LISA-7B baselines, respectively. These contributions establish a more practical and resilient framework for language-driven medical image analysis.

2 Med-ReasonSeg Dataset

Robust language-driven segmentation models require contextual reasoning beyond simple entity referral. While BiomedParse established a unified multi-modality parsing standard, its prompts rely on rule-based templates with GPT-4

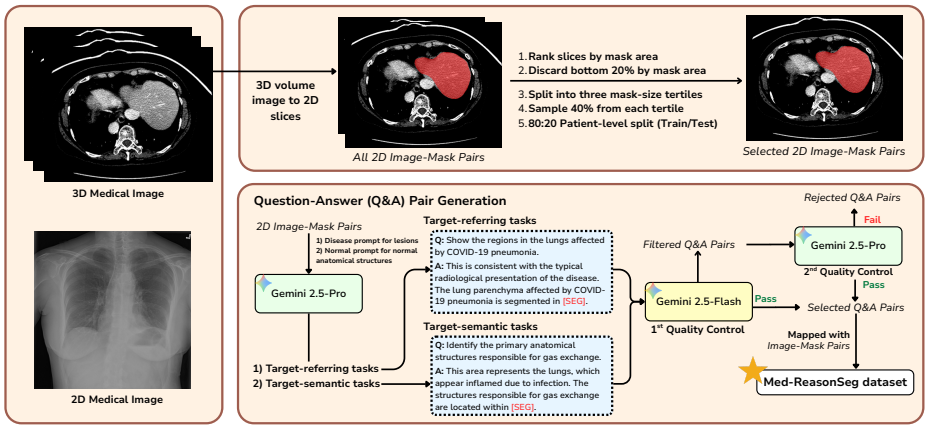

Fig. 1. Overview of the Med-ReasonSeg dataset construction pipeline. The process involves: (i) Data Curation, employing slice sampling from 3D volumes based on mask size; (ii) Instruction Generation, synthesizing target-referring and target-semantic prompts via Gemini 2.5-Pro; and (iii) LLM-Based Q&A Verification, a two-stage text quality control using Gemini 2.5-Flash and Pro.

used only for lexical rephrasing [28]. To address this, we constructed **Med-ReasonSeg**, aggregating 16 open-access datasets primarily curated from the BiomedParse collection [28] across 9 imaging modalities. The final dataset comprises 539,383 image–mask–Q&A triplets derived from 90,021 distinct scans. MRI (48.6%) and CT (27.3%) constitute the majority, followed by X-ray (9.9%), with the remaining 14.2% distributed across dermoscopy, fundus, endoscopy, OCT, mammography, and ultrasound. The construction process involves three key stages: (i) Data Curation, (ii) Question-and-Answer (Q&A) Pair Generation, and (iii) LLM-Based Q&A Verification (Fig. 1).

2.1 Data Curation

We adopted the data distribution protocol from BiomedParse [28] to align with established benchmarks. For 3D volumes, we processed them as 2D slices, discarding the bottom 20% based on mask area rankings to exclude uninformative samples with negligible anatomy. We then stratified the remaining slices into three mask-size tertiles and sampled 40% from each to mitigate inter-slice redundancy while balancing training efficiency. To strictly prevent data leakage, all data were split into training and test sets (80:20) by patient ID, ensuring that all slices from the same patient remained in the same split.

2.2 Question-and-Answer (Q&A) Pair Generation

We leveraged Multimodal Large Language Models (MLLMs) to generate Q&A pairs for model training and evaluation. We designed two task types adapted

for biomedical contexts: **(i) Target-referring tasks** directly name the segmentation target with linguistic variations (e.g., Q: "Show the regions affected by COVID-19", A: "This pattern is consistent with viral pneumonia. The affected lung parenchyma is segmented in `[SEG]`."). This task evaluates robustness to variation and ensures coverage of synonyms, abbreviations, and multi-word expressions for anatomical structures or diseases. **(ii) Target-semantic tasks** describe targets through their functional, spatial, or visual properties without explicit naming (e.g., Q: "Identify the structures for gas exchange", A: "This area represents the lungs, which appear inflamed. These structures are located within `[SEG]`."). This task type allows the model to learn implicit reasoning. By jointly training on both tasks, we aim to align explicit and implicit textual prompts with segmentation targets across diverse clinical inputs.

