This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

IBISAgent: Reinforcing Pixel-Level Visual Reasoning in MLLMs for Universal Biomedical Object Referring and Segmentation

Yankai Jiang[1,2★], Qiaoru Li[1*], Binlu Xu[1*], Haoran Sun[2], Chao Ding[2], Junting Dong[2], **Yuxiang Cai**[1,3,4✉], Xuhong Zhang[1,3,4], Jianwei Yin[1]

[1]Zhejiang University [2]Shanghai Artificial Intelligence Laboratory
[3]Ningbo Global Innovation Center, Zhejiang University
[4]Zhejiang Key Laboratory of Digital-Intelligence Service Technology

jiangyankai@pjlab.org.cn, caiyuxiang@zju.edu.cn

Abstract

Recent research on medical MLLMs has shifted its focus from image-level understanding to fine-grained, pixel-level comprehension. Although segmentation serves as the foundation for pixel-level understanding, existing approaches face two major challenges. First, they introduce implicit segmentation tokens and require joint fine-tuning of the MLLM and external pixel decoders, increasing the risk of catastrophic forgetting and limiting out-of-domain generalization. Second, most methods rely on single-pass reasoning and lack the ability to iteratively refine segmentation results. To overcome these limitations, we propose IBISAgent—a novel agentic MLLM that reformulates segmentation as a vision-centric, multi-step decision-making process. IBISAgent enables MLLMs to generate interleaved reasoning and text-based click actions, invoke segmentation tools, and produce high-quality masks without architectural modifications. We also design a two-stage training framework consisting of cold-start SFT and agentic RL with tailored, fine-grained rewards. Through iterative multi-turn visual reasoning, IBISAgent naturally facilitates mask refinement and enhances robustness in complex medical referring and reasoning segmentation tasks. Extensive experiments demonstrate that IBISAgent consistently outperforms both closed-source and open-source SOTA methods.

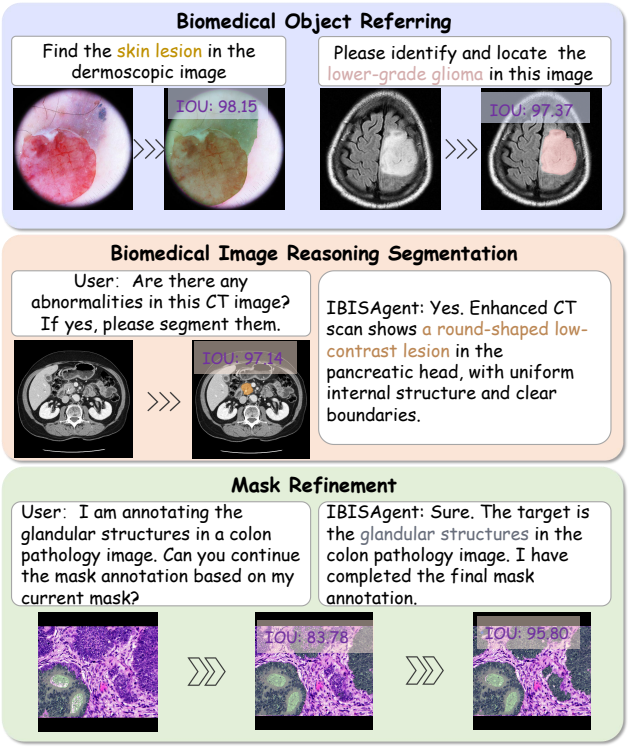


Figure 1. IBISAgent flexibly supports a wide range of fine-grained biomedical image understanding tasks, including referring and reasoning segmentation. It also handles a novel mask-refinement task that assists annotators in completing partially labeled masks.

1. Introduction

Multimodal large language models (MLLMs) has yielded notable advancements in developing powerful medical AI assistants [4, 21, 33, 43]. These models now achieve high scores on exam-style medical question-answering benchmarks [8, 16, 19, 29]. However, daily clinical diagnosis is far more complex than a single structured question-and-answer interaction, since holistic medical image analysis comprises multiple subtasks, such as segmentation and detection of biomedical objects [46]. Prevailing medical MLLMs focus on visual question answering (VQA) and are

*Equal contribution. ★Project leader. ✉Corresponding authors. The codes and trained models are available at https://github.com/Yankai96/IBISAgent.

largely confined to image-level understanding tasks, thus failing to achieve fine-grained pixel-level comprehension.

Recently, increasing efforts have been devoted to equipping MLLMs with pixel-level reasoning segmentation capabilities [3, 18, 23, 32, 45]. These models typically introduce additional task-specific segmentation tokens (*e.g.*, <SEG>), which are decoded through external pixel decoders to generate segmentation masks. Building on this foundation, many studies [1, 13, 38, 41] have adopted similar strategies to adapt MLLMs for medical image segmentation. Despite their effectiveness, the segmentation capabilities of existing medical MLLMs are still limited. Their dependence on joint fine-tuning of the MLLM and external pixel decoders heightens the risk of *catastrophic forgetting*, resulting in strong in-domain performance but weak cross-domain generalization. Moreover, the introduction of implicit segmentation tokens disrupts the MLLM's natural text output space, thereby weakening its reasoning ability and failing to reflect the model's intrinsic pixel-level understanding [35].

These limitations motivate a re-examination of how to better elicit pixel-level visual reasoning in MLLMs for biomedical image segmentation (BIS). Unlike natural images, biomedical images often exhibit subtle and complex visual semantics, such as faint lesion cues and nuanced pathological patterns. A single forward pass for segmentation is often inadequate. In contrast, human experts typically perform segmentation in a multi-step, interactive manner. For instance, annotators iteratively refine masks through positive and negative clicks using interactive segmentation tools [17, 26, 30]. It is natural to ask whether current MLLMs can observe an image multiple times, re-evaluate their intermediate decisions, and adapt to feedback to perform self-evolving segmentation—thereby emulating the strategies and interactive behaviors of human annotators through the use of segmentation tools.

