Beyond Visual Cues: CoT-Enhanced Reasoning for Semi-supervised Medical Image Segmentation

Yuming Chen[1,2], Yuxin Xie[1,2], Tao Zhou[3], and Yi Zhou[1,2(✉)]

[1] School of Computer Science and Engineering, Southeast University, Nanjing, China
[2] Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications, Ministry of Education, Nanjing, China
[3] Nanjing University of Science and Technology, Nanjing, China
{220252314@seu.edu.cn, yizhou.szcn@gmail.com}

Abstract. Semi-supervised medical image segmentation has emerged as a dominant research problem in medical image analysis, mitigating annotation scarcity by leveraging consistency regularization on unlabeled data. However, existing approaches operate predominantly via visual pattern matching, relying heavily on pixel-level similarities. This visual-centric dependency often falters in clinical scenarios characterized by the visual-semantic mismatch, where visually similar lesions warrant distinct diagnostic conclusions, thus failing to capture the underlying diagnostic logic used by experts. To address this, we move beyond visual cues and propose CERS (CoT-Enhanced Reasoning Segmentation), a framework that integrates Chain-of-Thought (CoT) reasoning to distinguish pathologically distinct cases. Specifically, we construct a knowledge pool enriched with linguistic reasoning descriptions generated by large language models (LLMs). A semantic-aware reference selection strategy is introduced to identify historical evidence, filtering candidates first by morphology, and then refining them via CoT consistency to eliminate hard negatives. Furthermore, a multi-scale coordinate attention module (MCAM) is designed to effectively fuse this reasoning-derived context into the decoding process. Extensive experiments demonstrate the superiority of CERS against state-of-the-art approaches, particularly in resolving boundary ambiguities and semantic inconsistencies. The code is available at https://github.com/cymasuna/CERS.

Keywords: Semi-supervised Medical Image Segmentation · Beyond Visual Cues · Reasoning-derived Context.

1 Introduction

Medical image segmentation is a cornerstone of computer-aided diagnosis, enabling precise delineation of organs and lesions [9,21]. However, the efficacy of deep learning models heavily relies on large-scale, high-quality annotations, which are difficult and laborious to obtain [7]. To address this, Semi-Supervised Learning (SSL) has emerged as a widely adopted paradigm, leveraging limited labeled data alongside abundant unlabeled data [7,18]. Although existing SSL

approaches, ranging from consistency regularization [19,1,24,10] to pseudo label supervision [16,28], have achieved promising results, they predominantly focus on pixel-level visual features.

A fundamental limitation of current SSL methods is their reliance on visual pattern matching. Although effective for standard cases, this approach often fails when faced with ambiguous boundaries, low-contrast lesions, imaging artifacts caused by respiration, or anatomically similar structures that mimic pathological appearance, leading to false positives or missed detections [27]. In clinical practice, physicians overcome such challenges by recalling similar historical cases and applying diagnostic reasoning to infer lesion boundaries beyond purely visual evidence [23]. Following this observation, recent studies have suggested that introducing external reference cues can significantly enhance model performance [29]. However, a critical challenge remains that clinical scenarios often exhibit a visual-semantic mismatch, where images with high visual similarity warrant distinct radiological reading conclusions [5]. Consequently, existing text-driven segmentation methods, which rely solely on superficial and incomplete medical text descriptions, often fail to capture the actual radiological reading logic [10,26,25]. This highlights the necessity to incorporate high-level reasoning capabilities to distinguish pathologically distinct cases that appear visually similar.

To address this challenge, we propose CERS (CoT-Enhanced Reasoning Segmentation), a framework that integrates Chain-of-Thought (CoT) [22] reasoning into the semi-supervised segmentation task. Instead of relying on unreliable visual matching, CERS generates linguistic reasoning descriptions to capture the semantic context of lesions. Using these CoT descriptions, our method performs a semantic-aware retrieval that bridges the gap between visual appearance and diagnostic logic. This ensures that the utilized historical references are not just visually similar, but logically consistent with the target case. These reasoning-guided cues are then effectively injected into the segmentation network to resolve ambiguities in unlabeled data.

