

INCENTIVIZING TOOL-AUGMENTED THINKING WITH IMAGES FOR MEDICAL IMAGE ANALYSIS

Anonymous authors

Paper under double-blind review

ABSTRACT

Recent reasoning based medical MLLMs have made progress in generating step by step textual reasoning chains. However, they still struggle with complex tasks that necessitate dynamic and iterative focusing on fine-grained visual regions to achieve precise grounding and diagnosis. We introduce **Ophiuchus**, a versatile, tool-augmented framework that equips an MLLM to (i) decide when additional visual evidence is needed, (ii) determine where to probe and ground within the medical image, and (iii) seamlessly weave the relevant sub-image content back into an interleaved, multimodal chain of thought. In contrast to prior approaches limited by the performance ceiling of specialized tools, **Ophiuchus** integrates the model’s inherent grounding and perception capabilities with external tools, thereby fostering higher-level reasoning. The core of our method is a three-stage training strategy: cold-start training with tool-integrated reasoning data to achieve basic tool selection and adaptation for inspecting key regions; self-reflection fine-tuning to strengthen reflective reasoning and encourage revisiting tool outputs; and Agentic Tool Reinforcement Learning to directly optimize task-specific rewards and emulate expert-like diagnostic behavior. Extensive experiments show that **Ophiuchus** consistently outperforms both closed-source and open-source SOTA methods across diverse medical benchmarks, including VQA, detection, and reasoning-based segmentation. Our approach illuminates a path toward medical AI agents that can genuinely “think with images” through tool-integrated reasoning. Datasets, codes, and trained models will be released publicly.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) have exhibited remarkable performance across a wide range of medical image analysis tasks such as visual question answering (VQA) (Li et al., 2023; Chen et al., 2024a), disease diagnosis (Sun et al., 2025b; Liu & Song, 2025), segmentation (Bai et al., 2024; Wang et al., 2025a), and report generation (Bassi et al., 2025; Wang et al., 2025b). Recent breakthroughs in chain-of-thought (CoT) techniques (Wei et al., 2022) have further advanced the reasoning capabilities of medical MLLMs. Numerous studies employ supervised fine-tuning (SFT) (Sun et al., 2025a) or reinforcement learning (RL) (Lai et al., 2025b; Pan et al., 2025; Su et al., 2025b) to push these models beyond the direct-prediction paradigm, enabling step-by-step reasoning to address complex clinical challenges and improve diagnostic decision support.

Despite these advancements, current medical MLLMs still show critical limitations in how they interact with visual information during reasoning processes. First, they often attend to irrelevant regions while missing fine-grained evidence, such as tiny structures, intricate abnormalities, subtle lesion cues, and nuanced pathological semantics (Wang et al., 2025a). Such omissions severely degrade diagnostic performance. In essence, these weaknesses stem from a static, global-perception paradigm: the models primarily rely on image-level representations and lack the key ability to actively and adaptively probe and explore localized, fine-grained visual details. Moreover, prevailing medical MLLMs express intermediate reasoning steps exclusively in text and lack a “look-again” mechanism during thinking, leading to the loss of critical visual information. Ideally, MLLMs should autonomously perform dynamic, iterative, fine-grained interactions with task-relevant image regions throughout the reasoning process, revisiting and revising earlier answers in light of emerging visual cues to support more accurate perceptual decision-making, as exemplified in Figure 1.