Therefore, we propose IBISAgent, a novel agentic MLLM that reformulates segmentation as a multi-step Markov Decision Process. IBISAgent decouples pixel-level visual grounding and mask prediction. It iteratively generates interleaved reasoning and text-based click commands, invokes segmentation tools, and refines the current segmentation based on evolving visual features. Compared with previous medical MLLMs that rely on implicit tokens and additional pixel decoders, IBISAgent preserves the LLM's inherent internal language representations and extends segmentation beyond mere pixel prediction to encompass fine-grained visual reasoning and action planning. This design facilitates vision-centric, multi-step decision-making and enables genuine reasoning to support advanced tasks such as automatic mask refinement (Fig. 1). Moreover, by treating segmentation models as plug-and-play tools controllable through language, IBISAgent enhances flexibility and extensibility by eliminating rigidly defined input–output templates (*e.g.*, "It's <SEG>."), thereby facilitating the implementation of a unified framework across diverse tasks.

To develop IBISAgent, we first employ an automated iterative click-simulation algorithm to transform existing BIS datasets into annotation trajectories. Then, we construct a high-quality dataset comprising $400K+$ samples, each annotated with step-wise reasoning traces, click action trajectories, and corresponding masks. We further propose a two-stage training protocol. First, we perform cold-start supervised fine-tuning (SFT) to enable IBISAgent to learn visual reasoning and plan click actions according to user instructions. Then we employ reinforcement learning (RL) with novel fine-grained rewards—particularly, a region-based click placement reward and a progressive segmentation improvement reward—to further enhance decision-making, allowing the model to autonomously discover efficient and advanced action policies rather than merely imitating the click trajectories learned during the SFT stage.

We conduct a comprehensive evaluation on multiple benchmarks, spanning both in-domain and zero-shot scenarios, to rigorously assess the performance of IBISAgent. Results show that our model exhibits strong pixel-level visual reasoning abilities and significantly exceeds the SOTA MLLMs. Our contributions can be summarized as follows:

- We present IBISAgent, a novel agentic framework that equips MLLMs with fine-grained pixel-level visual reasoning, enabling high-quality segmentation without architectural modifications or implicit tokens.
- We construct a comprehensive dataset and introduce effective training strategies—cold-start SFT and RL with tailored fine-grained rewards—to foster strong decision-making and advanced action planning.
- We conduct extensive held-in and held-out experiments to evaluate IBISAgent, and the results verify its superiority on biomedical object referring and segmentation tasks.

2. Related Work

Pixel-Level Understanding in Medical MLLMs. Recent advancements [12, 40, 41] in medical MLLMs have increasingly focused on enhancing the models' fine-grained, pixel-level understanding to enable accurate detection and segmentation of biomedical structures. Inspired by the pioneering general-domain model LISA [18], which introduced a "reasoning segmentation" task to enable models to parse complex, implicit text queries and generate corresponding masks, numerous recent studies [12, 14, 39, 40] have adapted and extended this paradigm to the medical domain to address its unique challenges. These LISA-style MLLMs follow an innovative "embedding-as-mask" paradigm, in which the hidden-state embeddings of a special <SEG> token are projected by the LLM and subsequently decoded into segmentation masks via a vision decoder. While promising, these methods require MLLMs to

learn task-specific implicit tokens and undergo additional joint fine-tuning with a segmentation decoder. This process disrupts the MLLM's original text output space and increases the risk of catastrophic forgetting, thereby compromising semantic generalization by deviating from language-based outputs. Moreover, existing methods are limited to single-turn reasoning and grounding, lacking an inherent mechanism for autonomous, self-evolving, multi-step refinement of mask predictions. These challenges represent the key issues our work aims to address.

MLLMs with Segmentation Tools. Recent studies [11, 20, 24, 25] have explored activating the intrinsic pixel-level understanding capabilities of MLLMs via RL, enabling the models to generate bounding boxes or point prompts that precisely localize target regions. These spatial coordinates are subsequently passed to SAM as prompts to generate the corresponding segmentation masks. However, these methods are constrained to single-turn reasoning and grounding. In complex real-world scenarios, MLLMs often struggle to precisely localize target regions in a single step. In contrast, we reformulate segmentation as a multi-step Markov Decision Process, enabling the MLLM to perform iterative mask refinement and exhibit self-reflective behavior—capabilities absent in prior work. This formulation effectively mitigates error accumulation and substantially improves performance. Furthermore, during the RL process, we design step-wise rewards that provide segmentation-quality–guided feedback at each iteration, further enhancing the agent's stability and generalization ability.

3. Method

IBISAgent is a unified multimodal agent capable of pixel-space reasoning for biomedical image segmentation and mask refinement by adaptively generating spatial prompts and invoking segmentation tools. The ability is inherited from the model's native capability of visual grounding and action planning, and further incentivized and enhanced via end-to-end SFT (Sec. 3.2) and RL training (Sec. 3.3).

3.1. Architecture Overview.

As illustrated in Fig. 2, given a user question Q and an input image I, the agent generates a multi-step, interleaved reasoning path P to derive the final answer. Unlike the predominant textual reasoning paradigm, this pixel-space reasoning path P can be represented as an T-step chain: $P = \{(r_t, a_t, o_t)\}_{t=1}^{T}$, where each step comprises textual thinking r_t, an action a_t which refers to spatial click operations, and the resulting segmentation observation o_t generated based on a_t. This iterative thought-action-observation loop continues until the model reaches a conclusive answer or when predefined limits on context length or interaction turns are reached. The core components are detailed below:

- Textual Thinking: The internal reasoning process used by the policy model to select the next action, conditioned on the interaction history and the current observation.
- Action: The action space comprises two options: (1) click operations and (2) emitting a final answer. For click operations, each action is parameterized by three components: `Target`, which specifies the class name of the current segmentation objective; `Attribute` $\in \{+1, -1\}$, which indicates whether the click is positive or negative; and `Coordinate_2d` $\in [0,1]^2$: represents the relative position of the point within the image, scaled to the $[0, 1]$ range for both x and y axes. Note that our method supports performing click operations on multiple targets at once. That is, in the action, it can contain multiple triples consisting of `Target`, `Attribute`, and `Coordinate_2d`. For clarity, in the following exposition, we focus on the single-target setting as a representative example to describe the method.
- Observation: The observation produced by executing a_t in the environment. Concretely, it corresponds to the segmentation mask generated by a segmentation tool (*e.g.*, MedSAM2 [27]) given the inputs of the click prompts in a_t, the previous step's output mask M_{t-1} as a spatial prior, and the original image. The initial observation o_0 is an empty mask, whereas the final observation represents the optimal target mask. This observation is appended to the interaction history and fed back to the model.

