MediRound: Multi-Round Entity-Level Reasoning Segmentation in Medical Images

Qinyue Tong[1], Ziqian Lu[2], Jun Liu[1], Rui Zuo[1], Zhe-Ming Lu[1(✉)], and Yueming Jin[3(✉)]

[1] Zhejiang University, Hangzhou, Zhejiang, China
{qinyuetong,junliu0930,ruizuo,zheminglu}@zju.edu.cn
[2] Zhejiang Sci-Tech University, Hangzhou, Zhejiang, China
ziqianlu@zstu.edu.cn
[3] National University of Singapore, Singapore
ymjin@nus.edu.sg

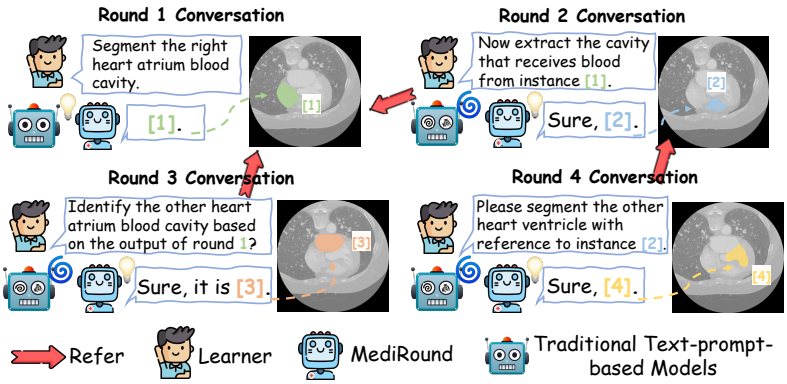

Fig. 1: A medical education dialogue illustrating the proposed MediRound. Our model can comprehend user queries that refer to the mask results from previous rounds (*e.g., the Round 2 query refers to the Round 1 mask result*), enabling cross-round entity-level reasoning in multi-round medical conversations. This iterative querying paradigm allows learners to progressively develop their understanding of medical entities. In contrast, conventional text-prompt-based medical segmentation methods struggle with such a complex and context-dependent task.

Abstract. Despite notable progress in text-guided medical image segmentation nowadays, these methods are limited to single-round dialogues and fail to support multi-round reasoning, which is important for **medical education** scenarios. In this work, we introduce Multi-Round Entity-Level Medical Reasoning Segmentation (**MEMR-Seg**), a new task that requires generating segmentation masks through multi-round queries with entity-level reasoning, helping learners progressively develop their understanding of medical knowledge. To support this task, we construct **MR-MedSeg**, a large-scale dataset of 177K multi-round medical segmentation dialogues, featuring entity-based reasoning across rounds. Furthermore, we propose **MediRound**, an effective baseline

model designed for multi-round medical reasoning segmentation. To mitigate the inherent error propagation within the chain-like pipeline of multi-round segmentation, we introduce a lightweight yet effective **Judgment & Correction Mechanism** during model inference. Experimental results demonstrate that our method effectively addresses the MEMR-Seg task and outperforms conventional medical referring segmentation methods. *The project is available at https://github.com/Edisonhimself/MediRound.*

Keywords: Multi-Round · Medical Segmentation · Benchmark Dataset

1 Introduction

Medical image segmentation focuses on identifying and outlining key anatomical or lesion-related regions across diverse medical imaging modalities [29, 32, 43]. This capability underpins a broad range of downstream medical applications, including clinical analysis, medical education, and biomedical research [4, 15, 27, 45, 48]. Most existing studies focus predominantly on task-specific segmentation [9, 26, 35], commonly referred to as "specialist models". Despite their remarkable performance, these models generally suffer from limited adaptability and lack the interactive capabilities necessary for practical, real-world usage.

To address this issue, recent text-prompt-based medical image segmentation methods [5, 11, 20, 53] enable users to specify target regions through natural language instructions, allowing segmentation to be guided more flexibly according to user intent. Building upon these developments, several MLLM-based approaches [13, 41, 49] further enhance the interactivity and practicality by enabling reasoning-driven medical segmentation, allowing users to obtain the desired masks through implicit queries.

However, these models primarily focus on single-round medical segmentation and are incapable of supporting multi-round and continuous text-based interactions. Although the one-shot querying paradigm can be efficient in clinical use, it presumes high-quality prompts and substantial user expertise. As a result, it is well suited to medical experts but less conducive to the learning process of **non-expert users**. In **medical education** settings, visual understanding is typically developed **progressively**. As illustrated in Figure 1, learners tend to formulate new segmentation queries based on the mask results obtained from previous rounds. For example, in *Round 2 Conversation*, the user's query is based on the mask generated in *Round 1 Conversation*, and the results of *Round 2 Conversation* in turn serve as a reference for *Round 4 Conversation*. Such iterative questioning helps learners progressively develop a better visual understanding of different medical entities and grasp the relationships among them more effectively. Moreover, these follow-up queries are often relation-driven and logically constrained, reflecting how trainees learn anatomy and pathology through relational concepts rather than precise standalone descriptions, which aligns with how learning occurs in real-world medical education. Nevertheless, due to the lack of multi-round entity-level reasoning capabilities, existing methods often

fail to produce the expected responses when dealing with inputs that involve cross-round logical dependencies.

In this work, we define a new medical vision-language task, termed **MEMR-Seg** (Multi-Round Entity-Level Medical Reasoning Segmentation), which entails generating binary segmentation masks based on multi-round medical image queries, involving entity-level reasoning across dialogue rounds. Notably, the queries are not only multi-round in nature, but each round is also a derivative and extended inquiry based on the entity results from the previous round. To successfully perform this new task, the model should possess two crucial abilities: *(1) supporting multi-round dialogue and cross-round entity-level reasoning*; and *(2) generating accurate segmentation results in response to user queries.*

As data scarcity remains a critical challenge in accomplishing MEMR-Seg, we first propose a Multi-Round Entity-Level Medical Reasoning Segmentation dataset, termed **MR-MedSeg**, which contains a large number of medical conversations centered on entity-level reasoning segmentation with multi-round queries and answers. We collect all metadata from the publicly accessible SA-Med2D-20M [51]. Subsequently, inspired by SegLLM [44], we construct multi-round question–answer pairs through a combination of manual annotation and GPT-5-based generation. Consequently, the MR-MedSeg dataset comprises a large and diverse collection of 177K conversations. It is worth noting that a major difference between our MR-MedSeg and traditional medical segmentation datasets [2, 10, 36–38, 41] lies in the fact that our dataset not only features multi-round dialogues but also incorporates entity-based reasoning logic across different dialogue rounds. This design enhances the educational utility of MR-MedSeg and facilitates more interactive medical image segmentation.