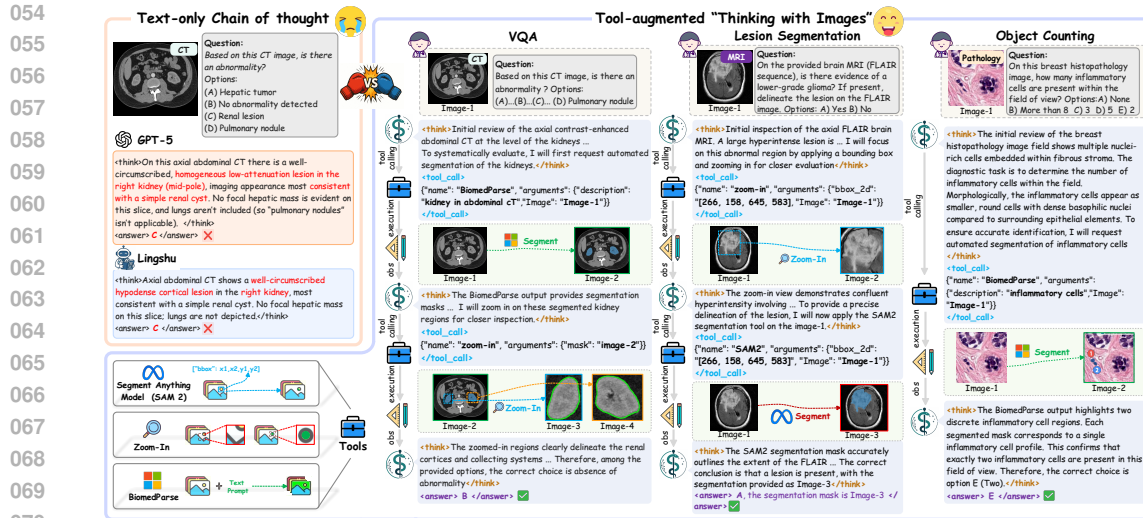


Figure 1: Overview of the tool-augmented “thinking with images” paradigm. Compared with (a) text-only CoT, which fails to analyze fine-grained, task-critical image regions and thus limits understanding, increases hallucinations, and produces false positives, (b) our tool-augmented, interleaved vision-language reasoning adaptively generates effective tool-invocation strategies, inspects fine-grained regions, and integrates the resulting evidence into subsequent reasoning, yielding more accurate diagnostic cognition across diverse medical tasks.

The aforementioned challenges motivate a fundamental rethinking of how medical MLLMs engage more seamlessly with fine-grained visual information during reasoning. This leads us to propose **Ophiuchus**, a novel and versatile framework capable of interleaved vision-language reasoning through thinking with tools across diverse medical tasks, including VQA, detection, and segmentation. **Ophiuchus** can decide whether to actively integrate external visual tools (e.g., SAM2 (Ravi et al., 2024), BiomedParse (Zhao et al., 2024), and zoom-in) directly into the reasoning loop to iteratively manipulate and interpret key visual content. This paradigm shift from pure text-based reasoning to a more grounded and interpretable tool-augmented visual cognition—a new frontier where reasoning is continuously intertwined with ongoing visual perception. To develop **Ophiuchus**, we construct a high-quality dataset of 64k samples with annotations for grounding key visual regions and explicit tool-invocation trajectories, and we propose a three-stage training protocol (Figure 2). First, we introduce cold-start SFT with tool-integrated reasoning data, enabling the model to learn how to invoke appropriate external vision tools and extract useful information from their outputs to solve questions beyond its internal capabilities. Second, we conduct fine-tuning with a new self-reflection sampling (SRS) strategy that further strengthens self-correction, enabling the model to assess whether its tool choices meet expectations based on intermediate results and to adjust accordingly. Finally, we introduce agentic tool reinforcement learning (ATRL) with fine-grained rewards that further strengthens reasoning, enabling the model to autonomously discover effective tool-use policies rather than merely imitating the static tool-use trajectories from the SFT stage.

We conduct a comprehensive evaluation on 8 benchmarks, spanning both in-domain and zero-shot scenarios, to rigorously assess the performance of **Ophiuchus**. Results show that our model exhibits strong interleaved vision-language reasoning with tools and significantly exceeds the SOTA MLLMs across all tasks. Our contributions can be summarized as follows:

- We propose tool-augmented “thinking with images”, an interleaved vision-language reasoning paradigm that enables medical MLLMs to adaptively process localized visual evidence and integrate it into the evolving reasoning chain.
- We develop a principled, multi-stage training framework that cultivates tool-augmented reasoning: the model progresses from early tool exploration to advanced orchestration, enabling it to solve complex diagnostic problems that require detailed visual inspection.
- We achieve SOTA performance across multiple medical benchmarks. **Ophiuchus** exhibits emergent reasoning patterns, demonstrating robust tool-assisted problem solving.

2 RELATED WORKS

Medical MLLMs for Fine-Grained Vision Understanding. Multimodal large language models have demonstrated promising performance in handling a wide range of medical imaging modalities and tasks (Li et al., 2023; Chen et al., 2024a; Yang et al., 2024; Lin et al., 2025). To enable region-level perception and assessment, numerous approaches (Chen et al., 2024b; Xie et al., 2024; Shui et al., 2025; Deng et al., 2025) introduce region-of-interest (ROI) supervision, teaching medical MLLMs to ground critical regions. Moreover, some studies (Bai et al., 2024; Wang et al., 2025a; Huang et al., 2025) have also explored pixel understanding in MLLMs by combining these models with segmentation tasks. These methods require MLLMs to learn an implicit $[seg]$ token and involve additional fine-tuning with a separate pixel decoder. Despite direct fine-grained supervision, these models still exhibit limited grounding accuracy and mask precision (Zhu et al., 2025b; Nath et al., 2025), and they often underperform task-specific models. In contrast, we train the MLLM to act as a competent agent that effectively invokes external tools across diverse queries, enabling it to surpass inherent limitations in pixel-level grounding and to solve tasks beyond its native capacity.