Rollout Formulation. Our objective is to train a policy π_θ that emulates the annotation behavior of human experts using an interactive segmentation tool. The policy iteratively generates textual reasoning with a series of positive and negative click points conditioned on the current image I and the evolving mask M, progressively refining the segmentation to achieve a high-quality result. At each step t, the policy is defined as:

$$r_{t+1}, a_{t+1} \sim \pi_\theta(\cdot \mid I, Q, P_{<t})\,. \tag{1}$$

We instruct the model to mark its textual thinking, action, and final answers in the output using the special tokens `<think>`, `<action>`, and `<answer>`. When the model output includes `<action>`, we automatically parse the action into a format compatible with the interactive segmentation model. Subsequently, all previous actions $a_{0:t}$, the current mask M_t, and the original image I are fed into the interactive segmentation model F_{seg}, which generates the updated mask M_{t+1}. Furthermore, we overlay M_{t+1} as a semi-transparent mask on the original image I to create a new composite image o_{t+1}. This image o_{t+1} are then inserted into the `<obs>` field and appended to the ongoing trajectory, serving as the input for the MLLM in next step, allowing the model to simultaneously perceive information about M_{t+1} and I in a single frame. This iterative reasoning process continues until the model determines that the segmentation has reached satisfactory quality. It then outputs

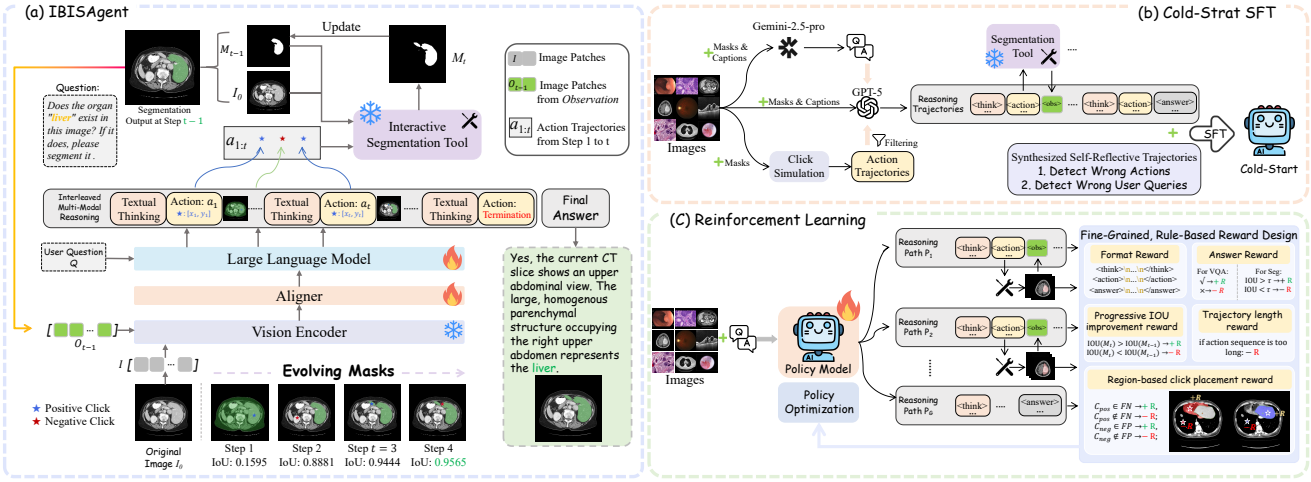


Figure 2. **Overview of IBISAgent.** (a) Overall architecture of the agent; (b) illustration of the cold-start SFT training process; and (c) illustration of the RL training process.

a designated end token in <action> and generates the final <answer>. The system and user prompts are provided in Appendix E. Based on this formulation, IBISAgent supports flexible usage scenarios, including both from-scratch segmentation and refinement of pre-existing masks.

3.2. Cold-Start Supervised Fine-Tuning

We aim to cultivate a novel pixel-space reasoning paradigm within IBISAgent. Pure prompting, however, is insufficient to endow multimodal agents with the precision and robustness required to reliably perform iterative visual operations in real-world medical scenarios. To establish a strong foundation for subsequent reinforcement learning, we first initialize IBISAgent's fine-grained pixel understanding and mask-refinement capability through supervised fine-tuning (SFT) on a cold-start dataset $\mathcal{D}_{\text{cold}}$, which provides ground-truth reasoning and click-annotation trajectories.

Collect Seed Datasets. Existing biomedical segmentation datasets contain only final masks without the step-by-step annotation process. Collecting such data by re-hiring human annotators would incur substantial cost. This motivates us to explore whether trajectory supervision can be automatically derived from existing mask annotations using a rule-based procedure. To this end, we utilize the large-scale BiomedParseData [46], containing 3.4 million image–mask–label tuples that cover 82 biomedical object types across 9 imaging modalities. The dataset provides diverse, high-quality masks that supply rich pixel-level details for fine-grained analysis and serve as reliable supervision for synthesizing click-based annotation trajectories.

Trajectory Generation. Inspired by prior research on interactive segmentation, we adopt the click-simulation strategy F_{cs} proposed in [42]. Given the current mask M_t and the GT mask M_{gt}, F_{cs} outputs the next click action $a_{t+1} = F_{\text{cs}}(M_t, M_{\text{gt}})$. Specifically, this function computes the false positive and false negative regions between the current mask M_t and the GT mask M_{gt}, placing the next click action at the center of the error region based on the size and position of these regions. Based on F_{sim}, we can simulate high-quality trajectories $[M_0, a_0, M_1, a_1, \ldots, M_T, a_T]$. The full trajectory generation algorithm and an illustrative example are presented in the Appendix B.1.

Question, Answer, and Reasoning Generation. We first filter the data based on trajectory quality by removing (1) overly long trajectories and (2) those whose final mask Dice score falls below a predefined threshold. For the remaining samples, we prompt Gemini-2.5-Pro [6] to generate question–answer pairs conditioned on each image, its GT mask, and the corresponding mask description. This process produces fine-grained queries that explicitly focus on grounding and segmenting the specified mask regions. The question set Q includes diverse query types, ranging from those that explicitly specify the segmentation target to those that require the model to first reason over the image and adaptively identify the target region on its own. The prompts for QA generation and examples are provided in Appendix B.2. We further use GPT-5 [28] to synthesize reasoning for each click action, conditioning on the QA, the correct next action, and pixel-level TP/FP/FN information of the current mask. All generated reasoning traces are post-filtered for format and correctness by human annotations. More details are provided in Appendix B.3.