The main contributions are summarized as follows. (1) We propose CERS, a CoT-enhanced framework that generates segmentation-aware reasoning descriptions to provide high-level semantic guidance for semi-supervised segmentation. **(2)** We design a Multi-scale Coordinate Attention Module (MCAM) with a dual-decoder consistency scheme to effectively fuse reasoning-derived context into the decoder, significantly improving segmentation accuracy in challenging scenarios. **(3)** We conduct extensive experiments on multiple public datasets, showing that CoT-enhanced reasoning moves segmentation beyond purely visual cues and significantly improves semi-supervised performance.

2 Method

We study semi-supervised medical image segmentation with a labeled set $\mathcal{S}_l = \{(x_i^l, y_i^l, t_i^l)\}_{i=1}^{N_l}$, where $x_i^l \in \mathbb{R}^{H\times W\times C}$ is a medical image, $y_i^l \in \{0,1\}^{H\times W}$ is the corresponding mask and t_i^l is the associated text, and an unlabeled set

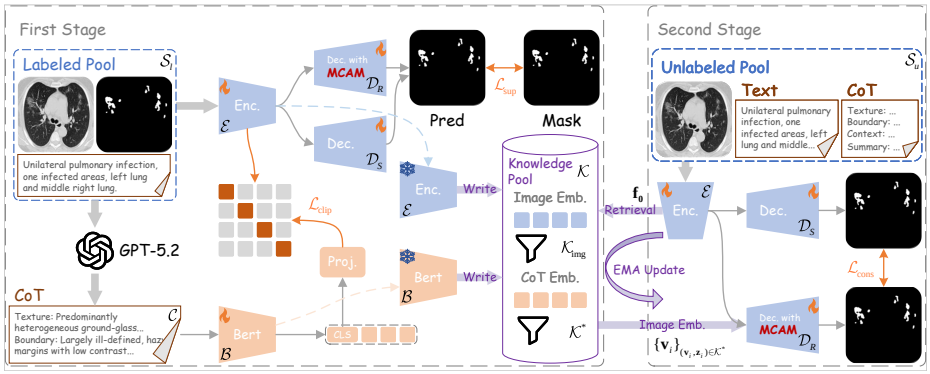

Fig. 1. Overview of CERS. A frozen LLM generates CoTs, and labeled CoTs form a knowledge pool for image–CoT hybrid retrieval. Retrieved context is fused with image features via MCAM and used by a retrieval-aware decoder to enable reasoning-guided semi-supervised segmentation beyond visual cues.

$\mathcal{S}_u = \{(x_j^u, t_j^u)\}_{j=1}^{N_u}$. For every sample, we generate a Chain-of-Thought (CoT) via a frozen LLM, denoted c_i^l for labeled and c_j^u for unlabeled samples. **Only labeled CoTs** are used to populate the knowledge pool $\mathcal{K}$, avoiding leakage from unlabeled data. As illustrated in Fig. 1, our multi-modal CERS model is $f_\theta = \{\mathcal{E}, \mathcal{D}_S, \mathcal{D}_R\}$, where $\mathcal{E}$ is a shared encoder, $\mathcal{D}_S$ a standard segmentation decoder and $\mathcal{D}_R$ a retrieval-aware decoder that fuses retrieved context. CERS employs an image–CoT hybrid retrieval in two stages: a preliminary screening retrieves visually similar candidates to ensure morphological consistency, and a fine-grained re-ranking uses CoT similarity to remove hard negatives that are visually similar but semantically distinct. Retrieved features are fused with the the current features via the Multi-scale Coordinate Attention Module to yield a retrieval-aware enhanced representation, which is consumed by $\mathcal{D}_R$.

