Skill-Evolving Grounded Reasoning for Free-Text Promptable 3D Medical Image Segmentation

Tongrui Zhang[1], Chenhui Wang[1], Yongming Li[1], Zhihao Chen[1], Xufeng Zhan[2], and Hongming Shan[1] (✉)[0000-0002-0604-3197]

[1] Institute of Science and Technology for Brain-inspired Intelligence; MOE Frontiers Center for Brain Science; Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence (Ministry of Education), Fudan University, Shanghai, China
[2] School of Information Science and Technology, University of Science and Technology of China, Anhui, China
hmshan@fudan.edu.cn

Abstract. Free-text promptable 3D medical image segmentation offers an intuitive and clinically flexible interaction paradigm. However, current methods are highly sensitive to linguistic variability: minor changes in phrasing can cause substantial performance degradation despite identical clinical intent. Existing approaches attempt to improve robustness through stronger vision-language fusion or larger vocabularies, yet they lack mechanisms to consistently align ambiguous free-form expressions with anatomically grounded representations. We propose **S**kill-**E**volving ground**E**d **R**easoning (**SEER**), a novel framework for free-text promptable 3D medical image segmentation that explicitly bridges linguistic variability and anatomical precision through a reasoning-driven design. First, we curate the SEER-Trace dataset, which pairs raw clinical requests with image-grounded, skill-tagged reasoning traces, establishing a reproducible benchmark. Second, SEER constructs an evidence-aligned target representation via a vision-language reasoning chain that verifies clinical intent against image-derived anatomical evidence, thereby enforcing semantic consistency before voxel-level decoding. Third, we introduce SEER-Loop, a dynamic skill-evolving strategy that distills high-reward reasoning trajectories into reusable skill artifacts and progressively integrates them into subsequent inference, enabling structured self-refinement and improved robustness to diverse linguistic expressions. Extensive experiments demonstrate superior performance of SEER over state-of-the-art baselines. Under linguistic perturbations, SEER reduces performance variance by 81.94% and improves worst-case Dice by 18.60%. Project resources are available at: https://seer-medseg.github.io.

Keywords: Free-Text Promptable Medical Image Segmentation · Vision-Language Models · Grounded Reasoning

1 Introduction

Text-guided 3D medical image segmentation [22, 24–26] allows users to specify targets via predefined class names, offering a more intuitive interface for

clinical workflows compared to traditional point- or bounding-box-based models [3,7,12,14,15]. Yet, the dominant language signal available in real-world care is not a standardized label set but free-text: unstructured, free-form language originating from imaging orders, clinical questions, or radiology reports. Free-text promptable segmentation broadens clinical applicability for diverse diagnostic scenarios and also serves as a natural interface for future AI agent integration.

However, directly conditioning segmentation models on free-text introduces severe challenges regarding linguistic variability, as clinical text frequently contains abbreviations, synonymous phrasing, and institution-specific conventions [8]. While recent free-text promptable segmentation work attempts to alleviate this through larger training corpora, stronger vision-language fusion [19], or more complex multi-agent systems [11], they largely overlook that standard natural language embedding spaces are not inherently aligned with complex clinical terminology. Clinically equivalent synonyms or shorthand frequently map to disparate embedding neighborhoods [4], resulting in drastically different conditioning signals and ultimately unstable segmentation masks. Overcoming this instability requires moving beyond direct text-to-feature mapping; it necessitates an intermediate step of explicit reasoning to disambiguate the raw text and resolve the underlying clinical intent before executing the task.

Yet, pure language reasoning is insufficient for medical imaging; it must be heavily grounded in the patient's visual evidence [21, 23]. A dependable text-guided segmentation system needs to comprehend the unstructured clinical request and align it with the anatomical evidence [9]. Rather than relying on rigid text-to-text synonym normalization, this alignment is best achieved by decomposing the vision-language reasoning into granular, modular steps, known as skills. By applying skills (e.g., synonym normalization or spatial relation reasoning) that cross-reference the textual prompt with visual features or volume renderings, the system can synthesize an explicit, evidence-aligned target representation. This provides the downstream segmentation backbone with a canonical signal, substantially mitigating instability from diverse linguistic expressions.