Tool-Augmented Medical Agents. Augmenting medical MLLMs with external tools is an active direction that extends capabilities beyond standalone models by leveraging dedicated functions or expert models for quantitative assessment. MMedAgent (Li et al., 2024) curates an instruction-tuning corpus covering six medical tools, enabling task-specific tool invocation and result aggregation. Similarly, VILA-M3 (Nath et al., 2025) trains models to trigger medical expert models for tasks such as segmentation and classification. AURA (Fathi et al., 2025) integrates different medical models as tools for different tasks, including MedSAM (Ma et al., 2024) for segmentation, and CheXAgent (Chen et al., 2024c) for VQA. However, these methods remain restricted to fixed perception behaviors due to fragmented reasoning composed of disconnected tool invocations that undermine coherence and holistic planning. For multi-turn, multi-step reasoning, MedAgent-Pro (Wang et al., 2025c) builds a hierarchical multi-agent framework integrating GPT-4o (Achiam et al., 2023), a retrieval-augmented generation agent, and specialized tools to support disease diagnosis under disease-specific workflows. SMR-Agents (Wang et al., 2026) adopts a large language model as a general practitioner, which automatically selects and coordinates a team of specialized medical expert agents to refine the final answer. Although these approaches support sequential composition of multiple tools, their reliance on predefined workflows limits generalization and blocks recovery from tool failures. In contrast, **Ophiuchus** attains tool-augmented “thinking with images” through a three-stage training framework, enabling the model to learn diverse, context-aware tool-use strategies and to reflect on tool outputs via an iterative reasoning process.

3 METHODS

We first present an overview of **Ophiuchus**. Then we introduce our three-stage training framework including cold-start SFT, self-reflection fine-tuning, and Agentic Tool Reinforcement Learning.

3.1 OPHIUCHUS

Ophiuchus is a unified multimodal agent that can “think with images” by adaptively invoking and sequencing tools throughout CoT reasoning. As illustrated in Figure 1, given a user question Q and an input image I , **Ophiuchus** generates a multi-step reasoning path R to derive the final answer. This reasoning path R can be represented as an N -step chain: $R = \{(r_n, t_n, o_n)\}_{n=1}^N$, where each step comprises natural-language thoughts r_n , a tool invocation t_n for further image inspection, and the resulting observation o_n . The iterative thought-tool-observation loop continues until the model reaches a conclusive answer in the final reasoning thoughts r_N or when predefined limits on context length or interaction turns are reached. The core components are detailed below.

Available Tools for Image Analysis. We equip **Ophiuchus** with three useful tools, including:

- SAM2 (Ravi et al., 2024). We employ SAM2 as one of the segmentation tools due to its superior performance and efficient inference speed. Leveraging the bounding boxes provided by **Ophiuchus**, SAM2 can generate precise, fine-grained mask for the target object.
- BiomedParse (Zhao et al., 2024), which takes an image and a text prompt as inputs. The text prompt specifies the object type for segmentation, then BiomedParse outputs a segmentation mask.

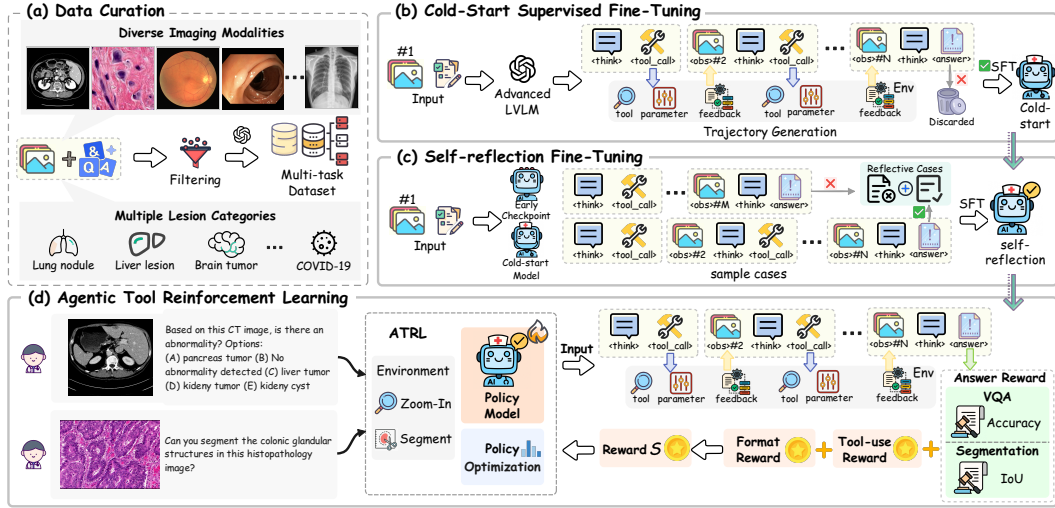


Figure 2: **Overview of the Ophiuchus.** A versatile “think with images” framework that integrates zoom and segmentation tools, trained through cold-start SFT, self-reflection fine-tuning and ATRL.

Unlike SAM2, the text-driven BiomedParse enables complete delegation of region localization to the tool, whereas SAM2 still relies on the MLLM to produce bounding boxes (bboxes) as prompts. Although both support segmentation, BiomedParse complements rather than duplicates SAM2; providing an effective alternative when the MLLM cannot reliably generate bboxes (Appendix C.2).

- An image zoom-in function, which accepts an image together with either bboxes or masks as input. With bboxes, the function returns zoom-in crops of the specified regions; with masks, it returns the corresponding zoom-in crops and additionally delineates the target object contours (making boundary morphology explicit in this way facilitates recognition of lesion-specific shape cues).