Furthermore, building upon existing reasoning-based segmentation works [18, 42, 44, 50], we introduce **MediRound**, an effective baseline model designed for multi-round medical reasoning segmentation. MediRound integrates the segmentation feature representations from the reference round and the textual information from previous dialogues with the current query. These are jointly embedded into the input sequence of the inner-MLLM. This design enables the model not only to comprehend the current query but also to access the core information from the reference round while maintaining awareness of the entire conversational history. Moreover, to address the inevitable issue of error accumulation inherent in multi-round segmentation (*as illustrated in Figure 1, if the result of Round 1 contains errors, these errors may propagate to Round 2, and the resulting inaccuracies in Round 2 can further affect Round 4*), we introduce a lightweight yet effective **Judgment & Correction Mechanism** during the model inference stage. Applied after the end-to-end model training phase, this mechanism helps MediRound mitigate error propagation in practical multi-round inference and effectively enhances the overall accuracy of multi-turn entity-level reasoning. Extensive experimental results demonstrate the superior performance of MediRound in both multi-round and single-round medical segmentation, while confirming the effectiveness of the Judgment & Correction Mechanism. Overall, we summarize our main contributions as follows:

- We introduce the Multi-Round Entity-Level Medical Reasoning Segmentation (**MEMR-Seg**) task, which requires reasoning over multi-round queries that refer to results from earlier rounds in medical images.
- We construct the **MR-MedSeg** dataset, which comprises 177K multi-round medical segmentation dialogues that require multi-round reasoning.
- We introduce an effective baseline model, **MediRound**, and propose a **Judgment & Correction Mechanism** to address the inherent error accumulation during multi-round reasoning segmentation. Comprehensive experiments verify the effectiveness of our model and the utility of the proposed mechanism.

2 Related Works

Medical Image Segmentation. As one of the foundational architectures in medical image segmentation, the U-Net family [7, 14, 35, 54] has inspired numerous extensions and variants. Despite their strong performance on specific tasks and imaging modalities, these fully automated and task-specific models often lack the flexibility and user interactivity necessary for adaptability in practical, real-world medical environments. Text-prompt-based segmentation models address this by allowing users to guide segmentation via language prompts [5,8,24,53], and recent works further support complex implicit queries [13,41,49], improving flexibility and interactivity. However, they are unable to handle continuous, multi-turn queries or perform cross-round reasoning over medical entities, both of which are common and crucial in medical education scenarios. This work primarily centers on multi-round entity-based medical reasoning segmentation, enabling entity-centered reasoning across dialogue rounds while generating accurate segmentation masks.

Reasoning Segmentation. General reasoning segmentation is introduced by LISA [18], which interprets implicit textual prompts to generate segmentation masks by combining MLLMs' vision-language understanding [21] with SAM's segmentation capabilities [17]. More recently, several follow-up studies extend reasoning segmentation [16,30,33,52]. Among them, SegLLM [44] represents the initial effort to address multi-round reasoning segmentation in natural images. Compared to natural image domains, multi-round reasoning segmentation in medical imaging could provide enhanced practical utility. This is because such a segmentation paradigm aligns well with the process through which trainees learn medical knowledge. However, despite recent progress in medical reasoning segmentation [13, 41, 47, 49], this aspect remains unaddressed. Concurrent with our work, MTurn-Seg [28] introduces a bilingual benchmark for multi-turn medical image dialogue that incorporates segmentation. In this work, building upon prior works [19, 25, 44, 51], we develop a method for multi-round reasoning in medical image segmentation.

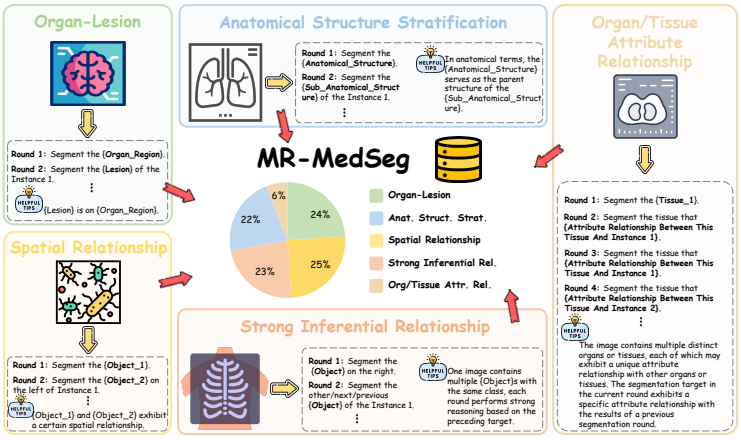

Fig. 2: Overview of MR-MedSeg. Our dataset comprises five types of medical reasoning dialogues, each characterized by a specific form of inter-instance relationship, encompassing nearly all multi-round segmentation scenarios in medical education settings.

3 MR-MedSeg

Dataset Overview. The overview of our proposed MR-MedSeg dataset is shown in Figure 2. As illustrated in the figure, we design five representative types of interaction scenarios tailored for multi-round entity-level medical reasoning segmentation, each reflecting realistic use cases in medical education practice. These scenarios each involve distinct forms of inter-instance relationships, including organ–lesion dependency, anatomical hierarchy, organ/tissue attribute association, spatial relationship, and strong inferential relationship. Figure 2 further presents the overall composition and proportions of these scenarios in MR-MedSeg.

Data Construction Pipeline. As shown in Figure 3, we design a semi-automatic pipeline for constructing the MR-MedSeg dataset.

In *Step 1*, we manually select appropriate medical entities according to the five types of multi-round medical interaction scenarios illustrated in Figure 2. Specifically, we traverse the SA-Med2D-20M dataset [51] and identify sub-datasets that meet the criteria of each scenario. From these selected sub-datasets, we further manually select the medical entities that best match the five designed interaction settings. MR-MedSeg is built upon the curated image, mask, and label annotations of SA-Med2D-20M, which substantially alleviates the burden of data cleaning, preprocessing, and manual annotation. Next, in *Step 2*, we construct diverse inter-entity relationships for the selected medical entities based on the entity relationships defined across five scenarios, employing a hybrid strategy that combines *manual annotation* with *GPT-5-based generation*. Notably, when generating relationships that are independent of image information, such as biological or anatomical attributes shared among medical entities which are generally consistent across images, we omit the reference to the im-