To this end, we propose a **s**kill-**e**volving ground**e**d **r**easoning (SEER) framework, which bridges variable free-form requests to robust 3D medical segmentation by formulating the required vision-language reasoning as explicit, executable skills. When presented with a clinical request and a 3D scan, SEER leverages these skills to generate a task specification for the segmentation backbone. Furthermore, instead of relying on a static reasoning module, we introduce a dynamic skill-evolving strategy, which continuously distills high-reward grounded reasoning episodes into auditable, reusable skill artifacts. By tracking skill usage and performance gains, SEER self-refines its reasoning capabilities, progressively enhancing its robustness against complex or unseen linguistic perturbations.

Our contributions are summarized as follows. **(i)** We introduce SEER, a novel framework that bridges diverse free-text clinical requests to 3D medical segmentation. **(ii)** We curate the SEER-Trace dataset, containing raw clinical requests with image-grounded and skill-tagged reasoning traces, to support reproducible research on free-text promptable 3D medical segmentation. **(iii)** We formalize

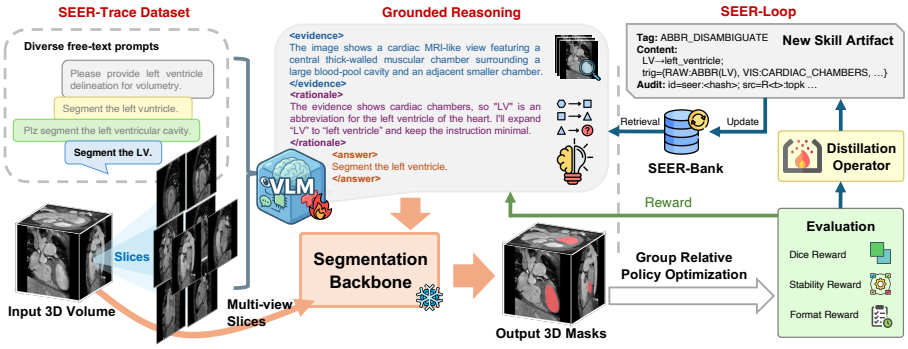

Fig. 1. Illustration of the proposed SEER framework. SEER makes free-text promptable 3D medical segmentation robust by grounding clinical language in image evidence and evolving reusable reasoning skills.

grounded visual–language reasoning as explicit skills, enabling the framework to map ambiguous free-text into explicit target representations. **(iv)** We propose a dynamic skill-evolving strategy that distills high-reward reasoning episodes into reusable skill artifacts, allowing the model to self-refine and adapt to complex linguistic variations over time. **(v)** Extensive experimental results demonstrate that SEER achieves superior segmentation performance over state-of-the-art baselines and strong transferability across segmentation backbones.

2 Methodology

Fig. 1 presents an overview of the proposed SEER framework for free-text promptable 3D medical segmentation. Building upon the supervisory foundation of *SEER-Trace*, our framework employs *Grounded Reasoning* to accurately align text with multi-view visual evidence; to overcome the limitations of static reasoning, we further introduce a dynamic skill-evolving strategy (*SEER-Loop*) to distill successful episodes into reusable skills stored in SEER-Bank, thereby establishing a synergistic system that continuously enhances robustness.

2.1 SEER-Trace Dataset Curation

SEER-Trace is built from established 3D medical segmentation benchmarks [2,6,16,17,20] and augments each case with clinician-like requests and grounded, skill-tagged traces. It essentially aligns three core elements: (i) linguistically diverse requests, (ii) evidence-aligned intermediate reasoning, and (iii) an executable task specification defined by the downstream segmentation system.

Formally, SEER-Trace consists of tuples:

$$\mathcal{D}_{\text{TRACE}} = \{(V_i,\ q_i,\ y_i,\ \tau_i)\}_{i=1}^{N}, \qquad y_i = (e_i, r_i, a_i), \tag{1}$$

where V_i is a 3D volume, q_i is a raw free-text request, and y_i is a structured reasoning response containing `<evidence>` e_i, `<rationale>` r_i, and an executable `<answer>` a_i. Specifically, each raw request q_i is generated under one of eight predefined clinical styles (e.g., radiology note and consult question), producing diverse styles and levels of underspecification. We structure the response y_i as this triplet to explicitly decouple the cognitive steps of intent resolution. Within y_i, the evidence e_i enforces perception by isolating observable visual cues from V_i, the rationale r_i enforces logical reasoning by linking the ambiguous text q_i to these visual cues, and the answer a_i formulates the final target representation. To augment y_i with modular steps, we also design the skill bank τ_i, which includes explicit skills such as spatial relation reasoning and synonym normalization.