To generate these pairs, we employed Gemini 2.5-Pro, selected for its superior multimodal reasoning [4], using both images and their corresponding masks as inputs. This visual grounding allowed us to synthesize six distinct instructions per unique mask (three referring, three semantic) to enhance generalization, scaling the dataset to 541,752 pairs (429,798 training and 111,954 testing).

2.3 LLM-Based Q&A Verification

To ensure dataset integrity, we verified 541,752 pairs via a two-stage pipeline. Following a rule-based check to exclude corrupted images, null masks, or malformed responses missing the `[SEG]` token, we narrowed the pool to 541,662 candidates and employed a hierarchical filtering strategy. First, Gemini 2.5-Flash served as an efficient screener to process the extensive dataset, flagging 5,186 pairs for potential logical inconsistencies or hallucinations. Subsequently, leveraging its advanced reasoning capabilities, Gemini 2.5-Pro acted as a specialist to re-evaluate these flagged instances. This expert review recovered 2,907 over-penalized samples, while discarding 2,279 noisy pairs. This rigorous process yielded a final high-quality dataset of 539,383 pairs (427,861 training and 111,522 testing), each comprising the question, answer, modality, and target organ.

3 MedFuse-Seg Architecture

MedFuse-Seg architecture integrates three foundation models including MedSigLIP [24], MedGemma-4B [24] and MedSAM [17] through two core mechanisms: (i) Multi-Level Context Fusion and (ii) Reasoning-Guided Mask Decoding (Fig. 2). First, for reasoning-enhanced text encoding, we utilize MedSigLIP's visual features inherently coupled with MedGemma's language backbone (Based on Gemma 3 [6]) and fine-tune this architecture to generate reasoning tokens. This enables the model to interpret clinical language by grounding text prompts in image content. Second, for spatial segmentation, we introduce a Multi-Level Context Fusion Block that injects MedSigLIP's hierarchical features directly into MedSAM's encoder at multiple scales. Our design maximizes the utility of specialized medical representations while maintaining parameter efficiency through shared visual encoding.

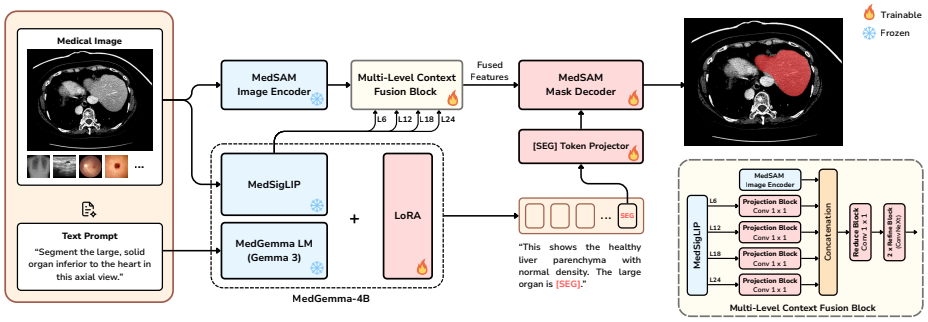

Fig. 2. Overview of the MedFuse-Seg architecture. The proposed framework integrates three foundation models through two core mechanisms: (i) Multi-Level Context Fusion for integrating MedSigLIP's medical knowledge with MedSAM's spatial features, and (ii) Reasoning-Guided Mask Decoding via MedGemma-4B for reasoning-enhanced segmentation.

3.1 Multi-Level Context Fusion

Recent work shows that intermediate vision encoder layers retain spatial details often lost in final outputs [10], aligning with hierarchical medical segmentation architectures like UNETR [8] and Swin UNETR [7]. We translate this into a hierarchical fusion strategy (Fig. 2) that integrates MedSigLIP semantics with MedSAM spatial priors. We extract multi-level features from MedSigLIP layers $L6$, $L12$, $L18$, and $L24$, where early layers capture low-level details (edges, shapes) and deeper layers encode high-level medical semantics. Each feature is projected via 1×1 *convolution* to align dimensions, then concatenated with MedSAM embeddings to create a unified representation. A Reduce Block (1×1 *convolution*) synthesizes these concatenated features through cross-channel interaction, combining semantic and spatial cues while maintaining compatibility with the MedSAM encoder. Finally, a Refine Block utilizing two cascaded large-kernel ConvNeXt blocks [26] captures long-range dependencies to refine feature representations before mask decoding. This strategy mitigates semantic ambiguity by injecting MedSigLIP's medical priors into MedSAM's boundary delineation, establishing a robust foundation for reasoning-guided decoding.