Reflective Behavior Synthesis. In complex scenarios, if the model cannot backtrack or undo previous actions, an inaccurate decision at any step may propagate and negatively impact subsequent predictions, ultimately degrading the final segmentation quality. To enhance robustness, we synthesize additional samples with self-reflection trajectories, covering two types of error correction: (1) Self-correction,

where the agent detects a wrong action, reverts to the previous state, and re-reasons over the interaction history to produce the correct action; and (2) User inconsistency correction, where, in mask-refinement scenarios, if the segmentation target described in the instruction does not align with the initial mask, the agent first discards the erroneous mask and re-segments according to the user instruction.

Cold-Start Training Objective. Through systematic curation, we obtain a dataset $\mathcal{D}_{\text{cold}}$ containing 456K samples, including both gold-standard and error-induced self-correction trajectories for cold-start SFT. The training objective is to minimize the average negative log-likelihood over all reasoning and action tokens. We employ the standard SFT loss for training. Specifically, we apply loss masks to tokens corresponding to segmentation outputs from executed actions as well as to designated erroneous actions within the self-correction trajectories. Masking the erroneous actions prevents the policy from learning to execute the incorrect actions. Our SFT strategy endows the model with strong pixel-level reasoning and self-reflection abilities, providing a solid foundation for subsequent RL.

3.3. Agentic Reinforcement Learning

We further optimize IBISAgent through RL with carefully designed rewards, enabling it to adaptively discover new action strategies and achieve higher-level decision-making, thereby moving beyond the constraints of mimicking the action trajectories learned during SFT.

Dataset Curation. Unlike $\mathcal{D}_{\text{cold}}$, the RL training data includes only images, GT masks, and QA pairs, without click trajectories or reasoning trace annotations. This design encourages the model to autonomously explore and strengthen its pixel-level reasoning ability during RL.

Specifically, we randomly sample image–mask pairs from the BiomedParseData [46]. Following the same QA generation procedure as in the SFT stage, we obtain 564K VQA instances. In addition, we incorporate widely used biomedical VQA datasets that do not require fine-grained pixel-level reasoning or segmentation. This hybrid data composition enables IBISAgent to selectively activate pixel-space reasoning only when necessary. In total, 888K VQA samples are used for RL training, forming the dataset $\mathcal{D}_{\text{rl}}$. More details of $\mathcal{D}_{\text{rl}}$ are in the Appendix A.2.

Reward Design. Unlike prior works [11, 24, 25, 44] that rely on overly simplified, outcome-only reward designs, we introduce a novel, fine-grained, rule-based reward framework that delivers dense feedback throughout the reasoning process. This enables the model to develop more efficient and effective decision-making strategies. Formally, the reward framework consists of the following components:

- Format reward $\mathcal{S}_{\text{format}}$, which evaluates the structural validity of the model's output R. It checks whether all required special tokens appear in the correct order and whether the `<action>` fields can be successfully parsed according to the predefined schema.
- Final-answer reward $\mathcal{S}_{ans}$, which encompasses multiple task types. For close-ended QA questions, we simply check the exact match between the predicted and answers. For segmentation task, we compute the IoU between the predicted masks and GT masks and assign piecewise rewards based on predefined IoU thresholds.
- **Region-based click placement reward** $\mathcal{S}_{\text{click}}$ is defined as a bonus granted only when the model produces a reasonable click action. Specifically, given the model-predicted click a_t, we use the segmentation tool to generate the corresponding mask M_t and compute the false-positive (FP) and false-negative (FN) regions between M_t and the GT mask M_{gt}. A positive click is expected to fall within an FN region, while a negative click should lie within an FP region. Rewards and penalties are assigned accordingly, encouraging the model to place clicks in semantically meaningful locations rather than arbitrarily.
- **Progressive segmentation improvement reward** $\mathcal{S}_{\text{pseg}}$. This reward enforces that each action a_t leads to a segmentation improvement over the previous step. In other words, the mask produced after executing a_t must achieve a higher quality than the mask at step $t-1$. This mechanism encourages the agent to consistently refine the segmentation rather than performing redundant actions or oscillating among repetitive operations. Concretely, we compute the IoU of the generated mask at each step t; if the score surpasses that of the mask from the previous action a_{t-1}, the agent receives a reward; otherwise, no reward is given.
- **Trajectory length reward** $\mathcal{S}_{\text{len}}$. If the action sequence to complete segmentation is shorter than a predefined threshold, a reward is given; otherwise, penalties increasing with trajectory length are applied to encourage efficiency.

The final reward $\mathcal{S}$ is derived as: $\mathcal{S} = \frac{1}{5}(\mathcal{S}_{ans} + \mathcal{S}_{format} + \mathcal{S}_{\text{click}} + \mathcal{S}_{\text{pseg}} + \mathcal{S}_{\text{len}})$. Formal equations for each reward component are provided in Appendix C. Our fine-grained reward scheme better reflects the complexity of iterative segmentation, guiding the model to produce actions that are both spatially valid and semantically accurate.

Optimization. Based on the rollout formulation and rewards defined above, we optimize the policy using GRPO [7] without the KL penalty term [9] on dataset $\mathcal{D}_{rl}$:

$$\mathcal{L}_{\text{RL}} = \mathbb{E}_{\substack{(I,Q,A)\sim\mathcal{D}_{\text{rl}} \\ \{P_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|I,Q)}} \Big(-\frac{1}{G}\sum_{i=1}^{G}\frac{1}{N_i}\sum_{t=1}^{T_i}\min\big(\pi_{\theta_{i,t}}\mathcal{A}_i, \text{clip}(\pi_{\theta_{i,t}}, 1-\epsilon, 1+\epsilon)\,\mathcal{A}_i\big)\Big), \tag{2}$$

$$\pi_{\theta_{i,t}} = \frac{\pi_\theta\left(r_{i,t}, a_{i,t} \mid I, Q, P_{i,<t}\right)}{\pi_{\theta_{\text{old}}}\left(r_{i,t}, a_{i,t} \mid I, Q, P_{i,<t}\right)}. \tag{3}$$