2.1 Segmentation-Aware Reasoning Knowledge Pool Construction

CoT Generation. We generate Chain-of-Thought (CoT) descriptions for each sample by prompting GPT-5.2 [17]. For labeled samples we provide the image x_i^l, its mask y_i^l (as a translucent overlay), and text t_i^l; for unlabeled samples we provide only the image x_j^u and text t_j^u. The resulting CoTs for labeled and unlabeled sets are denoted by $\mathcal{C}^l$ and $\mathcal{C}^u$ respectively.

Prompts elicit four diagnostic aspects—*texture*, *boundary*, *context*, and a concise *summary*—thereby encouraging the LLM to produce semantically discriminative cues for segmentation. We omit explicit spatial coordinates because free-text locations are often ambiguous; morphology-aware image embeddings already supply spatial priors. To verify the reliability of the generated prompts, we randomly sampled 100 cases for expert manual review, which confirmed the high clinical accuracy (more than 95%) of the LLM-generated CoT.

Warm-up to Build Knowledge Pool. To initialize the model and ensure stable retrieval in the subsequent SSL phase, we first conduct a supervised warm-up (Eq. 2) using **only the labeled set** $\mathcal{S}_l$. During this phase the multi-modal model $f_\theta = \{\mathcal{E}, \mathcal{D}_S, \mathcal{D}_R\}$ is trained on labeled samples to learn reliable segmentation and cross-modal embeddings. The supervised segmentation loss is defined as:

$$\mathcal{L}_{\text{sup}} = \frac{1}{|\mathcal{S}_l|} \sum_{(x,y,t)\in\mathcal{S}_l} \Big[\mathcal{L}_{\text{seg}}\big(\mathcal{D}_S(\mathcal{E}(x,t)), y\big) + \mathcal{L}_{\text{seg}}\big(\mathcal{D}_R(\mathcal{E}(x,t)), y\big)\Big], \tag{1}$$

where $\mathcal{L}_{\text{seg}}$ combines standard cross-entropy loss and Dice loss. We obtain an image embedding $\mathbf{v}_i = \mathcal{E}(x_i^l, t_i^l)$ and a CoT embedding $\mathbf{z}_i = \mathcal{B}(c_i^l)$ using a texture encoder $\mathcal{B}$ and apply contrastive objective [14] $\mathcal{L}_{\text{clip}}$ to align these embeddings. The overall warm-up loss is defined as:

$$\mathcal{L}_{\text{warm}} = \mathcal{L}_{\text{seg}} + \lambda_{\text{clip}}\mathcal{L}_{\text{clip}}, \tag{2}$$

where λ_{clip} balances segmentation supervision and cross-modal alignment. After warm-up convergence, the trained image encoder $\mathcal{E}$ and texture encoder $\mathcal{B}$ are fixed and used to encode all labeled samples. These paired representations are then stored in the knowledge pool $\mathcal{K} = \{(\mathbf{v}_i, \mathbf{z}_i)\}_{i=1}^{N_l}$, which serves as the retrieval database in the subsequent semi-supervised learning stage.

2.2 Image–CoT Joint Cues

In the second training stage, we continue optimization based on the warm-up model and leverage the constructed knowledge pool $\mathcal{K}$ to retrieve reliable cues for both labeled and unlabeled samples. Given an input sample x and the associated text with its generated CoT c, we first extract the image embedding $\mathbf{v} = \mathcal{E}(x,t)$ and the CoT embedding $\mathbf{z} = \mathcal{B}(c)$. CERS performs retrieval in a coarse-to-fine manner: **(1) Image-based coarse filtering**: To ensure visual consistency, we conduct an image-based coarse retrieval over the knowledge pool using cosine similarity, $s_i^{\text{img}} = \cos(\mathbf{v}, \mathbf{v}_i)$, where $(\mathbf{v}_i, \mathbf{z}_i) \in \mathcal{K}$. The top-$K$ samples with the highest s_i^{img} are selected to form a candidate set $\mathcal{K}_{\text{img}}$, which contains morphologically similar samples. **(2) CoT-based fine re-ranking**: Among the candidates in $\mathcal{K}_{\text{img}}$, we further compute CoT similarity to refine retrieval results, $s_i^{\text{cot}} = \cos(\mathbf{z}, \mathbf{z}_i)$, where $(\mathbf{v}_i, \mathbf{z}_i) \in \mathcal{K}_{\text{img}}$. The candidates are re-ranked according to s_i^{cot}, and the top-M samples are retained as the final retrieved evidence set $\mathcal{K}^*$. This fine-grained re-ranking effectively filters out *hard negatives* that are visually similar but diagnostically inconsistent with the query.