3.2 Reasoning-Guided Mask Decoding

While the fusion block provides rich visual features, the segmentation target is defined by the LLM. Following LISA [13], we add a special `[SEG]` token to the vocabulary. When the model identifies a target region, it generates this token, whose hidden state is projected as the prompt embedding. The MedSAM mask decoder then receives the refined image features and reasoning prompt to generate the final mask. By conditioning on the LLM's output, MedFuse-Seg aligns clinical reasoning to segmentation.

4 Experiments and Results

Implementation Details. We implemented our framework using PyTorch [19] and DeepSpeed [21] on 4 NVIDIA A100 GPUs. We adopted a parameter-efficient fine-tuning strategy via Low-Rank Adaptation (LoRA) [9], applying adapters ($r = 64, \alpha = 128$) to all linear layers in the MedSigLIP encoder and MedGemma model. The pre-trained backbones were frozen, while we optimized the LoRA parameters, projectors, and segmentation decoder. The model was trained for 5 epochs with a global batch size of 32 using the AdamW optimizer [15] configured with an initial learning rate of 10^{-4} and a cosine annealing scheduler with a 3% linear warmup. To maximize computational throughput, we utilized mixed precision training (bfloat16) and gradient checkpointing. All fine-tuned models used identical training settings.

Data Augmentation. We applied augmentations to enhance generalization while preserving consistency between images and Q&A pairs, such as directional terms (e.g., "left", "top"). We used soft geometric transformations ($p = 0.5$) with limited ranges: rotation ($\pm 15°$), shift (5%), and scale (20%). We added Gaussian blur and noise ($p = 0.15$), and brightness/contrast adjustments ($p = 0.5$).

Loss Function and Evaluation Metrics. We minimized a joint objective for text generation and mask segmentation: $\mathcal{L} = \mathcal{L}_{text} + \mathcal{L}_{mask}$. For text generation, we employed standard auto-regressive cross-entropy ($\mathcal{L}_{text}$) to ensure semantic consistency. For segmentation, we combined Focal ($\gamma = 2, \alpha = 0.25$) and Dice losses to address class imbalance while maximizing structural precision: $\mathcal{L}_{mask} = 4\mathcal{L}_{Focal} + 2\mathcal{L}_{Dice}$. For evaluation on the Med-ReasonSeg test set, we assessed performance using the Dice Similarity Coefficient (DSC) to measure region overlap and the 95th percentile Hausdorff Distance (HD95) to capture boundary adherence and shape discrepancies [18]. All metrics are computed at the model's standard resolution (1024^2 pixels) and reported as macro-averages.

4.1 Main Results

We benchmark MedFuse-Seg against state-of-the-art approaches in Table 1. The generalist foundation model (SAM 3) and the hybrid approach (G-DINO + MedSAM) struggle significantly in the zero-shot setting (DSC < 0.15), likely due to the substantial domain gap. While the medical specialist BiomedParse shows improved generalization, it lags behind our method, particularly in semantic tasks. Overall, MedFuse-Seg outperforms zero-shot BiomedParse by 13.4% in DSC and 54.04 pixels in HD95. Notably, even when the LISA-7B baseline is fully fine-tuned on our dataset, our model maintains superior margins of 4.89% in DSC and 15.29 pixels in HD95. These gains show our Multi-Level Context Fusion effectively integrates visual and textual features, outperforming bottleneck-only fusion in existing reasoning architectures.

Qualitative analysis on the test set shows strong model performance in both segmentation and text output reasoning. The model can handle both descriptive (Fig. 3A and 3B) and direct segmentation requests (Fig. 3C). Moreover,

Table 1. Main Benchmarking Results. Performance comparison on Target-referring and Target-semantic tasks across various architectures. Metrics: DSC (↑) and HD95 (↓, px). Experimental settings are tailored to each model's design: *Oracle* (MedSAM w/ GT boxes), *Zero-shot* for foundation models (SAM 3, G-DINO + MedSAM, BiomedParse) to evaluate generalization, and *Fine-tuning* for LISA-7B to ensure a competitive reasoning baseline. Best results are in **bold**, second best are underlined.