We render each volume into a multi-view 2D representation $\mathcal{I}_i = \mathcal{R}(V_i) = \{I_{i,m}\}_{m=1}^{M}$. We then use a large language model with constrained prompting to generate q_i and a high-capacity vision-language model (VLM) conditioned on $(\mathcal{I}_i, q_i)$ to synthesize y_i and τ_i, generating intent-preserving explanations and skill tags that are strictly consistent with the volumes and requests.

To avoid hallucinations, we constrained each request q_i by target aliases, visual priors, and forbidden terms to preserve clinical intent and reduce trivial label leakage. We also checked each candidate y_i for semantic equivalence with the intended target and compatibility with the frozen backbone. Furthermore, we randomly sampled 5% of tuples for expert audit to verify clinical plausibility. In the expert audit, only 7 out of 1,120 randomly sampled instances were flagged as abnormal, yielding a 0.63% abnormal rate, which supports the reliability of the full SEER-Trace dataset. The final SEER-Trace dataset contains 22,330 multimodal instruction instances from 1,811 cases.

2.2 Grounded Reasoning

Building upon the supervised format of SEER-Trace, this component executes the core reasoning process at inference. Rather than solely leveraging a fixed mapping from text input to segmentation type, SEER actively decomposes complex input prompts into granular, executable cognitive operations, using explicit *skills* such as spatial relation reasoning and synonym normalization.

When processing a clinical request q and rendered views $\mathcal{I} = \mathcal{R}(V)$, the VLM π_θ follows the learned structured reasoning format and generates $y = (e, r, a)$:

$$\pi_\theta(y \mid \mathcal{I}, q, \mathcal{B}) = \pi_\theta(e, r, a \mid \mathcal{I}, q, \mathcal{B}), \tag{2}$$

where $\mathcal{B}$ denotes optional *skill artifacts* retrieved from SEER-Bank, supplying reusable, high-reward reasoning patterns. To achieve visually grounded reasoning, three components operate synergistically. First, the model *extracts* anatomical evidence e as a visual grounding constraint in $\mathcal{I}$. Second, it *formulates* a rationale r that logically links the ambiguous free-text request q to this explicit visual evidence, thereby disambiguating the intended target. Finally, based on the visual evidence and rationale, it *synthesizes* an executable answer a, which is executed by the frozen segmentation system $\mathcal{S}$, yielding a predicted mask $\widehat{G}$.

This structural separation prevents the model from relying on unconstrained latent text conditioning, forcing it to explicitly validate clinical intent against anatomical evidence before generating the final segmentation.

To make our framework robust against diverse linguistic expressions of the same clinical intent, we explicitly optimize the model to produce stable segmentation masks. Let $\Omega(q)$ denote a set of clinically equivalent rephrasings of q, SEER aims to improve both expected segmentation quality and stability under $\Omega(q)$ by optimizing the following objective:

$$\mathcal{J}(\theta) = \mathbb{E}_{(V,q,G)}\big[\mathbb{E}_{q'\sim\Omega(q)}\big[D_\theta(V,q',G)\big] - \lambda \mathrm{Var}_{q'\sim\Omega(q)}\big[D_\theta(V,q',G)\big]\big], \quad (3)$$

where $D_\theta(V,q',G) = \mathrm{Dice}\left(\mathcal{S}\left(V, a_\theta(V,q')\right), G\right)$, G is the ground truth segmentation mask, $a_\theta(V,q')$ is the generated `<answer>` for (V,q'), and λ controls the quality–stability trade-off.

To effectively maximize $\mathcal{J}(\theta)$, the learning process proceeds in two stages. First, supervised fine-tuning (SFT) aligns the VLM with the structural operations of SEER-Trace. Second, group relative policy optimization (GRPO) further refines the policy via a composite reward function R. By jointly rewarding the execution Dice score and stability across equivalent queries $q' \in \Omega(q)$, while using a β-weighted KL penalty to the SFT policy to preserve format compliance, GRPO enforces robust, high-fidelity generation.