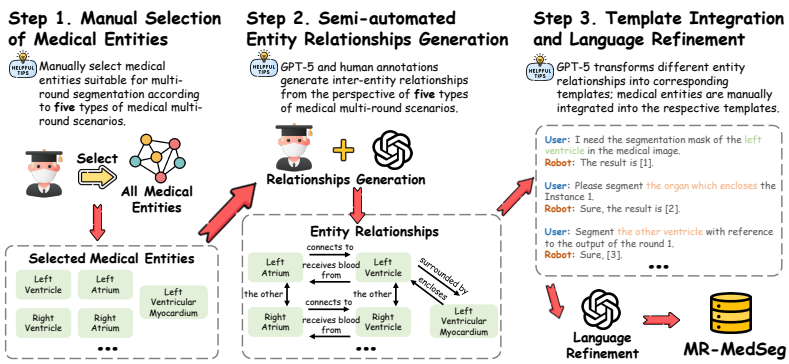

Fig. 3: Semi-automatic pipeline for constructing MR-MedSeg dataset. The process includes three stages: entity selection, relationship generation, and template integration. We construct the pipeline with manual annotation, augmented by GPT-5 generation.

age context. In contrast, when generating relationships that depend on image information, such as the relative spatial positions of anatomical structures that vary between images, we explicitly incorporate the corresponding image context to ensure both accuracy and contextual relevance. Subsequently, in *Step 3*, we employ GPT-5 [39] to transform each entity relationship into 50–80 diverse yet semantically equivalent templates. The corresponding medical entities are then manually integrated into these templates to construct the initial version of the multi-round medical segmentation data. Finally, we perform an additional automatic refinement process using GPT-5 to further enhance the fluency and accuracy of the generated dialogues. Moreover, we engage three inspectors with professional medical expertise to screen and filter the data in MR-MedSeg, thereby maximizing the validity and reliability of the dataset.

Data Statistics. The MR-MedSeg dataset comprises over 177K multi-round medical segmentation dialogues that require cross-round reasoning, constructed from an initial candidate pool of 118K images and 569K masks. We assign 174,934 conversations to the training set, 1,270 to the validation set, and 1,273 to the test set. MR-MedSeg further covers 168 distinct medical entities, spanning a wide range of anatomical structures, pathological findings, and tissue types. The dataset also covers 9 medical imaging modalities, further highlighting its diversity.

4 MediRound

We propose an effective baseline for multi-round medical reasoning segmentation, termed MediRound. MediRound demonstrates the ability to perform multi-round reasoning over segmented medical entities by understanding information from preceding rounds to achieve fine-grained segmentation in the current round. To address the inherent drawback that errors in previous rounds can severely affect the performance of subsequent rounds in the multi-round segmentation

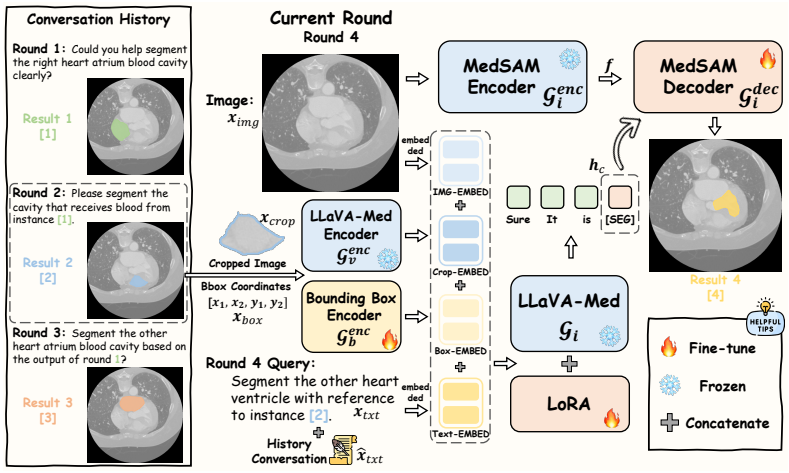

Fig. 4: Overview of the MediRound framework. The figure illustrates the model's workflow in processing the Round 4 conversation, referring to the Round 2 mask. The encoder $\mathcal{G}_v^{enc}$ consists of the vision encoder and visual projection layer from LLaVA-Med.

pipeline, we introduce a novel Judgment & Correction Mechanism (JCM). Details of MediRound and JCM follow.

4.1 Model Architecture

Figure 4 illustrates the model architecture of our proposed MediRound, which is composed of the following modules: MedSAM [25] serving as the vision backbone with an image encoder $\mathcal{G}_i^{enc}$ and a mask decoder $\mathcal{G}_i^{dec}$; LLaVA-Med [19] functioning as the multimodal large language foundation model $\mathcal{G}_i$; the LLaVA-Med visual encoder $\mathcal{G}_v^{enc}$ for extracting and projecting image embeddings; and a linear projection acting as the bounding box encoder $\mathcal{G}_b^{enc}$.

4.2 Multi-Round Reasoning and Segmentation

Inspired by existing MLLM-based segmentation methods [16, 50, 52] and the pioneer in general multi-round segmentation method SegLLM [44], we propose MediRound tailored for multi-round medical reasoning segmentation. Specifically, we first extend the vanilla vocabulary of LLM by introducing a special token, [SEG], which indicates a request to produce the segmentation mask. As shown in Figure 4, **three rounds of conversations have already been established**. Given the **Round 4** query x_{txt} together with the original input image x_{img} and history conversation $\hat{x}_{txt}$ (including the question–answer pairs from Rounds 1, 2, and 3), we first utilize the mask output from the **reference round (Round 2)** to crop the corresponding object x_{crop} from the original image. The cropped region and its bounding box coordinates x_{box} in the original

Algorithm 1 Multi-Round Inference of MediRound with JCM

Require: Image x_{img}; multi-round queries $\{x_{txt}^{(t)}\}_{t=1}^{T}$; $\mathcal{G}_i$: LLaVA-Med; $\mathcal{G}_i^{dec}$: MedSAM decoder; h_c: [SEG] feature; MLP_J: Quality Judgment Module; MLP_C: Correction Module; β: the threshold used by Quality Judgment Module to judge the quality of [SEG] feature
Ensure: Predicted masks $\{\hat{\mathbf{M}}^{(t)}\}_{t=1}^{T}$, with initialization $\hat{\mathbf{M}}^{(0)} \leftarrow \emptyset$
for $t = 1$ to T **do**
  $r \leftarrow$ In this round, user selects the previous round r mask from $\{0, 1, ..., t-1\}$ for reference; $r = 0$ means no mask is referenced
  $\tilde{x}_{txt}^{(t)} \leftarrow \text{IntegrateMaskInfo}(x_{txt}^{(t)}, \hat{\mathbf{M}}^{(r)})$
  $h_c \leftarrow \mathcal{G}_i(x_{img}, \tilde{x}_{txt}^{(t)})$
  $q \leftarrow \text{MLP}_J(h_c)$
  if $q > \beta$ **then**
    $\hat{\mathbf{M}}^{(t)} \leftarrow \mathcal{G}_i^{dec}(h_c)$
  else
    $h_c' \leftarrow \text{MLP}_C(h_c)$; $\hat{\mathbf{M}}^{(t)} \leftarrow \mathcal{G}_i^{dec}(h_c')$
  end if
end for