		Referring		Semantic		Average	
Method	**Type**	**DSC ↑**	**HD95 ↓**	**DSC ↑**	**HD95 ↓**	**DSC ↑**	**HD95 ↓**
Traditional / Oracle							
MedSAM (w/ GT Box) [17]	Specialist	-	-	-	-	0.6329	58.41
Foundation / Zero-shot							
SAM 3 [2]	Foundation	0.1425	373.52	0.1167	370.50	0.1296	372.01
BiomedParse [28]	Specialist	0.6703	105.23	0.6344	115.97	0.6524	110.60
Detection-Guided Models							
G-DINO [14] + MedSAM [17]	Hybrid	0.1424	494.06	0.1461	488.56	0.1442	491.31
Reasoning Segmentation							
LISA-7B (ft) [13]	Generalist	0.7398	71.55	0.7370	72.13	0.7384	71.84
MedFuse-Seg (Ours)	**Specialist**	**0.7879**	**56.46**	**0.7867**	**56.65**	**0.7873**	**56.55**

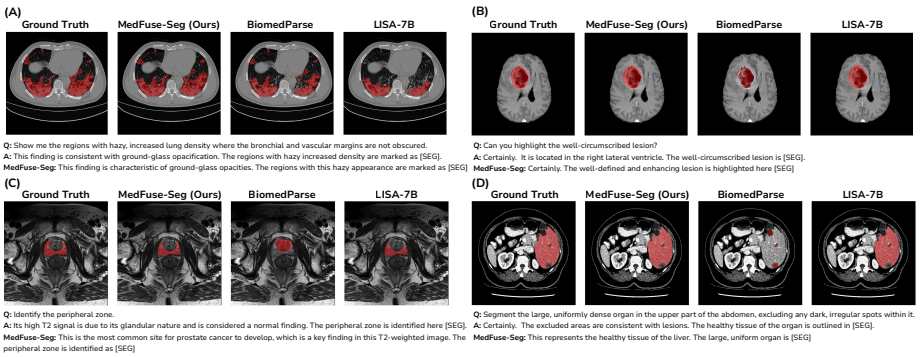

Fig. 3. Sample segmentation results of MedFuse-Seg, BiomedParse, and LISA-7B against ground truth, with queries, key answers, and MedFuse-Seg's responses.

the model can perform complex segmentation and incorporate clinical disease knowledge into its responses (Fig. 3C and 3D). Preliminary expert review of these outputs suggests overall clinical soundness. However, performance remains limited for some underrepresented irregular lesions. Incorporating more diverse, disease-focused training data may further improve performance on these challenging cases.

4.2 Ablation Studies

Impact of Visual Feature Fusion Strategy. We observe that low-level features ($L6$, $L12$) contribute more to DSC from 0.7373 to 0.7865, whereas high-level features ($L18$, $L24$) contribute to HD95 from 69.74 to 56.55. Incorporating both

Table 2. Comprehensive MedFuse-Seg Ablation Study. The default configuration uses MedSAM (Frozen), Multi-Level Fusion, Verified Data, and End-token positioning. The best results are in **bold**, second best are underlined.

Component	Variant / Setting	DSC ↑	HD95 ↓
Feature Fusion	Baseline (LISA-Style)	0.7373	69.74
	Spatial Focus ($L6, L12$)	0.7865	58.96
	Semantic Focus ($L18, L24$)	0.7786	58.81
	Multi-Level ($L6, L12, L18, L24$) (Ours)	**0.7873**	**56.55**
Visual Backbone	SAM (Unfrozen)	0.7090	98.27
	SAM (Frozen)	0.7740	70.24
	MedSAM (Unfrozen)	0.7844	57.22
	MedSAM (Frozen) (Ours)	**0.7873**	**56.55**
Data Verification	Raw / Unverified	0.7853	59.07
	LLM-Verified (Ours)	**0.7873**	**56.55**
[SEG] Position	Start	0.7851	57.63
	End (Ours)	**0.7873**	**56.55**

yields the highest performance, achieving DSC and HD95 of 0.7873 and 56.55, respectively (Table 2). This demonstrates that the fusion mechanism successfully captures intricate boundary precision in complex anatomical structures, rather than relying solely on text reasoning.

Impact of Visual Backbone and Training Strategy. MedSAM outperforms the generalist SAM, likely due to pretraining on large-scale medical images. Keeping the MedSAM frozen achieves higher performance (DSC = 0.7873) than full fine-tuning (DSC = 0.7844), indicating that preserving pretrained domain-specific representations is crucial for instruction-tuning stability.

Impact of LLM-Based Verification. Models trained on our LLM-verified dataset achieve superior segmentation stability over raw synthetic data (Table 2). LLM-based verification significantly improves DSC from 0.7853 to 0.7873 ($t = 6.91$, $p < 0.001$) and HD95 from 59.07 to 56.55 ($t = 8.99$, $p < 0.001$), confirming data quality is critical for minimizing noise and enhancing robustness.