Here, G is the number of rollout reasoning paths; $P_i = \{(r_{i,t}, a_{i,t}, o_{i,t})\}_{t=1}^{T_i}$ denotes the i-th reasoning path; N_i is the total length of P_i excluding observation tokens; $\mathcal{S}_i$ is

Benchmarks \ Methods	General-Purpose MLLMs with Segmentation Capability				Medical MLLMs with Segmentation Capability				Ours
	LISA [18]	LISA++ [45]	SAM4MLLM [5]	VisionReasoner [25]	MedPLIB [13]	Citrus-V [40]	UniBiomed [41]	MMedAgent [20]	IBISAgent
In-domain testset $\mathcal{D}_{\text{test}}$									
IOU ↑	9.44 (20.46)	9.49 (20.76)	15.85 (27.84)	16.11 (29.11)	22.29	30.61	50.74	36.13	**85.58**
DSC ↑	14.11 (25.73)	14.30 (25.94)	21.16 (33.04)	22.05 (35.50)	27.35	37.63	58.31	42.85	**92.21**
F1-score ↑	20.18 (32.15)	20.75 (32.34)	32.53 (42.75)	34.78 (46.72)	38.94	53.75	69.22	56.64	**96.39**
Out-of-domain testset MeCOVQA-G+									
IOU ↑	10.07 (15.24)	9.87 (15.01)	16.99 (21.19)	18.27 (24.46)	33.36	46.54	24.88	26.54	**80.63**
DSC ↑	15.44 (21.30)	14.70 (21.26)	21.85 (26.35)	25.08 (30.24)	41.19	52.65	31.74	33.81	**89.27**
F1-score ↑	21.69 (28.04)	21.25 (27.96)	32.94 (38.57)	37.83 (42.08)	53.47	69.84	43.63	44.17	**95.24**
Held-out in-house testset									
IOU ↑	5.23 (9.12)	5.46 (9.45)	8.28 (14.00)	10.10 (17.66)	20.12	32.08	35.62	27.39	**72.09**
DSC ↑	9.58 (14.33)	9.69 (14.80)	13.59 (18.04)	15.88 (24.57)	27.80	38.63	41.55	34.26	**83.78**
F1-score ↑	13.03 (17.15)	13.17 (17.72)	19.07 (25.26)	22.49 (30.07)	39.42	50.76	54.97	45.88	**91.76**

Table 1. Comparison of segmentation performance on both in-domain and out-of-domain benchmarks. LISA [18], LISA++ [45], VisionReasoner [25], and SAM4MLLM [5] are re-implemented using their official codebases. We evaluate two settings: (1) directly loading their publicly released model weights (shown in blue); and (2) further fine-tuning these models on our SFT and RL training datasets, which include the same images, masks, and QA pairs used for IBISAgent (shown in "()"), ensuring fair comparison. MedPLIB [13], Citrus-V [40], MMedAgent [20], and UniBiomed [41] are also re-implemented following their official repositories. For these models, we directly load the released weights without additional fine-tuning, as they were pretrained on large-scale public datasets that partially overlap with ours.

the reward of P_i; and π_θ and $\pi_{\theta_{\text{old}}}$ represent the current and old policy distributions, respectively. The normalized score $\mathcal{A}_i = [\mathcal{S}_i - \text{mean}(\{\mathcal{S}_j\}_{j=1}^G)]/\text{std}(\{\mathcal{S}_j\}_{j=1}^G)$, reflects the relative quality of each reasoning path within the rollout group, enabling the model to distinguish between learnable and poor reasoning trajectories. Through RL training, the agent learns to adaptively reason over pixel features and plan click actions when necessary, achieving superior autonomous interactive segmentation.

4. Experiment

4.1. Experimental Setup

Evaluation Benchmarks. For segmentation performance evaluation, we conduct experiments on three datasets: (1) In-domain test set D_{test}. We randomly sample 9K image–mask pairs from the official BiomedParseData [46] test split. (2) Out-of-domain benchmark. We adopt the MeCOVQA-G+ [12, 40] test set, which includes 3K samples across 5 imaging modalities. This benchmark pairs biomedical images with text queries that explicitly request segmentation of specific anatomical structures or lesions. (3) Held-out in-house dataset. Since many foundation models and MLLMs are trained on large-scale public datasets, training–testing overlap may vary across prior works. To ensure fair evaluation and avoid unintentional data leakage, we additionally evaluate on a private held-out dataset collected from three medical centers, comprising 1K CT, MRI, and pathology images across seven cancer types. For VQA performance evaluation, we use four widely adopted medical VQA benchmarks: PathVQA [8], SLAKE [22], VQA-RAD [19], and OmniMedVQA [10]. More benchmark details are provided in Appendix A.3.

Metrics. For the segmentation task, we report mIOU and Dice score for mask segmentation, and F1 Score for mask-to-entity correspondence accuracy. For the VQA task, we evaluate the accuracy of the model's responses.

Implementation Details. We implement IBISAgent based on Qwen2.5-VL-7B [2]. MedSAM2 [27] is used as the interactive segmentation tool. The training is conducted on a cluster of 16 NVIDIA A100 GPUs. For the cold-start SFT stage, we optimize the model with a learning rate of 1×10^{-5} for 10 epochs. The total batch size is 256. The subsequent RL optimization is implemented using the VERL [34] framework, where we set the training batch size to 256 and generate 4 candidate reasoning paths per question, up to a maximum of 20 times of actions. We use a constant learning rate of 1×10^{-6} and set the maximum context length to 32K tokens. RL training runs for 12 epochs.

4.2. Comparison with Previous SOTA Methods

Segmentation Performance. We compare IBISAgent against two groups of baselines in Tab. 1: (1) General-purpose MLLMs with segmentation capability, including LISA-7B [18], LISA++ 7B [45], SAM4MLLM-8B [5], and VisionReasoner-7B [25] (using SAM2 [30]). (2) Medical MLLMs that support segmentation, including MedPLIB-7B [13], Citrus-V 8B [40], UniBiomed-1B [41], and MMedAgent-7B [20] (using MedSAM [26]).