2.3 Dynamic Skill-Evolving Strategy (SEER-Loop) via SEER-Bank

While grounded reasoning resolves ambiguity using existing skills, the linguistic variability in real-world clinical settings is unbounded. Therefore, we propose a dynamic skill-evolving strategy (SEER-Loop) implemented via SEER-Bank. This memory mechanism captures recurring ambiguity patterns as explicit skill artifacts, retrieves relevant skills during generation, and updates the bank over time through experience distillation and downstream feedback; see Fig. 1 for the interaction–update loop.

At evolution round t, SEER-Bank stores a set of skill artifacts, $\mathcal{B}^{(t)} = \{s_j = (\kappa_j, z_j, \eta_j)\}_{j=1}^{K_t}$, where κ_j is a skill tag, z_j is the skill content, and η_j records audit metadata. Given rendered views $\mathcal{I}$ and a request q, a retrieval operator selects a relevant subset $\mathcal{B}^{(t)}(\mathcal{I}, q) \subset \mathcal{B}^{(t)}$ and uses it to guide the model's interaction with the environment by conditioning the generation on retrieved skills:

$$\pi_\theta(y \mid \mathcal{I}, q, \mathcal{B}^{(t)}) \approx \pi_\theta\big(y \mid \mathcal{I}, q, \mathcal{B}^{(t)}(\mathcal{I}, q)\big). \quad (4)$$

Unlike generic text retrieval, retrieved skills are reusable reasoning strategies. Retrieval ranks skills using both request–metadata compatibility, such as target, tag, and alias overlap, and consistency between stored visual cues and the current multi-view evidence. The selected skills are rendered as structured `<skill_bank>` blocks and appended to the VLM input.

Skill evolution is driven by high-reward interaction episodes with the frozen environment. Each episode records the rendered views I, request q, generated

Table 1. Performance comparison under label and free-text prompting modes.

Dataset	Method	Label Prompting	Free-text Prompting		
		Dice↑	Dice ↑	Worst Dice ↑	Std. ↓
BrainMet-Share	SAT [26]	22.16	0.69	0.00	2.53
	BiomedParseV2 [25]	18.66	2.53	0.00	7.27
	Text3DSAM [22]	0.10	0.41	0.00	0.93
	MedSAM3 [11]	11.33	16.62	10.56	5.17
	VoxTell [19]	48.19	52.15	46.71	3.35
	SEER (**Ours**)	**51.70**	**53.83**	**51.44**	1.67
PENGWIN	SAT [26]	96.05	0.01	0.00	0.13
	BiomedParseV2 [25]	1.35	8.53	0.00	7.50
	Text3DSAM [22]	24.75	0.01	0.00	0.16
	MedSAM3 [11]	18.26	5.75	3.67	6.40
	VoxTell [19]	**97.59**	92.26	79.34	7.49
	SEER (**Ours**)	97.56	**97.39**	**95.47**	0.98

executable answer $y = (e, r, a)$, execution result $\widehat{G} = \mathcal{S}(V, a)$, and reward R. Let Ξ^+ denote the set of top-reward episodes. The distillation extracts successful, novel reasoning patterns from Ξ^+ into new skill artifacts. To prevent memory bloat and redundancy, we apply two utility-based operations. First, Dedup$(\cdot)$ groups skills by source, skill tag, and target, merges semantically identical or highly overlapping skills, and keeps the one with higher empirical utility. Second, Prune$(\cdot)$ further caps the bank size by removing skills with negative or low empirical utility. The bank update is written as:

$$\Delta\mathcal{B}^{(t)} = \text{Distill}(\Xi^+), \qquad \mathcal{B}^{(t+1)} = \text{Prune}\Big(\text{Dedup}\Big(\mathcal{B}^{(t)} \cup \Delta\mathcal{B}^{(t)}\Big)\Big). \tag{5}$$

The empirical utility used by both Dedup$(\cdot)$ and Prune$(\cdot)$ is estimated from each skill's marginal reward gain, which is defined as:

$$\Delta(s) = \mathbb{E}\big[R \,\big|\, s \in \mathcal{B}(\mathcal{I}, q)\big] - \mathbb{E}\big[R \,\big|\, s \notin \mathcal{B}(\mathcal{I}, q)\big]. \tag{6}$$

By tracking these metrics, SEER-Bank identifies and retains only empirically verified reasoning skills. In effect, this continuous distillation, deduplication, and pruning process stabilizes generation, allowing the framework to self-refine and adapt to complex, unseen linguistic variations over time.