image are used as reference information for the referred medical entity. Subsequently, the LLaVA-Med visual encoder $\mathcal{G}_v^{enc}$ and the bounding box encoder $\mathcal{G}_b^{enc}$ are employed to extract their respective features, which are then concatenated with x_{txt}, $\hat{x}_{txt}$ and x_{img} to form the complete input. This input design enables the model not only to comprehend the current query but also to access key information from the reference round while maintaining awareness of the full conversational history. We then feed the complete input into LLaVA-Med $\mathcal{G}_i$, which in turn outputs a text answer $\hat{y}_{txt}$. It can be formulated as (where x_{img}, x_{txt}, $\hat{x}_{txt}$ also denote their embeddings):

$$\hat{y}_{txt} = \mathcal{G}_i(x_{img}, \mathcal{G}_v^{enc}(x_{crop}), \mathcal{G}_b^{enc}(x_{box}), x_{txt}, \hat{x}_{txt}). \tag{1}$$

The output $\hat{y}_{txt}$ includes the special token [SEG] when the model attempts to segment a specific medical entity within x_{img}. We then extract the last-layer embedding of the LLM corresponding to the [SEG] token, denoted as h_c, which encapsulates the rich semantic representation of the medical entity inferred by the MLLM. The vision backbone $\mathcal{G}_i^{enc}$ first extracts dense visual representations f from the image x_{img}. These features are then integrated with h_c by the decoder $\mathcal{G}_i^{dec}$ to predict the target mask $\hat{\mathbf{M}}$. The process can be formulated as:

$$f = \mathcal{G}_i^{enc}(x_{img}), \quad \hat{\mathbf{M}} = \mathcal{G}_i^{dec}(f, h_c). \tag{2}$$

4.3 Judgment & Correction Mechanism

During the training of MediRound, we adopt teacher forcing [46] and directly use the ground-truth masks to obtain the referred information. While this strategy effectively optimizes the model parameters, it introduces a discrepancy between the training and testing scenarios: *during multi-round evaluation, the trained model can only rely on its own outputs from previous rounds as reference information.* As illustrated in Figure 4, the mask information from the referred round provides essential contextual guidance for the current round.

This dependency introduces a potential limitation: errors from earlier segmentation rounds may accumulate and propagate, ultimately degrading the accuracy of subsequent predictions during evaluation. This effect may become increasingly pronounced as the number of segmentation rounds increases, due to the progressive accumulation and amplification of errors. Therefore, we introduce a lightweight yet effective Judgment & Correction Mechanism that enables the model to transfer higher-quality masks from previous rounds to subsequent ones during evaluation, thereby mitigating the accumulation of segmentation errors. Motivated by prior studies [31, 41] demonstrating that the [SEG] embedding encodes rich target semantics and that semantic drift within this embedding directly degrades mask quality, we intervene at the [SEG] embedding level to mitigate error propagation. Specifically, as illustrated in Figure 5, during the MediRound evaluation process, we employ a **Quality Judgment Module** to assess the quality of the hidden features associated with the output [SEG] token in the current round. If the feature quality is deemed unsatisfactory, the features are then passed to the **Correction Module** for feature refinement; otherwise, they are directly decoded into masks as reference candidates. In this way, when later rounds refer to the current round result, they receive a higher-quality reference mask, which facilitates more coherent cross-round reasoning. To better illustrate how JCM assists MediRound, Algorithm 1 provides the pseudocode. Notably, both the Quality Judgment and Correction Modules are lightweight MLPs, with training details in Sec. 4.4.

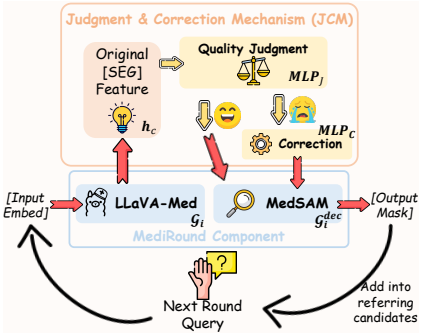

Fig. 5: Illustration of how the Judgment & Correction Mechanism assists MediRound in multi-round reasoning. This mechanism evaluates and optimizes the quality of the [SEG] hidden layer features in each round, effectively preventing current-round errors from being propagated to later dialogues.

4.4 Training and Optimization

Training the MediRound. In end-to-end training for MediRound, we use LoRA to conduct effective fine-tuning on the LLM in $\mathcal{G}_i$, while freezing the vision backbone $\mathcal{G}_i^{enc}$ and LLaVA-Med vision encoder $\mathcal{G}_v^{enc}$. The mask decoder $\mathcal{G}_i^{dec}$ and the bounding box encoder $\mathcal{G}_b^{enc}$ are fully trainable.

For optimization, MediRound is trained through two supervisions. For the text generation objective, we optimize the Auto-Regressive Cross-Entropy loss $\mathcal{L}_{txt}$ between the predicted textual response $\hat{y}_{txt}$ and the ground-truth text answer y_{txt}. For the mask generation, the mask loss $\mathcal{L}_{mask}$ is calculated between the predicted mask $\hat{\mathbf{M}}$ and the ground-truth mask $\mathbf{M}$. The mask loss $\mathcal{L}_{mask}$ is formulated as a weighted combination of the per-pixel Binary Cross-Entropy $\mathcal{L}_{bce}$ and the DICE loss $\mathcal{L}_{dice}$, controlled by the weighting factors λ_{bce} and λ_{dice}.