Compared with general-domain MLLMs, IBISAgent achieves substantially higher performance across all benchmarks. Since these models were trained solely on datasets containing natural images, we further fine-tune them using our cold-start SFT and RL datasets—which include the same images, masks, and QA pairs used for training IBISAgent—to ensure fairness. Even after this additional fine-tuning, IBISAgent still significantly outperforms these models, demonstrating that its superiority does not simply



Figure 3. **Qualitative Analysis.** We present the responses and segmentation outputs on a reasoning–segmentation example. Existing MLLMs exhibit incorrect reasoning and low-quality segmentation, highlighting their misaligned fine-grained vision–language understanding. In contrast, **IBISAgent** delivers substantially improved reasoning quality and segmentation performance.

Efficiency \ Method	UniBiomed [41]	MedPLIB [13]	Citrus-V [40]	MMedAgent [20]	IBISAgent
Inference Time (s)	5.82	3.14	8.25	10.42	28.70

Table 2. Computational efficiency comparison. All experiments are conducted on the same A100 GPU.

arise from the use of a specialized dataset. Instead, the improvements primarily stem from our novel and effective design, which emulates the annotation process of human experts, enabling multi-round reasoning and refinement. This domain-specific modification, tailored for medical imaging, ultimately yields markedly better segmentation results.

Compared with existing medical-specific MLLMs, IBISAgent achieves substantially superior segmentation performance. On average, it surpasses these models by at least 35.13% in IoU, 37.58% in DSC, and 29.79% in F1 score. Notably, Citrus-V [40] and UniBiomed [41] were trained on datasets far larger than ours; nevertheless, our method consistently outperforms them, further confirming that the improvements arise from the effectiveness of our proposed technical components rather than from the use of a specialized dataset. In particular, MMedAgent [20] is also a tool-augmented MLLM that employs MedSAM [26] for segmentation. IBISAgent still markedly outperforms this model, demonstrating that its advantages do not merely arise from the integration of segmentation tools, but from more accurate grounding and iterative reasoning-based segmentation mask refinement.

Qualitative Analysis. As shown in Fig. 3, we present qualitative results that further demonstrate the merits of IBISAgent. For a case requiring reasoning to identify a liver tumor, the powerful closed-source GPT-4o [15] outputs seemingly correct textual reasoning but produces an incorrect localization bounding box. This indicates that the model fails to align textual reasoning with visual features—it may be hallucinating plausible answers rather than truly understanding the medical image, ultimately leading to incorrect localization. MedPLIB [13] provides coarse localization that roughly captures the lesion region but suffers from false-positive predictions. Moreover, its language output is highly constrained—when performing segmentation, the model rigidly outputs only the token "<SEG>", losing the rich language generation capability originally inherent to large language models. UniBiomed [41] produces incorrect textual reasoning and responses, describing objects that do not exist in the image and generating low-quality segmentation masks. This observation reinforces our motivation: current methods that rely on implicit "<SEG>" tokens for segmentation disrupt the MLLM's native text output space, compromising both its language capability and semantic generalization. Moreover, these approaches exhibit fundamentally limited pixel-level visual reasoning, preventing the model from truly understanding fine-grained visual features. In contrast, IBISAgent not only generates correct and coherent textual reasoning but also adaptively produces high-quality masks through step-by-step, precise pixel grounding. This capability arises from our design that decouples reasoning from segmentation, thereby preserving the MLLM's inherent language reasoning ability while simultaneously enabling adaptive, multi-round refinement.

Methods	MeCOVQA-G+			In-House Testset		
	IoU ↑	DSC ↑	F1 ↑	IoU ↑	DSC ↑	F1 ↑
GPT-4o [15] + MedSAM2 [27]	11.75	17.39	22.42	7.23	10.40	15.16
LLaVA-Med [21] + MedSAM2 [27]	24.54	31.38	35.70	20.03	26.94	37.75
HuatuoGPT-Vision [4] + MedSAM2 [27]	35.86	43.41	54.79	30.25	36.72	52.28
Lingshu [43] + MedSAM2 [27]	39.63	47.18	60.31	31.19	37.55	54.02
Chiron [37] + MedSAM2 [27]	39.58	47.16	60.25	30.75	36.90	53.37
IBISAgent	**80.63**	**89.27**	**95.24**	**72.09**	**83.78**	**91.76**

Table 3. Comparison of IBISAgent with tool-augmented MLLM agents. Except for the closed-source models, all competing methods in this table are implemented using their official open-source code and model weights, followed by further fine-tuning on our $\mathcal{D}_{\text{cold}}$ and $\mathcal{D}_{\text{rl}}$ datasets to ensure a fair comparison.

Efficiency Comparison. We randomly sampled $1,000$ cases (30%) from the out-of-domain testset MeCOVQA-G+ [12, 40] to measure the average time required by different models to process each sample. The results are reported in Tab. 2. Because IBISAgent performs multi-round mask refinement, its per-case inference time is longer than that of existing MLLMs. Fundamentally, such multi-step interactions between the agent and the environment are intended to trade additional computation time for improved performance—an inherent limitation of multi-round agent systems. Nevertheless, the inference overhead of IBISAgent remains within an acceptable range, especially in light of the substantial performance gains it delivers.

Method	Training Strategies			In-House Testset		
	SFT	Self-Reflection	RL	IoU ↑	DSC ↑	F1 ↑
prompt-driven $\mathcal{M}_{\text{base}}$				11.77	16.83	23.47
$\mathcal{M}_{\text{cold}}$	✓			53.42	62.01	68.61
$\mathcal{M}_{\text{cold+reflect}}$	✓	✓		57.16	67.73	74.52
$\mathcal{M}_{\text{rl}}$			✓	62.77	71.29	77.50
$\mathcal{M}_{\text{cold+rl}}$	✓		✓	68.92	78.08	85.44
IBISAgent	✓	✓	✓	**72.09**	**83.78**	**91.76**

Table 4. Ablation study on our training strategies. Checkmarks (✓) indicate which strategies are applied.

4.3. Ablation Studies

IBISAgent vs. Prompting-based MLLM Agents. To validate the effectiveness of formulating segmentation as an interleaved fine-grained visual reasoning and action planning Markov decision process, we design several baseline agent workflows that directly predict bounding boxes or points using MLLMs, followed by segmentation with MedSAM2 [27]. We then compare IBISAgent with these agent systems to evaluate the advantages of our iterative, multi-round, reasoning-driven formulation. Specifically, we construct different agent systems using GPT-4o [15], LLaVA-Med [21], HuatuoGPT-Vision [4], Lingshu [43], and Chiron [37]. The results are shown in Tab. 3. IBISAgent consistently outperforms all competing agent systems, demonstrating that our method effectively activates the model's intrinsic pixel-level reasoning ability, enabling advanced segmentation performance that cannot be achieved through simple workflow-style tool calling.