2.3 Multi-scale Coordinate Attention Module

Using the features $\mathbf{f}^0 = \mathcal{E}(x,t)$ from the encoder as queries and the retrieved features $\{\mathbf{v}_i\}_{(\mathbf{v}_i,\mathbf{z}_i)\in\mathcal{K}^*}$ as keys and values enables the model to selectively retrieve and integrate relevant external knowledge guided by the current visual

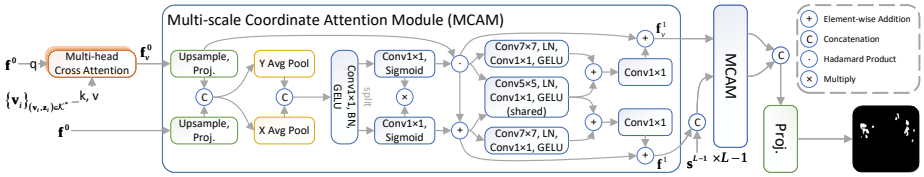

Fig. 2. Detailed structure of Multi-scale Coordinate Attention Module.

context, thereby enhancing feature discriminability and contextual consistency. It is formulated via multi-head cross attention [20] as:

$$\mathbf{f}_{\mathbf{v}}^{0} = \mathrm{MHCA}(\mathbf{f}^{0}, \{\mathbf{v}_i\}, \{\mathbf{v}_i\}),\ (\mathbf{v}_i, \mathbf{z}_i) \in \mathcal{K}^{*}. \tag{3}$$

As depicted in Fig. 2, Multi-scale Coordinate Attention Module (MCAM) is employed in the decoder to fuse the feature $\mathbf{f}^0$ with $\mathbf{f}_{\mathbf{v}}^{0}$ in a spatially aware manner. The resulting representation $\mathbf{f}^1$ combined with skip features $\mathbf{s}^{l-1}$ of $l-1^{\mathbf{th}}$ encoder layer and $\mathbf{f}_{\mathbf{v}}^{1}$ is subsequently fed into next MCAM in $\mathcal{D}_R$. The decoding and fusion for each stage can be expressed as follows:

$$(\mathbf{f}^{i+1}, \mathbf{f}_{\mathbf{v}}^{i+1}) = \mathrm{MCAM}(\mathrm{Concat}(\mathbf{f}^{i}, \mathbf{s}^{L-i}), \mathbf{f}_{\mathbf{v}}^{i}),\ i \in [1, L-1], \tag{4}$$

where L is the number of main stages in the encoder. Inspired by coordinate attention [8], MCAM enhances spatial sensitivity by modeling feature responses along horizontal and vertical directions, which is beneficial for preserving boundary structures during decoding. In addition, multi-scale feature processing allows the module to capture both local details and global context, improving the effectiveness of reference feature integration.

2.4 Knowledge-Guided Learning Objective

Knowledge Update. The knowledge pool $\mathcal{K}$ stores a visual prototype $\mathbf{v}_i^{\mathcal{K}}$ and a CoT prototype $\mathbf{z}_i^{\mathcal{K}}$ for each labeled sample i. The visual prototype are updated during training via exponential moving average (EMA). Concretely, given the current pooled image feature $\bar{\mathbf{f}}_i = \mathcal{E}(x_i^l)$, the pool entries are updated as:

$$\mathbf{v}_i^{\mathcal{K}} \leftarrow \alpha\, \mathbf{v}_i^{\mathcal{K}} + (1-\alpha)\, \bar{\mathbf{f}}_i, \tag{5}$$

where $\alpha \in [0, 1)$ is the EMA momentum. Updates are performed only for labeled entries, and stored prototypes are detached from the gradient graph to stabilize retrieval and allow inexpensive persistence.