Tool-Integrated Reasoning. At the n -th reasoning step, **Ophiuchus** \mathcal{M} can autonomously decide, after textual CoT thoughts r_n , whether to directly produce an answer, or invoke a tool t_n for further image inspection. The tool’s outputs, such as masks or cropped images, are sequentially indexed to maintain order and appended to the ongoing trajectory, allowing the model to reason over all previous context. This reasoning process can be formulated as: $(r_n, t_n) = \mathcal{M}(Q, I, R_{<n})$, where $R_{<n} = \{(r_i, t_i, o_i)\}_{i<n}$ denotes the reasoning history before step n . Each tool invocation t_i can target either the original input image or any previous observation along the trajectory via explicit indices. This design enables the model to learn dynamic tool composition across diverse tasks, coupling visual cues with textual reasoning to achieve more accurate perceptual decision-making.

3.2 TRAINING FRAMEWORK

Our training procedure consists of three stages (Figure 2): (1) cold-start SFT to establish basic tool-invocation and visual-interaction capabilities. (2) self-reflection fine-tuning that fosters self-correction behaviors, enabling the model to reassess tool outputs and, when necessary, to revise its tool-invocation decisions. (3) ATRL. We train with RL and carefully designed fine-grained rewards to further incentivize reasoning, driving tool use from early exploration to efficient, accurate exploitation and orchestration via trial-and-error interactions across diverse tasks.

Cold-Start Supervised Fine-Tuning. Prompting alone lacks the adaptability and robustness required for multimodal agents to reliably select and invoke appropriate tools for real-world medical tasks. Therefore, we initialize **Ophiuchus**’s ability to reason with tools via SFT on the cold-start dataset $\mathcal{D}_{\text{cold}}$, which contains ground-truth reasoning and tool-call trajectories. The training objective is to minimize the average negative log-likelihood over all reasoning and tool-call tokens:

$$\mathcal{L}_{\text{cold}} = \mathbb{E}_{(I, Q, A, R = \{(r_n, t_n, o_n)\}_{n=1}^N) \sim \mathcal{D}_{\text{cold}}} \left(-\frac{1}{T} \sum_{n=1}^N \log p(r_n, t_n \mid I, Q, R_{<n}) \right). \quad (1)$$

Here, $T = \sum_{n=1}^N (|r_n| + |t_n|)$ denotes the total number of tokens in reasoning steps and tool calls. A is the GT answer. The objective is to teach the model to generate valid tool-use trajectories with diverse, robust step-by-step reasoning patterns.

Self-Reflection Fine-Tuning. Moreover, considering the complexity of real-world medical tasks, errors can occur even when tools are invoked. It is therefore crucial for the model to exhibit reflective behavior over tool outputs. To this end, we introduce a novel **SRS strategy** that further strengthens the model’s self-correction capability. Specifically, we sample self-reflective cases from the model’s training dynamics. For challenging questions, when the agent moves from an incorrect answer at an early checkpoint to a correct answer at a later checkpoint, and its tool sequence differs between the two, we regard this as evidence of self-reflective strategy revision. Formally, we define a self-reflective reasoning process as a pair of trajectories $(R_{\text{early}}, R_{\text{late}})$ for the same input (I, Q) observed at different checkpoints, where the tool-invocation sequences differ and the answer correctness flips:

$$\exists (u, v) : (t_{\text{early}}^u \neq t_{\text{late}}^v) \wedge \text{Correct}(R_{\text{early}}) = 0 \wedge \text{Correct}(R_{\text{late}}) = 1, \quad (2)$$

where u and v denote arbitrary steps in two different reasoning trajectories. After each SFT epoch, we record the model’s output trajectories and answers, then examine them to identify such cases. We then curate a special training subset $\mathcal{D}_{\text{reflect}}$ of all identified cases. For each case in $\mathcal{D}_{\text{reflect}}$, we pair $(R_{\text{early}}, R_{\text{late}})$, reconcile and consolidate the trajectories, yielding a ground-truth reasoning trajectory. We further fine-tune **Ophiuchus** on $\mathcal{D}_{\text{reflect}}$ using a training objective analogous to Eq. 1. This selective training strategy encourages the model to learn from high-quality reasoning paths that exhibit both reflective thinking and correctness which is crucial for subsequent RL optimization.