The overall loss $\mathcal{L}_{MediRound}$ is formulated as:

$$\begin{aligned} \mathcal{L}_{MediRound} &= \mathcal{L}_{txt} + \mathcal{L}_{mask}, \\ \mathcal{L}_{mask} &= \lambda_{bce}\mathcal{L}_{bce} + \lambda_{dice}\mathcal{L}_{dice}. \end{aligned} \tag{3}$$

Training the JCM. The overall training pipeline of JCM is illustrated in Figure 6. As shown in Figure 6, LLaVA-Med takes x_{all} (denoting the full set of inputs) and produces the [SEG] token as output. The hidden feature of [SEG], h_c, is first fed into the Correction Module MLP_C, producing h'_c. The mask decoder $\mathcal{G}_i^{dec}$ then takes h'_c along with the visual features f (extracted by the vision backbone $\mathcal{G}_i^{enc}$) and outputs the predicted mask $\hat{\mathbf{M}}$. In parallel, $\mathcal{G}_i^{dec}$ takes h_c, f and outputs the uncorrected mask $\tilde{\mathbf{M}}$. The process can be formulated as:

$$h'_c = MLP_C(h_c), \quad \hat{\mathbf{M}} = \mathcal{G}_i^{dec}(f, h'_c), \quad \tilde{\mathbf{M}} = \mathcal{G}_i^{dec}(f, h_c). \tag{4}$$

Additionally, we also feed h_c into the Quality Judgment Module MLP_J, which outputs a quality score $q \in [0, 1]$ for h_c:

$$q = MLP_J(h_c), \quad q \in [0, 1] \tag{5}$$

We utilize the end-to-end trained MediRound weights to optimize the Correction Module MLP_C and the Quality Judgment Module MLP_J within the JCM. During the training of JCM, all parameters of the MediRound modules *(e.g., LLaVA-Med, MedSAM, etc.)* are frozen, and only the Correction Module MLP_C and the Quality Judgment Module MLP_J are trainable.

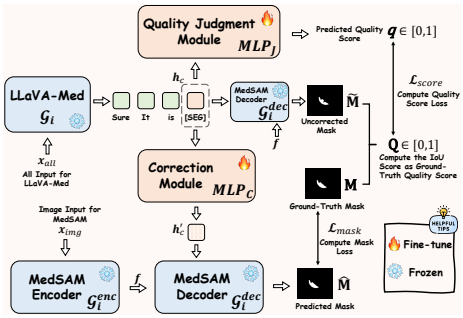

Fig. 6: Overview of the JCM training pipeline.

For optimization, as illustrated in Figure 6, the Correction Module MLP_C and the Quality Judgment Module MLP_J are jointly optimized using the mask loss $\mathcal{L}_{mask}$ and the quality score loss $\mathcal{L}_{score}$. For the quality score loss $\mathcal{L}_{score}$, we compute the Binary Cross-Entropy loss $\mathcal{L}_{bce}^{QJ}$ between the predicted quality score $q \in [0, 1]$ and the ground-truth quality score $\mathbf{Q} \in [0, 1]$. *The ground-truth quality score* $\mathbf{Q}$ *is obtained by computing the Intersection-over-Union (IoU) between the uncorrected mask* $\tilde{\mathbf{M}}$ *and the ground-truth mask* $\mathbf{M}$*, serving as a soft target in the Binary Cross-Entropy loss* $\mathcal{L}_{bce}^{QJ}$. The mask loss $\mathcal{L}_{mask}$ is also calculated between the predicted mask $\hat{\mathbf{M}}$ and the ground-truth mask $\mathbf{M}$, and it follows the same formulation as that used during MediRound training. The overall loss $\mathcal{L}_{JCM}$ can be formulated as:

$$\begin{aligned} \mathcal{L}_{JCM} &= \mathcal{L}_{bce}^{QJ} + \mathcal{L}_{mask}, \\ \mathcal{L}_{mask} &= \lambda_{bce}\mathcal{L}_{bce} + \lambda_{dice}\mathcal{L}_{dice}. \end{aligned} \tag{6}$$

Table 1: Comparison of multi-round entity-level medical reasoning segmentation performance between MediRound and existing relevant methods. *"hard case"* denotes a set of cases that are more challenging for models compared with *"regular case"*. SegLLM is fine-tuned on MR-MedSeg, whereas the other combined baseline methods are evaluated in a zero-shot manner. $\mathcal{T}$ denotes whether the method is an independent model (a single model or with an extension). This table reports the overall scores across all rounds in MR-MedSeg dataset.

Methods	val			test						$\mathcal{T}$
	overall			regular case			hard case			
	Dice	gIoU	cIoU	Dice	gIoU	cIoU	Dice	gIoU	cIoU	
Human-Thinking + BiomedParse [53]	21.1	17.0	12.1	28.6	22.1	19.0	11.3	10.2	8.4	✗
Human-Thinking + MedPLIB [12]	33.7	25.3	24.5	35.0	24.8	24.7	32.0	25.2	23.4	
Human-Thinking + IMIS-Net [5]	30.5	25.9	32.2	33.7	30.0	42.6	26.4	21.0	25.3	
Human-Thinking + MediSee [41]	43.0	33.5	35.4	45.6	34.5	38.6	45.1	37.1	39.7	
GPT-4o [1] + IMIS-Net [5]	19.3	15.8	19.6	23.8	18.7	20.8	17.4	15.2	21.2	✗
Gemini-2.5-Pro [40] + IMIS-Net [5]	26.9	22.6	28.0	33.6	29.5	35.2	21.9	17.0	19.2	
Qwen3-VL [3] + IMIS-Net [5]	26.6	22.8	26.0	25.1	22.3	26.2	23.4	18.3	20.1	
GPT-4o [1] + MediSee [41]	36.8	28.7	29.4	45.2	34.2	38.0	33.4	26.8	28.9	
Gemini-2.5-Pro [40] + MediSee [41]	33.5	26.1	27.6	37.9	31.4	35.1	38.6	28.6	30.8	
Qwen3-VL [3] + MediSee [41]	42.7	34.0	34.5	45.2	34.1	38.0	37.8	31.4	32.8	
SegLLM-7B [44]	36.3	27.5	36.8	42.1	32.9	42.4	30.4	22.2	31.0	✓
SegLLM-13B [44]	38.9	31.2	39.1	45.3	33.4	44.0	33.0	24.6	33.9	
MediRound + READ [31]	53.9	45.1	54.6	60.3	51.4	57.6	47.2	36.8	49.1	✓
MediRound	55.8	46.5	55.8	61.0	52.1	58.9	48.1	38.2	50.2	
MediRound + JCM	**58.4**	**49.0**	**58.9**	**63.3**	**54.5**	**60.2**	**50.3**	**40.5**	**53.4**	

5 Experiments

5.1 Experimental Setting

Network Architecture. For **MediRound**, as described in Sec. 4, we employ LLaVA-Med-v1.5-Mistral-7B [19] as the MLLM, and MedSAM [25] as the vision backbone. The $\mathcal{G}_b^{enc}$ is implemented as a linear projection from 4 dimensions to 4096. For **JCM**, the projections in the Quality Judgment Module and Correction Module are implemented as three-layer MLPs with dimensions [256, 512, 512, 1] and [256, 512, 512, 256], respectively.