Total Loss Function. For unlabeled samples we enforce cross-branch consistency between the two decoder outputs. The consistency loss is written as:

$$\mathcal{L}_{\mathrm{cons}} = \frac{1}{|\mathcal{S}_u|} \sum_{x \in \mathcal{S}_u} \left\| \mathcal{D}_S\left(\mathcal{E}(x,t)\right) - \mathcal{D}_R\left(\mathcal{E}(x,t), \{\mathbf{v}_i\}_{(\mathbf{v}_i, \mathbf{z}_i) \in \mathcal{K}^{*}}\right) \right\|_2^2. \tag{6}$$

Overall training objective is by weighting supervised and consistency terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sup}} + \lambda_{\text{cons}}\mathcal{L}_{\text{cons}}. \tag{7}$$

Notably, retrieval affects only the output of $\mathcal{D}_R$, so the consistency term encourages the base decoder $\mathcal{D}_S$ to absorb useful retrieval-induced cues without directly exposing $\mathcal{D}_S$ to retrieved features.

3 Experiments

3.1 Datasets and Experimental Details

Datasets. We conducted experiments on three public datasets, covering different modalities. (1) MosMedData+ [13] consists of 2 729 chest CT scans with COVID-19 infection annotations. (2) QaTa-COV19 [4] contains 9 258 chest X-ray images with COVID-19 infection segmentation masks. (3) BRISC 2025 [6] consists of 4 793 brain MRIs with annotated tumor regions and lesion masks. We split each dataset into train/val/test with an 8:1:1 ratio. The textual annotations for MosMedData+ and QaTa-COV19 follow LViT [10], while BRISC 2025 annotations were auto-generated from metadata and graphical features.

Implementation Details. All experiments including both training and evaluation phases were conducted on a single NVIDIA RTX 4090 GPU. We set $\lambda_{\text{clip}} = 0.5$ and $\alpha = 0.99$. The λ_{cons} was initialized as 0.003 and increased via a standard sigmoid ramp-up. For fair comparison, all methods adopted ConvNeXt [11] as the backbone and CXR-BERT [2] as CoT encoder whenever applicable. Each model was re-implemented and trained for 200 epochs under the same training protocol. The batch size was set to 16, and optimization was performed using SGD with an initial learning rate of $1e^{-2}$ and a weight decay of $1e^{-4}$. Performance was evaluated using Dice and IoU.

3.2 Comparison with State-of-the-Art Methods

We compare against six SOTA SSL baselines: BCP [1], DuCiSC [24], LeFeD [28], LViT [10], MT [19] and UCMT [16]. We evaluate semi-supervised performance under 50% and 25% labeled ratios across three public datasets. The quantitative comparison results are presented in Table 1. Compared with existing methods, CERS consistently outperforms SOTA methods in most metrics and datasets.

Fig. 3 provides qualitative examples where CERS better localizes lesions and preserves boundaries under ambiguous, low-contrast, artifact-corrupted, or visually confusing cases—demonstrating the value of moving beyond purely visual cues. Unlike pseudo-labeling or teacher-updating methods that suffer from confirmation bias, CERS mitigates noisy supervision by leveraging image–CoT joint cues. Our method remains effective even under fully supervised learning, significantly outperforming the U-Net [15] model that employ the same backbone.

Table 1. Quantitative comparison in terms of Dice (%) and IoU (%) on MosMedData+, QaTa-COV19 and BRISC 2025 with 50% and 25% labeled ratio. **Bold** and underlined entries indicate the best and second-best results respectively.