Agentic Tool Reinforcement Learning. We further optimize **Ophiuchus** through RL with carefully designed rewards, enabling the agent to adaptively discover effective tool-use strategies and achieve compositional tool use, thereby moving beyond the constraints of mimicking the static trajectories observed during SFT. ATRL process comprises three key components:

(1) Rollout Formulation. ATRL extends the formulation of traditional RL with text-only CoT by introducing tool-call tokens and observation tokens. The observation tokens come from external function calls rather than the model itself. According to the formulation in Section 3.1, given a user query Q and an image I , the reasoning trajectory up to step k is denoted as: $R_k = (r_1, t_1, o_1), (r_2, t_2, o_2), \dots, (r_k, t_k, o_k)$. At each step $k + 1$, the model must generate the next reasoning thoughts r_{k+1} and select a tool t_{k+1} to formulate a parameterized invocation of the tool to make progress toward solving Q . The model’s policy is defined as: $(r_{k+1}, t_{k+1}) \sim \pi_{\theta}(\cdot | I, Q, R_k)$. This rollout process continues to interleave until either an answer is generated or the maximum number of tool calls is reached. To prevent inefficient or circular behavior, we enable early termination: if a tool invocation duplicates a previously executed one, the rollout stops immediately. We instruct the model to mark its thoughts, tool calls, and final answers in the output using the special tokens `<think>`, `<tool_call>`, and `<answer>`. When the model output includes `<tool_call>`, we automatically parse the tool calls into individual invocations using the model-predicted parameters. The outputs from executions are then inserted into the `<obs>` field and appended to the ongoing trajectory. All observation tokens are considered as a whole, which does not contribute to the loss computation. The system and user prompt is provided in Appendix E.3.

(2) Reward Design. Unlike prior works (Guo et al., 2025; Lai et al., 2025b; Su et al., 2025b) that rely on overly simplified, answer-based rewards, we introduce a novel rule-based reward, decomposed into fine-grained signals that deliver dense feedback across various tool-use tasks, steering the model toward effective reasoning. Formally, the final reward comprises three components: reasoning format adherence, final answer quality, and strategic tool usage.

- The format reward $\mathcal{S}_{\text{format}}$ evaluates the structural validity of R by verifying that the model’s output includes all required special tokens in the prescribed order.
- The final-answer reward \mathcal{S}_{ans} encompasses multiple task types, thereby providing the agent with diverse feedback. For close-ended 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.
- The strategic tool-use reward $\mathcal{S}_{\text{tool}}$ is defined as a conditional bonus, granted only when the model both produces a correct answer and invokes at least one external perception tool during its trajectory. This design encourages the model to employ tools meaningfully-when they directly contribute to successful task completion-rather than using them arbitrarily or redundantly.

The final reward \mathcal{S} is derived as: $\mathcal{S} = \mathcal{S}_{ans} + \mathcal{S}_{format} + \mathcal{S}_{tool}$. Formal equations for each reward signal are provided in Appendix B. This diverse and fine-grained reward design better reflects the complexity of real-world tool use, guiding the model to generate outputs that are both syntactically valid and semantically faithful, which is crucial for final task success.

(3) Optimization. Based on the rollout formulation and rewards defined above, we optimize the policy using GRPO (Guo et al., 2025) without the KL penalty term (Hu et al., 2025) on dataset \mathcal{D}_{rl} :

$$\mathcal{L}_{RL} = \mathbb{E}_{\substack{(I,Q,A) \sim \mathcal{D}_{rl} \\ \{R_i\}_{i=1}^G \sim \pi_{\theta_{old}}(\cdot|I,Q)}} \left(-\frac{1}{G} \sum_{i=1}^G \frac{1}{T_i} \sum_{n=1}^{N_i} \min(\pi_{\theta_{i,n}} \mathcal{A}_i, \text{clip}(\pi_{\theta_{i,n}}, 1 - \epsilon, 1 + \epsilon) \mathcal{A}_i) \right), \quad (3)$$

$$\pi_{\theta_{i,n}} = \frac{\pi_{\theta}(r_{i,n}, t_{i,n} | I, Q, R_{i,<n})}{\pi_{\theta_{old}}(r_{i,n}, t_{i,n} | I, Q, R_{i,<n})}, \quad \mathcal{A}_i = \frac{\mathcal{S}_i - \text{mean}(\{\mathcal{S}_j\}_{j=1}^G)}{\text{std}(\{\mathcal{S}_j\}_{j=1}^G)}, \quad (4)$$

Here, G is the number of rollout reasoning paths; $R_i = \{(r_{i,n}, t_{i,n}, o_{i,n})\}_{n=1}^{N_i}$ denotes the i -th reasoning path; T_i is the total length of R_i excluding tool outputs; \mathcal{S}_i is the reward of R_i ; and π_{θ} and $\pi_{\theta_{old}}$ represent the current and old policy distributions, respectively. The normalized score \mathcal{A}_i reflects the relative quality of each reasoning path within the rollout group, enabling the model to distinguish between learnable and poor reasoning trajectories. Through ATRL training, the agent learns to utilize tools for key visual cues inspection, mimic expert-like efficient reasoning behaviors.