Impact of [SEG] Token Positioning. Our experiment shows that appending `[SEG]` at the end (e.g., `description... [SEG]`) reduces HD95 from 57.63 to 56.55 ($t = 9.89, p < 0.001$) compared to placing it at the start (e.g., `[SEG] description...`). This demonstrates that leveraging the autoregressive nature of MLLMs in a "think-before-act" style may have an effect on the outputs that should be further explored.

5 Conclusion

MedFuse-Seg bridges clinical reasoning and precise segmentation by injecting hierarchical visual features into MLLMs, outperforming state-of-the-art models on both DSC and HD95. With our large-scale Med-ReasonSeg dataset, the model establishes a robust, medically-aware foundation for language-driven medical image analysis. Future work should explore volumetric and video segmentation with inter-slice consistency, comparisons with a broader range of medical reasoning models, and policy optimization for refining clinical reasoning.

Acknowledgments. This research has received funding support from the National Science, Research and Innovation Fund (NSRF) via the Program Management Unit for Human Resources & Institutional Development, Research and Innovation under Grant B13F680099.

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

References

1. Bolya, D., Huang, P.Y., Sun, P., Cho, J.H., Madotto, A., Wei, C., Ma, T., Zhi, J., Rajasegaran, J., Rasheed, H.A., Wang, J., Monteiro, M., Xu, H., Dong, S., Ravi, N., Li, S.W., Dollar, P., Feichtenhofer, C.: Perception Encoder: The best visual embeddings are not at the output of the network. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2026)
2. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: SAM 3: Segment anything with concepts. In: The Fourteenth International Conference on Learning Representations (2026)
3. Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Sun, L.J.H., He, J., Zhang, S., Zhu, M., Qiao, Y.: SAM-Med2D. arXiv preprint arXiv:2308.16184 (2023)
4. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
5. 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 (2024)
6. Gemma Team: Gemma 3 technical report. arXiv preprint arXiv:2503.19786 (2025)
7. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D.: Swin UNETR: Swin transformers for semantic segmentation of brain tumors in mri images. In: International MICCAI brainlesion workshop. pp. 272–284. Springer (2021)
8. 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)
9. 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)

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. Kumar, S.: Advancements in medical image segmentation: A review of transformer models. Computers and Electrical Engineering **123**, 110099 (2025)
13. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9579–9589 (2024)
14. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding DINO: Marrying dino with grounded pre-training for open-set object detection. In: European conference on computer vision. pp. 38–55. Springer (2024)
15. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019)
16. Lu, Y., Cao, J., Wu, Y., Li, B., Tang, L., Ji, Y., Wu, C., Wu, J., Zhu, W.: RSVP: Reasoning segmentation via visual prompting and multi-modal chain-of-thought. In: Che, W., Nabende, J., Shutova, E., Pilehvar, M.T. (eds.) Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 14699–14716. Association for Computational Linguistics, Vienna, Austria (2025)
17. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
18. Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M.D., Buettner, F., Christodoulou, E., Glocker, B., Isensee, F., Kleesiek, J., Kozubek, M., et al.: Metrics reloaded: recommendations for image analysis validation. Nature methods **21**(2), 195–212 (2024)
19. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
20. Qian, R., Yin, X., Dou, D.: Reasoning to attend: Try to understand how <seg> token works. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24722–24731 (2025)
21. Rasley, J., Rajbhandari, S., Ruwase, O., He, Y.: DeepSpeed: System optimizations enable training deep learning models with over 100 billion parameters. In: Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 3505–3506 (2020)
22. Rokuss, M., Langenberg, M., Kirchhoff, Y., Isensee, F., Hamm, B., Ulrich, C., Regnery, S., Bauer, L., Katsigiannopulos, E., Norajitra, T., Maier-Hein, K.: VoxTell: Free-text promptable universal 3d medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 37538–37557 (2026)
23. 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)
24. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)

25. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
26. Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S.: ConvNeXt V2: Co-designing and scaling convnets with masked autoencoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16133–16142 (2023)
27. 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:2312.17240 (2023)
28. Zhao, T., Gu, Y., Yang, J., Usuyama, N., Lee, H.H., Kiblawi, S., Naumann, T., Gao, J., Crabtree, A., Abel, J., et al.: A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. Nature methods **22**(1), 166–176 (2025)