Method	Text	Backbone	Labeled	MosMedData+		QaTa-COV19		BRISC 2025	
				Dice	IoU	Dice	IoU	Dice	IoU
U-Net [15]		ConvNeXt	100%	76.99	66.03	75.70	64.90	82.57	74.21
Ours	✓	ConvNeXt	100%	79.21	68.53	81.40	71.60	84.40	76.28
BCP [1]		ConvNeXt	50%	72.57	60.79	72.59	60.92	76.66	67.42
DuCiSC [24]		ConvNeXt	50%	74.24	63.08	71.51	59.84	78.85	70.67
LeFeD [28]		V-Net [12]	50%	74.77	63.38	74.64	63.74	79.63	72.13
LViT [10]	✓	LViT	50%	75.47	63.27	79.08	69.65	83.85	75.88
MT [19]		ConvNeXt	50%	73.44	61.97	73.39	62.22	76.38	67.88
UCMT [16]		ConvNeXt	50%	75.95	64.65	74.42	63.33	80.04	71.70
Ours	✓	ConvNeXt	50%	**78.26**	**66.87**	**81.23**	**71.42**	**84.03**	**76.03**
BCP [1]		ConvNeXt	25%	71.31	59.37	72.44	60.81	72.46	62.43
DuCiSC [24]		ConvNeXt	25%	71.67	60.30	70.20	58.22	76.38	68.04
LeFeD [28]		V-Net [12]	25%	70.45	58.95	74.31	63.61	78.64	71.01
LViT [10]	✓	LViT	25%	71.16	58.13	77.48	67.30	**80.51**	**72.98**
MT [19]		ConvNeXt	25%	70.08	58.34	72.86	61.48	71.12	62.42
UCMT [16]		ConvNeXt	25%	72.00	60.47	72.15	60.69	74.32	66.19
Ours	✓	ConvNeXt	25%	**74.73**	**63.39**	**80.52**	**70.52**	80.34	71.80

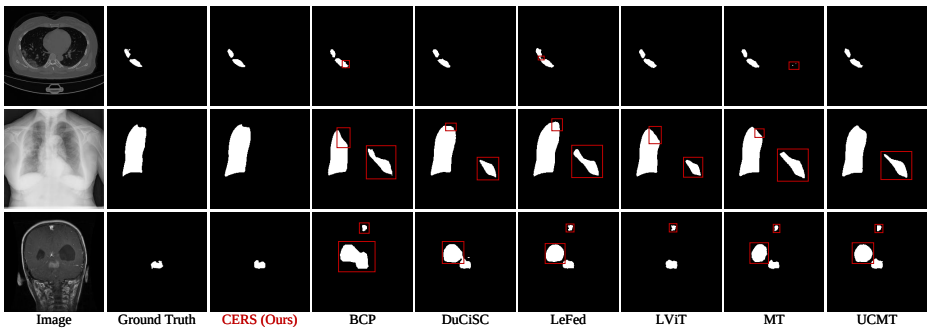

Fig. 3. Exemplar qualitative results of different approaches on MosMedData+ (row 1), QaTa-COV19 (row 2) and BRISC 2025 (row 3).

Table 2. Ablation study on different retrieval strategies of CERS. Latency refers to the time overhead incurred per training step due to the retrieval computation.

Text Retrieval	CoT Retrieval	Image Retrieval	MosMedData+		QaTa-COV19		Latency per Step
			Dice	IoU	Dice	IoU	
			73.51	62.36	74.07	62.86	0.0
✓			76.68	65.54	79.43	69.13	+0.048s
	✓		<u>77.22</u>	<u>66.07</u>	<u>80.86</u>	<u>71.05</u>	+0.055s
		✓	73.60	62.14	80.66	70.66	+0.047s
✓		✓	77.07	65.90	80.20	70.16	+0.099s
	✓	✓	**78.26**	**66.31**	**81.23**	**71.42**	+0.103s

3.3 Ablation Study

Impact of Different Retrieval. Table 2 presents the results of different search strategies on the MosMedData+ and QaTa-COV19 datasets with a 50% labeled ratio. The retrieval strategy provides a reference for the model to perform segmentation. Image retrieval defines a coarse boundary for the search space, while CoT further bridges the gap between visual appearance and diagnostic logic.