3 Experiments

3.1 Experimental Setup

Datasets and prompting protocols. We evaluate SEER on two datasets to comprehensively assess generalization: BrainMetShare [5] (representing a domain shift involving seen anatomy but unseen institutional sources) and PENGWIN [13] (serving as a strictly out-of-distribution (OOD) benchmark since its

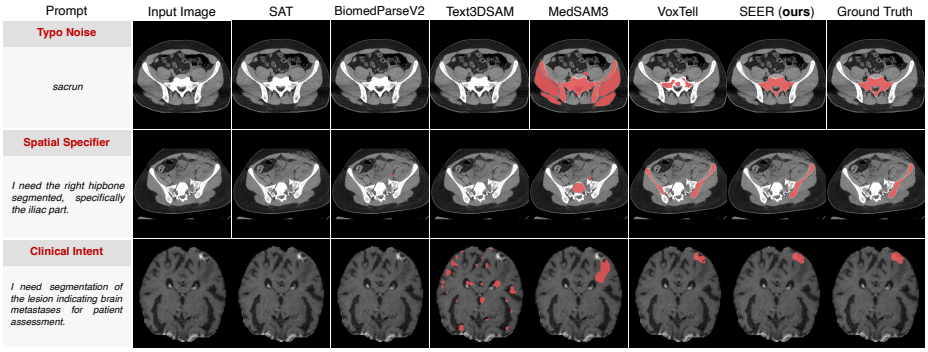

Fig. 2. Qualitative comparisons under three clinically plausible free-text prompts.

pelvic bone targets are absent from the SEER-Trace reasoning supervision and target coverage). Each volume is processed into the multi-view 2D representation introduced in SEER-Trace, paired with diverse, clinician-like free-text requests. To ensure a comprehensive and fair comparison, we evaluate all baselines under two settings: (i) their natively supported, predefined label prompting interfaces (label prompting mode), and (ii) the realistic, linguistically diverse free-text clinical requests introduced in SEER-Trace (free-text prompting mode).

Implementation details. All experiments are implemented in PyTorch [18] and trained on NVIDIA A100 GPUs. We use DeepSeek-V3.2 [10] as the high-capacity VLM during SEER-Trace construction and initialize our vision-language reasoning module with Qwen3-VL-4B-Instruct [1]. During both training stages, the visual encoder is kept frozen while the remaining parameters are fully updated. SEER-Loop via SEER-Bank is executed in discrete evolutionary rounds. Following each training phase, we distill high-reward reasoning episodes into reusable skill artifacts, deduplicate the memory bank, and inject these updated skill traces back into the subsequent training stage.

Evaluation metrics. We report Dice for 3D medical segmentation accuracy. Robustness analyses under sets of clinically equivalent rephrasings are also reported in terms of standard deviation (Std.) and worst-case Dice (worst Dice).

3.2 Comparison Results, Ablation Study, and Generalization

Quantitative comparison. Table 1 reports Dice in both label prompting and free-text prompting modes. In label prompting mode, SEER matches the strongest baseline on PENGWIN and improves on BrainMetShare. In free-text prompting mode, several prior methods collapse to near-zero Dice, reflecting poor robustness to realistic clinical phrasing. Among methods that remain functional, SEER consistently improves mean Dice while substantially reducing variance and improving worst-case Dice. The consistent gains across both partially and strictly OOD datasets demonstrate that SEER acquires a new capability—generalizing beyond seen anatomical targets by using grounded reasoning and skill-evolving

Table 2. Ablation study on the strictly OOD PENGWIN dataset.

Configuration	Dice ↑	Worst Dice ↑	Std. ↓
Baseline	92.26	79.34	7.49
+ Vanilla VLM	84.84	61.90	14.15
+ Fine-tuned VLM w/ Grounded Reasoning	95.92	88.27	3.84
+ SEER-Loop (our SEER)	**97.39**	**95.47**	**0.98**

strategy to resolve ambiguities—rather than merely normalizing long free-text into seen label tokens.