3.3 DATASET CURATION

Data Collection and QA Generation. Our data collection follows three core principles: (1) diverse tasks and imaging distributions; (2) scenarios where tool use yields measurable accuracy gains; and (3) comprehensive fine-grained annotations. Consequently, we leverage datasets from BiomedParseData (Zhao et al., 2024), which comprises 3.4 million triples of image, segmentation mask, and semantic label, encompass 82 major biomedical object types across 9 imaging modalities. We also include the Malenia dataset (Jiang et al., 2024), which provides 1,514 image-mask-report triplets with fine-grained descriptions of disease attributes across 12 lesion categories. To ensure that the tool invocations are genuinely necessary for resolving the vision-language queries, we prompt Gemini-2.5-pro (Comanici et al., 2025) to generate a QA pair conditioned on the image, mask, and the mask’s descriptions, yielding a fine-grained question that necessitates localizing the specified mask region. We further eliminate data that cannot be properly verified, such as questions with incorrect answers. The prompts for VQA generation and verification are provided in Appendix E.1.

Reasoning Trajectory Generation and Data Selection. We leverage GPT-5 (OpenAI, 2025) to generate reasoning paths following our reasoning paradigm described in Section 3.1. During generation, whenever GPT-5 issues a tool call, we provide the corresponding GT for that invocation. The generated paths are subsequently filtered based on format checking and answer correctness. Our data selection strategy encompasses two key steps: (1) We select instances where GPT-5 produces incorrect answers in single-turn interactions but achieves correct results when using tool invocation for fine-grained region inspection, highlighting scenarios where tool use is most beneficial. (2) We also include a small subset of questions that GPT-5 answers correctly without tools, aiming at teaching the model to adaptively employ tools only when necessary. Through systematic curation, we obtain 64k VQA samples spanning multiple task types and modalities. To ensure reliability, these samples are further cross-validated against human annotations. We split the data into 30k for \mathcal{D}_{cold} , 30k for \mathcal{D}_{rl} , and 4k for in-domain testing \mathcal{D}_{test} . Further details are provided in Appendix A.

4 EXPERIMENTS

4.1 EXPERIMENT SETTINGS

Evaluation Benchmark and Metrics. Beyond our in-domain-4k testing set, we also evaluate on several representative public benchmarks: (1) general medical VQA datasets—PathVQA (He et al., 2020), SLAKE (Liu et al., 2021), VQA-RAD (Lau et al., 2018), and OmniMedVQA (Hu et al., 2024); (2) more challenging medical-reasoning benchmarks—MMM(H&M) (Yue et al., 2024) and MedXpertQA Zuo et al. (2025); and (3) a held-out in-house dataset collected from three medical centers, comprising 1k CT, MRI, and pathology images across 7 cancer types for out-of-domain

Table 1: **Performance comparison on medical VQA benchmarks.** Gray-shaded rows denote large-sized models. **Bold** and underlined indicate the best and second-best results, respectively. **Improvement** in the last row denotes the absolute gain of **Ophiuchus** over the Qwen2.5-VL-7B (w/o tool use). Avg. is the arithmetic mean over the seven *Out-of-Domain* zero-shot benchmarks. To ensure fairness, since **Med-R1-2B** is trained on part of the OmniMedVQA test set, its Avg is computed only over the remaining six benchmarks (excluding OmniMed).

Methods	Tool	In-domain		Out-of-domain						Avg.
		\mathcal{D}_{test} -VQA	PathVQA	SLAKE	VQA-RAD	OmniMed	MMMU(H&M)	MedXpertQA	In-House-VQA	
<i>Close-Source SOTA</i>										
GPT-4.1	✗	36.4	58.3	71.6	65.2	75.6	73.6	40.8	26.1	58.7
GPT-5	✗	37.3	60.0	73.2	64.5	75.4	70.7	40.4	28.5	59.0
OpenAI-o3	✓	39.8	67.5	75.3	66.0	73.7	<u>74.5</u>	44.1	<u>30.7</u>	61.7
Gemini 2.5 Pro	✓	<u>40.2</u>	67.1	72.7	63.8	<u>76.9</u>	72.8	<u>46.6</u>	30.3	61.5
<i>Open-Source SOTA</i>										
InternVL3-8B	✗	34.9	53.2	70.4	65.6	72.2	62.3	23.8	20.4	52.6
LLaVA-Next-13B	✗	21.4	39.8	57.1	54.8	58.0	40.1	19.6	18.0	41.1
Qwen2.5-VL-32B	✗	36.5	47.7	70.1	<u>71.7</u>	69.5	60.1	26.8	25.5	53.1
<i>MLLMs can "Think with Images"</i>										
DeepEyes-7B	✓	37.1	52.9	68.2	65.9	64.8	57.8	23.6	20.9	50.6
Mini-o3-7B-v1	✓	37.7	53.4	67.8	65.7	65.1	57.4	24.3	21.5	50.7
PixelReasoner-RL-v1-7B	✓	37.4	52.6	67.3	66.0	64.9	58.0	23.5	21.2	50.5
<i>Medical MLLMs</i>										
LLaVA-Med-7B	✗	25.3	44.6	47.7	52.5	52.1	38.8	19.9	19.3	39.3
HuatuoGPT-Vision-34B	✗	34.8	50.7	68.3	61.7	70.6	60.1	23.6	20.8	50.8
<i>Medical MLLMs with CoT Reasoning</i>										
MedVLM-R1-2B	✗	25.8	38.3	54.3	45.2	68.3	45.9	21.7	19.8	41.9
Med-R1-2B*	✗	17.5	19.2	52.1	36.5	–	44.7	22.9	17.1	32.1
Lingshu-7B	✗	38.2	<u>68.4</u>	<u>77.8</u>	66.4	74.9	57.8	25.2	23.8	56.3
Chiron-o1-8B	✗	38.6	68.8	77.4	72.5	76.2	55.4	24.3	25.9	57.2
<i>Multimodal Medical Agents</i>										
MMedAgent-7B	✓	37.6	59.4	68.7	64.0	58.2	44.1	22.3	20.1	48.1
AURA	✓	37.9	59.8	68.4	64.5	59.7	49.3	23.5	21.8	49.6
SMR-Agents	✓	25.4	38.2	53.5	46.9	55.7	40.1	19.6	19.4	39.1
MedAgent-Pro	✓	36.7	58.5	69.4	63.3	59.5	52.9	27.8	23.7	50.7
MMedAgent-RL-7B	✗	36.0	58.5	67.9	66.1	60.3	58.9	22.6	21.4	50.8
VILA-M3-40B	✓	39.5	66.4	71.4	65.7	55.1	56.6	23.0	25.3	51.9
Qwen2.5-VL-7B	✗	33.6	50.4	66.5	63.4	64.7	56.7	23.5	20.2	49.3
Ophiuchus-7B (ours)	✓	59.4 ^(+25.8)	74.3 ^(+23.9)	83.9 ^(+17.4)	73.6 ^(+10.2)	78.6 ^(+13.9)	76.0 ^(+19.3)	39.3 ^(+15.8)	50.4 ^(+30.2)	68.0 ^(+18.7)