Impact of Top-K on Different Labeled Data Ratios. The performance of different numbers of CoT retrieval cues under varying labeled ratios is illustrated in Fig. 4. Utilizing too few cues fails to provide a robust reference for the model, whereas an excessive number increases the risk of introducing noise. For the MosMedData+ dataset, $K = 5$ proves to be a reasonable choice.

Impact of MCAM. We replace the MCAM with a standard U-Net [15] decoder and a Swin U-Net [3] decoder, with the results presented in Table 3. As demonstrated in Fig. 3, MCAM effectively preserves boundary structures and generates smoother, more precise segmentation masks compared to other decoders. To further validate the internal design of MCAM, we conduct ablation studies by replacing its sub-components with 1×1 convolutions. Specifically, removing

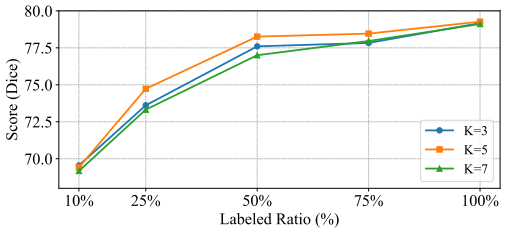

Fig. 4. CERS performance under different labeled ratios and top-K on MosMedData+.

Table 3. Ablation study on different decoders to fuse cues.

Decoder	Total Param	BRISC 2025	
		Dice	IoU
U-Net	51.36M	82.58	74.68
Swin U-Net	57.09M	79.22	70.26
Ours	48.49M	**84.03**	**76.03**

the Coordinate Attention module causes a 1.52% Dice drop, while removing the Multi-scale Branch leads to a 0.95% drop, confirming the individual effectiveness of both modules within MCAM. Moreover, MCAM efficiently fuses and leverages the image–CoT joint cues without introducing a number of additional parameters to the overall model.

4 Conclusion

In this paper, we present CERS, a CoT-enhanced semi-supervised segmentation framework that moves beyond purely visual supervision by incorporating reasoning-driven semantic guidance. Our results suggest that introducing structured diagnostic reasoning helps distinguish visually similar but semantically different regions, improving robustness in challenging scenarios where visual cues alone are insufficient. This finding highlights the importance of high-level semantic reasoning for medical image segmentation, especially under limited annotation settings. Despite its effectiveness, CERS relies on the quality of generated CoTs and an external retrieval mechanism, which may introduce additional computational overhead and dependency on the reasoning model. Future work will explore more efficient reasoning representations and end-to-end integration with medical foundation models to further improve scalability and generalization.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (Grant No 62476054 and Grant No 62576153).

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

References

1. Bai, Y., Chen, D., Li, Q., Shen, W., Wang, Y.: Bidirectional copy-paste for semi-supervised medical image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11514–11524 (2023)
2. Boecking, B., Usuyama, N., Bannur, S., Castro, D.C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., et al.: Making the most of text semantics to improve biomedical vision–language processing. In: European conference on computer vision. pp. 1–21. Springer (2022)
3. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European conference on computer vision. pp. 205–218. Springer (2022)
4. Degerli, A., Kiranyaz, S., Chowdhury, M.E., Gabbouj, M.: Osegnet: Operational segmentation network for covid-19 detection using chest x-ray images. In: 2022 IEEE International Conference on Image Processing (ICIP). pp. 2306–2310. IEEE (2022)
5. Fang, J., Fu, H., Zeng, D., Yan, X., Yan, Y., Liu, J.: Combating ambiguity for hash-code learning in medical instance retrieval. IEEE Journal of Biomedical and Health Informatics **25**(10), 3943–3954 (2021)