Implementation Details. MediRound is trained on the MR-MedSeg using four NVIDIA RTX A6000 GPUs (48 GB memory each) for about 1.5 days. We use a batch size of 15 on each device. To enhance training efficiency, we adopt the DeepSpeed [34] engine. We optimize the model parameters with AdamW [23] using a learning rate of 0.0003. The learning rate is adjusted using the WarmupDecayLR scheduler, which includes a 100-iteration warm-up stage. The LoRA rank is uniformly set to 8. For training **JCM**, we keep the same hyperparameter settings, and it can be completed in just half a day. All training is conducted on the MR-MedSeg training set, employing a teacher forcing strategy [46]. During training, the weights for the Dice loss λ_{dice} and BCE loss λ_{bce} are set to 0.5 and 2.0. All multi-round segmentation evaluations adopt an autoregressive free-running strategy, where subsequent rounds can only rely on the model's own mask outputs from previous rounds.

Table 2: Multi-round entity-level medical reasoning segmentation performance of MediRound and some relevant methods, reported for each round. The table presents the cIoU scores from rounds 2–8 across all samples in the MR-MedSeg validation set.

Methods	Conversation Rounds						
	Round 2	Round 3	Round 4	Round 5	Round 6	Round 7	Round 8
Human-Thinking + BiomedParse [53]	13.3	17.9	13.9	15.9	11.5	9.4	7.8
Human-Thinking + MedPLIB [12]	19.7	21.0	21.3	19.7	20.5	25.0	34.4
Human-Thinking + IMIS-Net [5]	34.0	27.1	30.1	33.4	37.5	37.7	32.8
Human-Thinking + MediSee [41]	43.6	37.0	34.4	31.9	29.7	37.0	25.1
GPT-4o [1] + IMIS-Net [5]	25.1	20.4	18.1	17.0	11.6	12.5	18.9
Gemini-2.5-Pro [40] + IMIS-Net [5]	27.2	28.6	29.4	26.2	22.6	20.1	30.5
Qwen3-VL [3] + IMIS-Net [5]	25.6	24.1	20.3	21.8	15.2	14.8	27.6
GPT-4o [1] + MediSee [41]	37.7	33.4	27.3	25.5	14.9	21.5	20.9
Gemini-2.5-Pro [40] + MediSee [41]	30.8	30.4	29.5	26.7	31.5	21.5	33.3
Qwen3-VL [3] + MediSee [41]	37.7	33.9	30.1	25.4	20.2	34.7	29.9
SegLLM-7B [44]	35.2	39.5	27.2	30.1	34.7	37.4	34.1
SegLLM-13B [44]	36.4	42.1	32.5	33.2	35.3	39.6	36.0
MediRound + READ [31]	49.1	58.3	55.0	56.2	62.4	63.3	44.5
MediRound	50.2	60.8	55.9	58.8	63.7	64.7	46.1
MediRound + JCM	**52.6**	**63.6**	**59.7**	**64.6**	**69.5**	**66.3**	**54.8**

5.2 Multi-Round Experimental Results

The results of multi-round medical reasoning segmentation are summarized in Tables 1 and 2. Table 1 reports the overall performance across all rounds, while Table 2 provides detailed scores for each individual round of interaction. Since existing text-prompt-based medical segmentation methods lack the capability for multi-round entity-level reasoning segmentation, we first incorporate human guidance to enable a broader and more equitable comparison. Specifically, humans with medical expertise quickly review the multi-round questions and historical outputs, infer the target at each round, then write a textual description as the prompt to the segmentation models, allowing these models to perform the task. As shown in Tables 1 and 2, human assistance can substantially improve the model's ability to perform multi-round segmentation. However, the overall performance remains suboptimal. This is primarily because, even with a certain level of medical knowledge, humans may still make errors when dealing with complex multi-round medical reasoning tasks, and their coordination with the model tends to be relatively rigid. In contrast, the proposed MediRound can effectively comprehend the current query while simultaneously integrating historical information and mask results from referred rounds. By achieving this capability, it yields an average improvement of approximately 15% across various evaluation metrics compared to other methods. Notably, despite its lightweight design, the proposed JCM remarkably improves the model's overall performance, particularly as the number of interaction rounds increases (refer to Table 2). This improvement is primarily attributed to the fact that JCM refines suboptimal mask outputs from earlier rounds before they are referred in subsequent ones, thereby effectively mitigating the error accumulation inherent in the chain-like pipeline of multi-round segmentation.

We further construct a hybrid paradigm for comparison, which integrates an MLLM with a medical segmentation model. Specifically, we leverage the strong text–image reasoning capabilities of MLLMs [1, 3, 40] to interpret complex multi-round queries, and subsequently feed their generated reasoning outputs into a medical segmentation model to produce visual predictions. Notably, all comparison methods are also allowed to access the historical mask outputs to ensure a fair comparison. As shown in Tables 1 and 2, despite being supported by powerful MLLMs, these combined approaches still underperform compared with MediRound. We attribute this advantage to our model's end-to-end training on extensive medical image–text pairs, which enables more coherent dynamic alignment between multimodal features than the two-stage combined paradigms. In addition, we include SegLLM [44]—the exploratory work on multi-round reasoning segmentation in natural images—in our comparison. The results indicate that while SegLLM performs well in natural images, it struggles when generalized to medical scenarios.

Table 3: Vanilla medical referring segmentation comparison results.

Methods	Metrics		
	Dice	gIoU	cIoU
Medical Interaction Seg. (bbox input)			
MedSAM [25]	77.7	67.1	81.0
IMIS-Net [5]	77.0	67.5	87.2
Medical Referring Seg.			
Grounding DINO [22] + SAM-Med2D [6]	18.4	12.5	10.9
Grounding DINO [22] + MedSAM [25]	18.5	11.0	13.5
BiomedParse [53]	24.9	20.1	14.5
MedPLIB [12]	40.6	39.1	45.2
IMIS-Net [5]	41.3	35.5	65.8
MediSee [41]	61.2	**50.4**	**70.8**
MediRound + READ [31]	60.7	48.0	65.2
MediRound	**62.1**	48.3	66.3

Table 4: Ablation study on the choices of vision and MLLM backbones. Experiments are conducted on the MR-MedSeg val set.

Settings	Dice	gIoU	cIoU
Seg. Head Backbone			
SAM	46.7	33.1	37.2
SAM-Med2D	53.9	45.3	55.5
MedSAM (LoRA)	54.6	44.9	54.4
MedSAM	55.8	46.5	55.8
MLLM Backbone			
LLaVA-v1.5 (LoRA)	53.4	42.1	53.3
LLaVA-Med (Frozen)	48.0	39.2	49.6
LLaVA-Med (LoRA)	55.8	46.5	55.8

Table 5: Ablation study on the effects of different strategies for extracting reference-mask information. The experimental results are reported on the val set of MR-MedSeg. $\mathcal{C}$: Cropped image input; $\mathcal{B}$: Bounding box input.