Qualitative comparison. Fig. 2 presents representative cases under three clinically plausible free-text prompt variants: *Typo Noise* (spelling perturbation with preserved intent), *Spatial Specifier* (laterality and subregion constraints), and *Clinical Intent Paraphrase* (high-level clinical phrasing referring to pathology). SEER produces consistent, on-target masks across all variants, reflecting evidence-aligned disambiguation from the image. In contrast, baseline outputs are highly prompt-sensitive in free-text prompting mode, frequently drifting to off-target regions or collapsing to degenerate predictions (e.g., empty/fragmented masks) under minor wording changes.

Ablation study. To evaluate the efficacy of our core architectural contributions, we conducted an ablation study on the strictly OOD PENGWIN dataset. We progressively integrated our proposed modules into the strongest baseline [19]. Table 2 shows that naively prepending a vanilla VLM (Qwen3-VL-4B-Instruct) to parse the text actually degrades performance. This validates our hypothesis that unconstrained, pure-text reasoning without strict anatomical grounding exacerbates instability. The fine-tuned VLM with grounded reasoning successfully rectifies this, lifting mean Dice to 95.92 and halving the standard deviation compared to the baseline. Finally, the integration of the SEER-Loop delivers the substantial improvement in robustness. By continuously retrieving empirically verified reasoning artifacts from the SEER-Bank, this full configuration achieves the highest mean accuracy, raises the worst-case Dice to 95.47, and crushes the standard deviation to a mere 0.98. This near-zero variance indicates that SEER effectively immunizes the segmentation backbone against complex linguistic perturbations.

Cross-backbone generalization. To evaluate the generalizability of SEER across different foundational architectures, we replace the downstream segmentation backbone with MedSAM3 while keeping our proposed front-end frozen. Specifically, on the strictly OOD PENGWIN dataset, integrating SEER substantially improves upon the standalone MedSAM3 baseline, increasing the mean Dice from 5.75 (std: 6.40) to 19.97 (std: 13.30) and elevating the worst-case minimum score from 3.67 to 3.94. This suggests that the reasoning capabilities of SEER are beneficial to other segmentation backbones, consistently elevating zero-shot generalization and worst-case robustness.

4 Conclusion

We presented SEER, a skill-evolving grounded reasoning framework that stabilizes free-text promptable 3D medical segmentation by converting linguistically diverse clinical requests into a single evidence-aligned, executable task specification for a frozen segmentation system. Across partially and strictly OOD benchmarks, SEER preserves strong label prompting performance while substantially improving free-text prompting robustness, reducing dispersion under linguistic variability and improving worst-case behavior; ablations further show that grounded visual reasoning and SEER-Loop contribute complementary gains. Future work should explore volumetric VLM encoders, expand the skill taxonomy and evaluate broader clinical tasks, while carefully auditing failure modes to ensure reliability in agent-mediated clinical pipelines.

Acknowledgments. This work was supported in part by National Natural Science Foundation of China (Nos. 62471148 and T2596013) and STI2030-Major Projects (No. 2021ZD0200204). The computations in this research were supported by the CFFF platform of Fudan University.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-VL technical report. arXiv:2511.21631 (2025)
2. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F.C., Pati, S., et al.: The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv:2107.02314 (2021)
3. Chen, T., Wang, C., Shan, H.: Autoregressive medical image segmentation via next-scale mask prediction. MICCAI 2025 (2025)
4. Deng, L., Chen, L., Liu, M., Wang, X., Qi, Y., Shao, C., Jiang, T.: Knowledge from medical ontology can significantly enhance mainstream text embedding models in medical information retrieval. Information Processing & Management **63**(2), 104435 (2026)
5. Grøvik, E., Yi, D., Iv, M., Tong, E., Rubin, D., Zaharchuk, G.: Deep learning enables automatic detection and segmentation of brain metastases on multisequence MRI. Journal of Magnetic Resonance Imaging **51**(1), 175–182 (2020)
6. Hernandez Petzsche, M.R., De La Rosa, E., Hanning, U., Wiest, R., Valenzuela, W., Reyes, M., Meyer, M., Liew, S.L., Kofler, F., Ezhov, I., et al.: ISLES 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. Scientific Data **9**(1), 762 (2022)
7. 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: ICCV. pp. 4015–4026 (2023)
8. Larson, D.B., Koirala, A., Cheuy, L.Y., Paschali, M., Van Veen, D., Na, H.S., Petterson, M.B., Fang, Z., Chaudhari, A.S.: Assessing completeness of clinical histories accompanying imaging orders using adapted open-source and closed-source large language models. Radiology **314**(2), e241051 (2025)