6. Fateh, A., Rezvani, Y., Moayedi, S., Rezvani, S., Fateh, F., Fateh, M.: Brisc: Annotated dataset for brain tumor segmentation and classification with swin-hafnet. arXiv e-prints pp. arXiv–2506 (2025)
7. Han, K., Sheng, V.S., Song, Y., Liu, Y., Qiu, C., Ma, S., Liu, Z.: Deep semi-supervised learning for medical image segmentation: A review. Expert Systems with Applications **245**, 123052 (2024)
8. Hou, Q., Zhou, D., Feng, J.: Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13713–13722 (2021)
9. Khan, W., Leem, S., See, K.B., Wong, J.K., Zhang, S., Fang, R.: A comprehensive survey of foundation models in medicine. IEEE Reviews in Biomedical Engineering (2025)
10. Li, Z., Li, Y., Li, Q., Wang, P., Guo, D., Lu, L., Jin, D., Zhang, Y., Hong, Q.: Lvit: language meets vision transformer in medical image segmentation. IEEE transactions on medical imaging **43**(1), 96–107 (2023)
11. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)
12. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. Ieee (2016)
13. Morozov, S.P., Andreychenko, A.E., Pavlov, N.A., Vladzymyrskyy, A., Ledikhova, N.V., Gombolevskiy, V.A., Blokhin, I.A., Gelezhe, P.B., Gonchar, A., Chernina, V.Y.: Mosmeddata: Chest ct scans with covid-19 related findings dataset. arXiv preprint arXiv:2005.06465 (2020)
14. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
15. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
16. Shen, Z., Cao, P., Yang, H., Liu, X., Yang, J., Zaiane, O.R.: Co-training with high-confidence pseudo labels for semi-supervised medical image segmentation. arXiv preprint arXiv:2301.04465 (2023)
17. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025)
18. Song, Y., Wang, T., Cai, P., Mondal, S.K., Sahoo, J.P.: A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. ACM Computing Surveys **55**(13s), 1–40 (2023)
19. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)
20. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
21. Wang, R., Lei, T., Cui, R., Zhang, B., Meng, H., Nandi, A.K.: Medical image segmentation using deep learning: A survey. IET image processing **16**(5), 1243–1267 (2022)

22. 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)
23. Welter, P., Deserno, T.M., Fischer, B., Günther, R.W., Spreckelsen, C.: Towards case-based medical learning in radiological decision making using content-based image retrieval. BMC medical informatics and decision making **11**(1), 68 (2011)
24. Wu, H., Wang, C., Cui, Z.: Dual cross-image semantic consistency with self-aware pseudo labeling for semi-supervised medical image segmentation. IEEE Transactions on Medical Imaging (2025)
25. Xie, Y., Chen, Y., Yang, Y., Zhou, Y., Zhou, T., Zhao, Z., Liu, J., Fu, H.: Coreseg: Reasoning-driven segmentation for complex lesions via reinforcement learning. arXiv preprint arXiv:2603.05911 (2026)
26. Xie, Y., Zhou, T., Zhou, Y., Chen, G.: Simtxtseg: Weakly-supervised medical image segmentation with simple text cues. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 634–644. Springer (2024)
27. Zeng, Q., Luo, H., Ma, X., Lu, Z., Hu, Y., Xia, Y.: Exploring text-enhanced mixture-of-experts for semi-supervised medical image segmentation with composite data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 226–236. Springer (2025)
28. Zeng, Q., Xie, Y., Lu, Z., Lu, M., Zhang, J., Xia, Y.: Consistency-guided differential decoding for enhancing semi-supervised medical image segmentation. IEEE Transactions on Medical Imaging **44**(1), 44–56 (2024)
29. Zhao, L., Chen, X., Chen, E.Z., Liu, Y., Chen, T., Sun, S.: Retrieval-augmented few-shot medical image segmentation with foundation models. IEEE Transactions on Neural Networks and Learning Systems (2025)