$\mathcal{C}$	$\mathcal{B}$	Dice	gIoU	cIoU
✗	✗	41.9	32.5	38.4
✓	✗	55.4	45.8	56.0
✓	✓	55.8	46.5	55.8

5.3 Vanilla Medical Referring Segmentation

To demonstrate MediRound's effectiveness in conventional single-round referring medical segmentation, we evaluate it against several prior related methods on the SA-Med2D-20M [51] benchmark. As summarized in Table 3, our model achieves strong performance, remaining highly competitive in traditional medical image

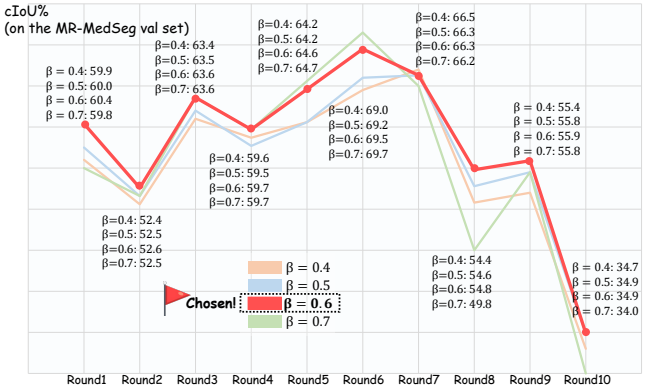

Fig. 7: Ablation study on the threshold β value in JCM.

segmentation. For context, we also report the segmentation results of IMIS-Net [5] and MedSAM [25] using conventional bounding box guidance.

5.4 Ablation Study

Choices of β in JCM. We investigate the impact of the threshold β in the Quality Judgment Module of JCM. As shown in Figure 7, the collaboration between JCM and MediRound achieves the best performance when $\boldsymbol{\beta = 0.6}$. This is likely because a smaller β causes many suboptimal results to lose the opportunity for refinement, while an excessively large β may instead disturb some high-quality predictions. *For the definition and role of β, see Algorithm 1.*
Choices of the Main Components. Table 4 summarizes the effects of different foundational MLLMs and segmentation head backbones. The results indicate that **LLaVA-Med** achieves superior performance owing to its strong multimodal alignment capability in medical contexts, while the **MedSAM** performs best as the segmentation head, benefiting from large-scale medical image pretraining.
Choices of Reference Information Strategies. We further examine the impact of different strategies for extracting information from the reference mask. Table 5 indicates that using both the cropped image and the bounding box as reference-mask information yields the best performance.

5.5 Qualitative Analysis

We present qualitative comparisons in Figure 8. Existing methods largely fail in multi-round reasoning for medical segmentation, whereas MediRound can understand the current query and dialogue history for accurate segmentation.

5.6 Human Evaluation

We recruited 15 medical students to rate our proposed multi-round entity-level medical reasoning segmentation task (MEMR-Seg) and its baseline, MediRound,

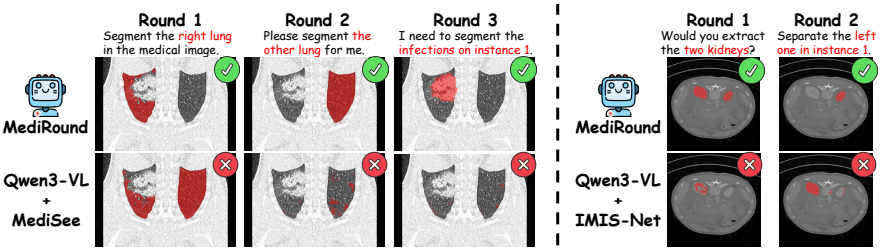

Fig. 8: Qualitative comparison with different methods.

on four criteria using a 5-point scale (1–5). The results in Figure 9 demonstrate the task and model's meaningfulness and practical value in medical education.

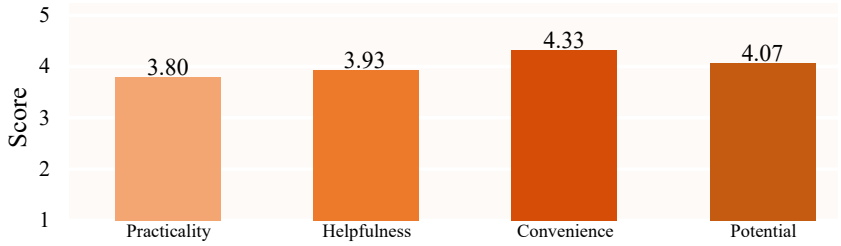

Fig. 9: Human evaluation on MEMR-Seg and MediRound (1–5 scores). This experiment is designed to demonstrate the potential value of our work in medical education scenarios.

6 Conclusion

This work introduces MEMR-Seg, a new task that advances medical image segmentation toward interactive, multi-round reasoning. By constructing the large-scale MR-MedSeg dataset and developing the MediRound baseline with a helpful Judgment & Correction Mechanism, this work demonstrates the feasibility and effectiveness of medical multi-round entity-level reasoning, providing insights for future research on interactive medical segmentation.