9. Lee, H.C., Liu, Z., Ahmed, H., Kim, S., Huver, S., Nath, V., Fayad, Z.A., Deyer, T., Mei, X.: Unified supervision for vision-language modeling in 3D computed tomography. In: ICCV. pp. 6725–6733 (2025)
10. Liu, A., Mei, A., Lin, B., Xue, B., Wang, B., Xu, B., Wu, B., Zhang, B., Lin, C., Dong, C., et al.: DeepSeek-V3.2: Pushing the frontier of open large language models. arXiv:2512.02556 (2025)
11. Liu, A., Xue, R., Cao, X.R., Shen, Y., Lu, Y., Li, X., Chen, Q., Chen, J.: MedSAM3: Delving into segment anything with medical concepts. arXiv:2511.19046 (2025)
12. Liu, S., Bao, L., Yang, Q., Geng, W., Zheng, B., Li, C., Chen, W., Peng, H., Yuan, Y.: MedSAM-Agent: Empowering interactive medical image segmentation with multi-turn agentic reinforcement learning. arXiv:2602.03320 (2026)
13. Liu, Y., Yibulayimu, S., Zhu, G., Shi, C., Liang, C., Zhao, C., Wu, X., Sang, Y., Wang, Y.: Automatic pelvic fracture segmentation: a deep learning approach and benchmark dataset. Frontiers in Medicine **12**, 1511487 (2025)
14. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
15. Ma, J., Yang, Z., Kim, S., Chen, B., Baharoon, M., Fallahpour, A., Asakereh, R., Lyu, H., Wang, B.: MedSAM2: Segment anything in 3D medical images and videos. arXiv:2504.03600 (2025)
16. Ma, J., Zhang, Y., Gu, S., Ge, C., Mae, S., Young, A., Zhu, C., Yang, X., Meng, K., Huang, Z., et al.: Unleashing the strengths of unlabelled data in deep learning-assisted pan-cancer abdominal organ quantification: the FLARE22 challenge. The Lancet Digital Health **6**(11), e815–e826 (2024)
17. Pace, D.F., Contreras, H.T., Romanowicz, J., Ghelani, S., Rahaman, I., Zhang, Y., Gao, P., Jubair, M.I., Yeh, T., Golland, P., et al.: HVSMR-2.0: A 3D cardiovascular MR dataset for whole-heart segmentation in congenital heart disease. Scientific Data **11**(1), 721 (2024)
18. 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. NeurIPS **32** (2019)
19. 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: CVPR (2026)
20. de Verdier, M.C., Saluja, R., Gagnon, L., LaBella, D., Baid, U., Tahon, N.H., Foltyn-Dumitru, M., Zhang, J., Alafif, M., Baig, S., et al.: The 2024 brain tumor segmentation (BraTS) challenge: Glioma segmentation on post-treatment MRI. arXiv:2405.18368 (2024)
21. Wu, C., Zhang, X., Zhang, Y., Hui, H., Wang, Y., Xie, W.: Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data. Nature Communications **16**(1), 7866 (2025)
22. Xin, Y., Ates, G.C., Shao, W.: Text3DSAM: Text-guided 3D medical image segmentation using SAM-inspired architecture. In: CVPR: Foundation Models for 3D Biomedical Image Segmentation (2025)
23. Yao, Z., Wang, B., Zhang, Y., Wang, J., Xia, I., Tang, Z., Han, S., Ouyang, F., Yang, Z., Cohan, A., et al.: Medical thinking with multiple images. In: ICLR (2026)
24. 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)

25. Zhao, T., Kiblawi, S., Usuyama, N., Lee, H.H., Preston, S., Poon, H., Wei, M.: Boltzmann attention sampling for image analysis with small objects. In: CVPR. pp. 25950–25959 (2025)
26. Zhao, Z., Zhang, Y., Wu, C., Zhang, X., Zhou, X., Zhang, Y., Wang, Y., Xie, W.: Large-vocabulary segmentation for medical images with text prompts. NPJ Digital Medicine **8**(1), 566 (2025)