testing. Benchmark details are provided in Appendix A. For evaluation metrics, we use answer accuracy for multiple-choice VQA benchmarks. For subsets in \mathcal{D}_{test} -4k and in-house-1k that require segmentation results as answers, we use the Dice Similarity Coefficient (DSC).

Implementation details. We implement **Ophiuchus** based on Qwen2.5-VL-7B (Bai et al., 2025). 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. We use the same settings as in the cold-start SFT and conduct self-reflection fine-tuning for 2 epochs. The subsequent RL optimization is implemented using the VERL (Sheng et al., 2025) framework, where we set the training batch size to 256 and generate 8 candidate reasoning paths per question, up to a maximum of 6 times of tool calling. 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. During inference, we expose external model tools (e.g., SAM2 and BiomedParse) as APIs using FastAPI (Ramírez & contributors, 2018–2025) for tool invocation acceleration.

4.2 MAIN RESULTS

Medical VQA Performance. We conduct a comprehensive evaluation across eight benchmarks; results are summarized in Table 1. (1) Against general-purpose SOTA MLLMs—both closed-source (GPT-series (OpenAI, 2025), Gemini-2.5-Pro (Comanici et al., 2025)) and open-source (Qwen-series (Bai et al., 2025), LLaVA-Next-13B (Liu et al., 2024), InternVL3-8B (Zhu et al., 2025a))—**Ophiuchus** achieves substantially better performance. (2) Compared with SOTA MLLMs that also can “think with images” via zoom-in operations (DeepEyes-7B (Zheng et al., 2025), Mini-o3-7B-v1 (Lai et al., 2025a), and PixelReasoner (Su et al., 2025a)), our method achieves consistently superior performance across all datasets, underscoring the effectiveness of incorporating external segmentation models as tools. (3) Compared with SOTA medical-specific MLLMs (LLaVA-Med-7B (Li et al., 2023) and HuatuoGPT-Vision-34B (Chen et al., 2024a)) and medical reasoning MLLMs (e.g., MedVLM-R1-2B (Pan et al., 2025), Med-R1-2B (Lai et al., 2025b), Chiron-o1-

378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431

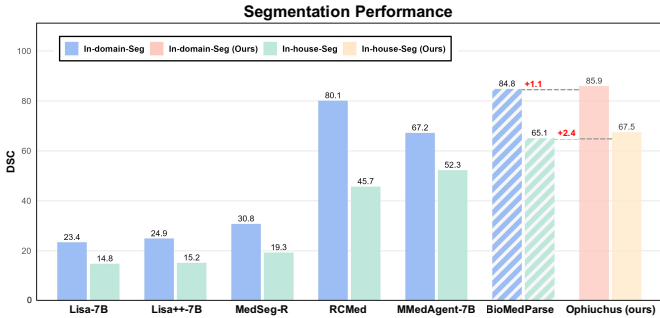


Figure 3: **Performance comparison on medical Segmentation benchmarks.** Colors indicate datasets (In-domain-Seg / In-house-Seg). We also provide performance of task-specific SOTA BiomedParse for reference.

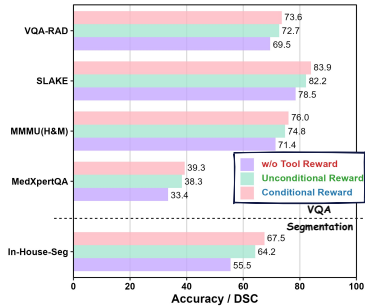


Figure 4: **Ablation on Reward Design.** Results are shown separately for VQA and Segmentation.