Effectiveness of Training Strategies. To evaluate the effectiveness of our training framework, we compare IBISAgent with several Qwen2.5-VL-7B–based variants: (1) $\mathcal{M}_{base}$, a prompt-driven baseline equipped with MedSAM2 [27]; (2) $\mathcal{M}_{cold}$, trained solely with cold-start SFT; (3) $\mathcal{M}_{cold+reflect}$, trained with cold-start SFT augmented by synthesized self-reflective trajectories; (4) $\mathcal{M}_{rl}$, trained exclusively with RL; and (5) $\mathcal{M}_{cold+rl}$. The results in Table 4 demonstrate the effectiveness of our training strategies. The prompt-only approach is neither adaptable nor robust, whereas progressively incorporating our proposed training stages yields substantial improvements over the base model, highlighting the necessity of each component. The RL stage provides the largest performance gain, indicating that RL's exploration–exploitation dynamics and reward feedback are crucial for acquiring context-aware decision-making policies and enabling genuine vision-centric multi-step reasoning. Furthermore, by integrating all training strategies, IBISAgent surpasses all baselines, further supporting our motivation for developing a versatile training framework that strengthens the model's pixel-level visual reasoning capabilities.

Significance of Different Reward Signals. We demonstrate the effectiveness of the proposed Region-based Click Placement Reward $\mathcal{S}_{\text{click}}$, Progressive Segmentation Improvement Reward $\mathcal{S}_{\text{pseg}}$, and Trajectory Length Reward $\mathcal{S}_{\text{len}}$ using the ablation results in Tab. 5. We begin by removing all three reward signals to establish a baseline that relies solely on the standard format reward and final answer reward (shown in the gray row). We then progressively incorporate different combinations of our proposed reward signals to examine their impact on segmentation performance. We also report the average predicted trajectory length (steps) to comprehensively evaluate how these rewards influence both segmentation quality and interaction efficiency. Each reward contributes substantially to the overall performance. The rewards $\mathcal{S}_{\text{click}}$ and $\mathcal{S}_{\text{pseg}}$ provide the greatest improvements in mask quality, as they ensure accurate click localization and encourage each interaction step to make a positive contribution to the segmentation outcome. Meanwhile, $\mathcal{S}_{\text{pseg}}$ and $\mathcal{S}_{\text{len}}$ improve interaction efficiency by discouraging redundant or uninformative clicks and preventing unnecessarily long trajectories, enabling the model to learn when to stop. By integrating all reward components, IBISAgent achieves an optimal balance between segmentation performance and interaction efficiency.

Reward Signals			MeCOVQA-G+				In-House Testset			
$\mathcal{S}_{\text{click}}$	$\mathcal{S}_{\text{pseg}}$	$\mathcal{S}_{\text{len}}$	IoU ↑	DSC ↑	F1 ↑	T_{avg} ↓	IoU ↑	DSC ↑	F1 ↑	T_{avg} ↓
			73.77	82.62	88.53	11.29	68.96	79.17	87.06	13.44
✓			76.60	85.77	91.23	10.61	70.45	81.23	89.30	12.74
	✓		77.64	86.85	92.31	8.59	70.62	81.40	89.56	10.07
		✓	74.19	82.88	89.05	5.94	69.03	79.65	87.48	7.22
✓	✓		80.61	89.19	**95.24**	8.12	72.05	83.74	91.71	9.68
✓		✓	77.73	86.88	92.34	5.03	70.75	81.63	89.78	6.49
	✓	✓	79.37	88.25	93.97	4.26	71.16	82.28	90.55	5.43
✓	✓	✓	**80.63**	**89.27**	**95.24**	**3.67**	**72.09**	**83.78**	**91.76**	**4.89**

Table 5. Ablation study on reward design. The gray row indicates the baseline that excludes all segmentation-tailored rewards and uses only the standard format and answer rewards.

5. Conclusion

In this paper, we present IBISAgent, a novel MLLM-based agent capable of pixel-level visual reasoning for unified biomedical object referring and segmentation. IBISAgent reformulates segmentation as a multi-step Markov Decision Process and enhances MLLMs' pixel-level understanding without introducing additional model components. We develop IBISAgent through a two-stage training protocol and design fine-grained rewards that incentivize continuous self-improvement reasoning. Additionally, we introduce a large-scale high-quality dataset with thinking and action trajectories. Through comprehensive empirical evaluation, we demonstrate the competitive performance of IBISAgent across diverse segmentation tasks. Our study paves the way for future exploration of vision-centric, multi-step decision-making agents for holistic medical image analysis.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant No. 62502429, and the Zhejiang Key Laboratory Project (2024E10001).

References

[1] Fan Bai, Yuxin Du, Tiejun Huang, Max Q-H Meng, and Bo Zhao. M3d: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*, 2024. 2

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 6

[3] Zechen Bai, Tong He, Haiyang Mei, Pichao Wang, Ziteng Gao, Joya Chen, Zheng Zhang, and Mike Zheng Shou. One token to seg them all: Language instructed reasoning segmentation in videos. *Advances in Neural Information Processing Systems*, 37:6833–6859, 2024. 2

[4] Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, et al. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*, 2024. 1, 7, 8

[5] Yi-Chia Chen, Wei-Hua Li, Cheng Sun, Yu-Chiang Frank Wang, and Chu-Song Chen. Sam4mllm: Enhance multi-modal large language model for referring expression segmentation. In *European Conference on Computer Vision*, pages 323–340. Springer, 2024. 6

[6] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 4

[7] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 5

[8] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020. 1, 6

[9] Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. *arXiv preprint arXiv:2503.24290*, 2025. 5

[10] Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22170–22183, 2024. 6

[11] Jiaqi Huang, Zunnan Xu, Jun Zhou, Ting Liu, Yicheng Xiao, Mingwen Ou, Bowen Ji, Xiu Li, and Kehong Yuan. Sam-r1: Leveraging sam for reward feedback in multimodal segmentation via reinforcement learning. *arXiv preprint arXiv:2505.22596*, 2025. 3, 5