Acknowledgements

This work was supported by Ministry of Education Tier 2 grant, Singapore (T2EP20224-0028), and Ministry of Education Tier 1 grant, Singapore (23-0651-P0001). This work was also supported by National Natural Science Foundation of China under Grant No.62501532; Zhejiang Provincial Natural Science Foundation of China under Grant No.LQN26F020039.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., et al.: The medical segmentation decathlon. Nature communications **13**(1), 4128 (2022)
3. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
4. 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 preprint arXiv:2107.02314 (2021)
5. Cheng, J., Fu, B., Ye, J., Wang, G., Li, T., Wang, H., Li, R., Yao, H., Cheng, J., Li, J., et al.: Interactive medical image segmentation: A benchmark dataset and baseline. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 20841–20851 (2025)
6. Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., et al.: Sam-med2d. arXiv preprint arXiv:2308.16184 (2023)
7. Drozdzal, M., Vorontsov, E., Chartrand, G., Kadoury, S., Pal, C.: The importance of skip connections in biomedical image segmentation. In: International workshop on deep learning in medical image analysis. pp. 179–187. Springer (2016)
8. 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)
9. Heller, N., Isensee, F., Trofimova, D., Tejpaul, R., Zhao, Z., Chen, H., Wang, L., Golts, A., Khapun, D., Shats, D., et al.: The kits21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase ct. arXiv preprint arXiv:2307.01984 (2023)
10. Hssayeni, M.D., Croock, M.S., Salman, A.D., Al-Khafaji, H.F., Yahya, Z.A., Ghoraani, B.: Intracranial hemorrhage segmentation using a deep convolutional model. Data **5**(1), 14 (2020)
11. Huang, X., Li, H., Cao, M., Chen, L., You, C., An, D.: Cross-modal conditioned reconstruction for language-guided medical image segmentation. IEEE Transactions on Medical Imaging **44**(4), 1821–1835 (2024)
12. Huang, X., Shen, L., Liu, J., Shang, F., Li, H., Huang, H., Yang, Y.: Towards a multimodal large language model with pixel-level insight for biomedicine. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 3779–3787 (2025)
13. Huang, Y., Peng, Z., Zhao, Y., Yang, P., Yang, X., Shen, W.: Medseg-r: Reasoning segmentation in medical images with multimodal large language models. arXiv preprint arXiv:2506.10465 (2025)
14. 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)
15. Jaffray, D.A., Knaul, F., Baumann, M., Gospodarowicz, M.: Harnessing progress in radiotherapy for global cancer control. Nature cancer **4**(9), 1228–1238 (2023)
16. Jang, D., Cho, Y., Lee, S., Kim, T., Kim, D.S.: Mmr: A large-scale benchmark dataset for multi-target and multi-granularity reasoning segmentation. arXiv preprint arXiv:2503.13881 (2025)

17. 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)
18. 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)
19. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36**, 28541–28564 (2023)
20. 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)
21. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
22. 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)
23. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
24. Lüddecke, T., Ecker, A.: Image segmentation using text and image prompts. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7086–7096 (2022)
25. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature communications **15**(1), 654 (2024)
26. 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)
27. Nauffal, V., Klarqvist, M.D., Hill, M.C., Pace, D.F., Di Achille, P., Choi, S.H., Rämö, J.T., Pirruccello, J.P., Singh, P., Kany, S., et al.: Noninvasive assessment of organ-specific and shared pathways in multi-organ fibrosis using t1 mapping. Nature medicine **30**(6), 1749–1760 (2024)
28. Nie, H., Zheng, Y., Ye, Y., Xie, B.: Mturn-seg: A large-scale bilingual medical benchmark for multi-turn reasoning segmentation. In: 2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). pp. 3946–3950. IEEE (2025)
29. Norouzi, A., Rahim, M.S.M., Altameem, A., Saba, T., Rad, A.E., Rehman, A., Uddin, M.: Medical image segmentation methods, algorithms, and applications. IETE Technical Review **31**(3), 199–213 (2014)
30. Pi, R., Yao, L., Gao, J., Zhang, J., Zhang, T.: Perceptiongpt: Effectively fusing visual perception into llm. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 27124–27133 (2024)
31. 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)
32. Ramesh, K., Kumar, G.K., Swapna, K., Datta, D., Rajest, S.S.: A review of medical image segmentation algorithms. EAI Endorsed Transactions on Pervasive Health & Technology **7**(27) (2021)

33. Rasheed, H., Maaz, M., Shaji, S., Shaker, A., Khan, S., Cholakkal, H., Anwer, R.M., Xing, E., Yang, M.H., Khan, F.S.: Glamm: Pixel grounding large multimodal model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13009–13018 (2024)
34. 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)
35. 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)
36. Rudyanto, R.D., Kerkstra, S., Van Rikxoort, E.M., Fetita, C., Brillet, P.Y., Lefevre, C., Xue, W., Zhu, X., Liang, J., Öksüz, I., et al.: Comparing algorithms for automated vessel segmentation in computed tomography scans of the lung: the vessel12 study. Medical image analysis **18**(7), 1217–1232 (2014)
37. Saha, A., Hosseinzadeh, M., Huisman, H.: End-to-end prostate cancer detection in bpmri via 3d cnns: Effects of attention mechanisms, clinical priori and decoupled false positive reduction. Medical image analysis **73**, 102155 (2021)
38. Setio, A.A.A., Traverso, A., De Bel, T., Berens, M.S., Van Den Bogaard, C., Cerello, P., Chen, H., Dou, Q., Fantacci, M.E., Geurts, B., et al.: Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge. Medical image analysis **42**, 1–13 (2017)
39. 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)
40. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
41. Tong, Q., Lu, Z., Liu, J., Zheng, Y., Lu, Z.M.: Medisee: Reasoning-based pixel-level perception in medical images. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 2742–2751 (2025)
42. Wang, J., Ke, L.: Llm-seg: Bridging image segmentation and large language model reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1765–1774 (2024)
43. 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)
44. Wang, X., Zhang, S., Li, S., Kallidromitis, K., Li, K., Kato, Y., Kozuka, K., Darrell, T.: Segllm: Multi-round reasoning segmentation. arXiv preprint arXiv:2410.18923 (2024)
45. Wang, Y., Zhao, L., Wang, M., Song, Z.: Organ at risk segmentation in head and neck ct images using a two-stage segmentation framework based on 3d u-net. IEEE Access **7**, 144591–144602 (2019)
46. Williams, R.J., Zipser, D.: A learning algorithm for continually running fully recurrent neural networks. Neural computation **1**(2), 270–280 (1989)
47. Xing, Y., Wu, J., Ozdemir, S., Zhang, Y., Yang, Y., Shao, W., Gong, K.: Medvlsam2: A unified 3d medical vision-language model for multimodal reasoning and prompt-driven segmentation. arXiv preprint arXiv:2601.09879 (2026)

48. Yan, K., Wang, X., Lu, L., Summers, R.M.: Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. Journal of medical imaging **5**(3), 036501–036501 (2018)
49. Yan, Z., Diao, M., Yang, Y., Jing, R., Xu, J., Zhang, K., Yang, L., Liu, Y., Liang, K., Ma, Z.: Medreasoner: Reinforcement learning drives reasoning grounding from clinical thought to pixel-level precision. arXiv preprint arXiv:2508.08177 (2025)
50. 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)
51. Ye, J., Cheng, J., Chen, J., Deng, Z., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., et al.: Sa-med2d-20m dataset: Segment anything in 2d medical imaging with 20 million masks. arXiv preprint arXiv:2311.11969 (2023)
52. Zhang, A., Yao, Y., Ji, W., Liu, Z., Chua, T.S.: Next-chat: An lmm for chat, detection and segmentation. arXiv preprint arXiv:2311.04498 (2023)
53. 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)
54. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: International workshop on deep learning in medical image analysis. pp. 3–11. Springer (2018)