[12] Xiaoshuang Huang, Lingdong Shen, Jia Liu, Fangxin Shang, Hongxiang Li, Haifeng Huang, and Yehui Yang. Towards a multimodal large language model with pixel-level insight for biomedicine, 2025. 2, 6, 7

[13] Xiaoshuang Huang, Lingdong Shen, Jia Liu, Fangxin Shang, Hongxiang Li, Haifeng Huang, and Yehui Yang. Towards a multimodal large language model with pixel-level insight for biomedicine. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3779–3787, 2025. 2, 6, 7

[14] Yu Huang, Zelin Peng, Yichen Zhao, Piao Yang, Xiaokang Yang, and Wei Shen. Medseg-r: Reasoning segmentation in medical images with multimodal large language models, 2025. 2

[15] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 7, 8

[16] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*, 2019. 1

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. 2, 8

[18] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024. 2, 6

[19] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018. 1, 6

[20] Binxu Li, Tiankai Yan, Yuanting Pan, Jie Luo, Ruiyang Ji, Jiayuan Ding, Zhe Xu, Shilong Liu, Haoyu Dong, Zihao Lin, et al. Mmedagent: Learning to use medical tools with multimodal agent. *arXiv preprint arXiv:2407.02483*, 2024. 3, 6, 7

[21] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023. 1, 7, 8

[22] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1650–1654. IEEE, 2021. 6

[23] Ye Liu, Zongyang Ma, Junfu Pu, Zhongang Qi, Yang Wu, Ying Shan, and Chang Wen Chen. Unipixel: Unified object

referring and segmentation for pixel-level visual reasoning. *arXiv preprint arXiv:2509.18094*, 2025. 2

[24] Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*, 2025. 3, 5

[25] Yuqi Liu, Tianyuan Qu, Zhisheng Zhong, Bohao Peng, Shu Liu, Bei Yu, and Jiaya Jia. Visionreasoner: Unified visual perception and reasoning via reinforcement learning. *arXiv preprint arXiv:2505.12081*, 2025. 3, 5, 6

[26] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1), 2024. 2, 6, 7, 8, 9

[27] Jun Ma, Zongxin Yang, Sumin Kim, Bihui Chen, Mohammed Baharoon, Adibvafa Fallahpour, Reza Asakereh, Hongwei Lyu, and Bo Wang. Medsam2: Segment anything in 3d medical images and videos. *arXiv preprint arXiv:2504.03600*, 2025. 3, 6, 7, 8, 9

[28] OpenAI. Gpt-5 chat. https://chat.openai.com, 2025. Accessed: 2025-09-17. 4

[29] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR, 2022. 1

[30] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2, 6

[31] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024. 8

[32] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26374–26383, 2024. 2

[33] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025. 1

[34] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297, 2025. 6

[35] Mennatullah Siam. Pixfoundation: Are we heading in the right direction with pixel-level vision foundation models? *arXiv preprint arXiv:2502.04192*, 2025. 2

[36] Amber L. Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram van Ginneken, Annette Kopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M. Summers, Patrick Bilic, Patrick F. Christ, Richard K. G. Do, Marc Gollub, Jennifer Golia-Pernicka, Stephan H. Heckers, William R. Jarnagin, Maureen K. McHugo, Sandy Napel, Eugene Vorontsov, Lena Maier-Hein, and M. Jorge Cardoso. A large annotated medical image dataset for the development and evaluation of segmentation algorithms, 2019. 1

[37] Haoran Sun, Yankai Jiang, Wenjie Lou, Yujie Zhang, Wenjie Li, Lilong Wang, Mianxin Liu, Lei Liu, and Xiaosong Wang. Enhancing step-by-step and verifiable medical reasoning in mllms. *arXiv preprint arXiv:2506.16962*, 2025. 7, 8

[38] Qinyue Tong, Ziqian Lu, Jun Liu, Yangming Zheng, and Zheming Lu. Medisee: Reasoning-based pixel-level perception in medical images. *arXiv preprint arXiv:2504.11008*, 2025. 2

[39] Quoc-Huy Trinh, Minh-Van Nguyen, Jung Zeng, Ulas Bagci, and Debesh Jha. Prs-med: Position reasoning segmentation with vision-language model in medical imaging, 2025. 2

[40] Guoxin Wang, Jun Zhao, Xinyi Liu, Yanbo Liu, Xuyang Cao, Chao Li, Zhuoyun Liu, Qintian Sun, Fangru Zhou, Haoqiang Xing, and Zhenhong Yang. Citrus-v: Advancing medical foundation models with unified medical image grounding for clinical reasoning, 2025. 2, 6, 7

[41] Linshan Wu, Yuxiang Nie, Sunan He, Jiaxin Zhuang, Luyang Luo, Neeraj Mahboobani, Varut Vardhanabhuti, Ronald Cheong Kin Chan, Yifan Peng, Pranav Rajpurkar, et al. Unibiomed: A universal foundation model for grounded biomedical image interpretation. *arXiv preprint arXiv:2504.21336*, 2025. 2, 6, 7

[42] Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas S Huang. Deep interactive object selection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 373–381, 2016. 4, 3

[43] Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*, 2025. 1, 7, 8

[44] Zhonghao Yan, Muxi Diao, Yuxuan Yang, Jiayuan Xu, Kaizhou Zhang, Ruoyan Jing, Lele Yang, Yanxi Liu, Kongming Liang, and Zhanyu Ma. Medreasoner: Reinforcement learning drives reasoning grounding from clinical thought to pixel-level precision, 2025. 5

[45] Senqiao Yang, Tianyuan Qu, Xin Lai, Zhuotao Tian, Bohao Peng, Shu Liu, and Jiaya Jia. Lisa++: An improved baseline for reasoning segmentation with large language model. *arXiv preprint arXiv:2312.17240*, 2023. 2, 6

[46] Theodore Zhao, Yu Gu, Jianwei Yang, Naoto Usuyama, Ho Hin Lee, Tristan Naumann, Jianfeng Gao, Angela Crabtree, Jacob Abel, Christine Moung-Wen, et al. Biomedparse: a biomedical foundation model for image parsing of everything everywhere all at once. *arXiv preprint arXiv:2405.12971*, 2024. 1, 4, 5, 6, 9