8B (Sun et al., 2025a), and Lingshu-7B (Xu et al., 2025)), **Ophiuchus** achieves at least a 10.8% average improvement, validating that tool-augmented reasoning enhances MLLMs’ ability to capture fine-grained information and leverage key visual evidence for medical question answering. (4) Relative to SOTA medical MLLM agents that also can invoke tools (MMedAgent-7B (Li et al., 2024), VILA-M3-40B (Nath et al., 2025)) or rely on multi-agent collaboration (MMedAgent-RL-7B Xia et al. (2025)), **Ophiuchus** outperforms competing agents by a substantial margin, further validating the effectiveness of our tool-augmented vision-language reasoning. Rather than merely calling tools to complete a task, we integrate tool use into step-by-step reasoning, enabling the model to compose multiple tools and reflect on tool outputs for better visual understanding.

Segmentation Performance. We also evaluate segmentation performance to further validate **Ophiuchus**’s ability to capture fine-grained regional features. We test on In-domain-Seg and In-house-Seg, two subsets whose VQA items explicitly require pixel-level segmentation outputs. Results are shown in Figure 3. We compare **Ophiuchus** with (i) general-purpose MLLMs specialized for segmentation (Lisa (Lai et al., 2024), Lisa++ (Yang et al., 2023)), (ii) medical MLLMs capable of reasoning-based segmentation (MedSeg-R (Huang et al., 2025), RCMed (Wang et al., 2025a)), and (iii) the medical MLLM agent that can invoke external segmentation models (MMedAgent-7B (Li et al., 2024)). On In-domain-Seg, **Ophiuchus** surpasses all baselines. Because existing MLLMs-and some foundation segmentation models (*e.g.*, SAM2)-are trained on large public datasets, we further evaluate on a fully held-out In-house-Seg set to mitigate data-leakage concerns. **Ophiuchus** remains ahead on this set as well, outperforming prior models by at least 15.2%. We make two key observations. (1) Models that leverage external segmentation models as tools outperform the vast majority of approaches that train the MLLM to perform segmentation independently. (2) Compared with other tool-enabled agents, **Ophiuchus**’s superior segmentation performance is not solely attributable to stronger tools; it arises from the ability to select among functionally similar tools and to reflect on their outputs. Given a query, **Ophiuchus** can choose different segmentation models, assess the results, and-if one model fails-switch to an alternative, capabilities that existing medical agents typically lack. Further analyses of the quantity and types of tools are provided in the Appendix C.2.

We also present a qualitative analysis of the diverse reasoning patterns that emerge in **Ophiuchus**, illustrating how the model integrates visual tools into its reasoning-analogous to expert visual cognition-and adapts their use to task demands. Cases are provided in the Appendix D.

4.3 ABLATION STUDIES

Effectiveness of training strategies. To assess the impact of our training framework, we compare **Ophiuchus** with the following Qwen2.5-VL-7B-based variants: (1) a prompt-driven tool-invocation baseline; (2) \mathcal{M}_{cold} , trained only with cold-start SFT; (3) $\mathcal{M}_{cold+reflect}$, trained with cold-start SFT followed by self-reflection fine-tuning; (4) \mathcal{M}_{ri} , trained only with ATRL; and (5) $\mathcal{M}_{cold+ri}$. The results in Table 2 validate the effectiveness of our three-stage training strategy. Prompt-only method for tool invocation is neither adaptable nor robust and yields minimal gains, whereas progressively adding our proposed training stages produces substantial improvements over the base model, underscoring the necessity of each stage. [To further demonstrate the effectiveness of self-reflection](#)

Table 2: **Impact of Training Stages.** Checkmarks (✓) indicating which stages are used.

Method	Training Stages			Benchmarks					
	SFT	Self-Reflection	ATRL	VQA-RAD	SLAKE	MMMU(H&M)	MedXpertQA	In-House-Seg	Avg.
<i>prompt-driven</i> \mathcal{M}_{base}				63.7	67.6	58.5	23.7	24.9	47.7
\mathcal{M}_{cold}	✓			67.6(+3.9)	76.7(+9.1)	66.2(+7.7)	30.3(+6.6)	52.6(+27.7)	58.7(+11.0)
$\mathcal{M}_{cold-12}$	✓			67.6(+3.9)	76.8(+9.2)	66.3(+7.8)	30.3(+6.6)	52.5(+27.6)	58.8(+11.1)
$\mathcal{M}_{cold+reflect}$	✓	✓		68.9(+5.2)	78.0(+10.4)	70.9(+12.4)	32.0(+8.3)	54.2(+29.3)	60.8(+13.1)
\mathcal{M}_{rl}			✓	71.3(+7.6)	79.5(+11.9)	71.5(+13.0)	34.5(+10.8)	57.3(+32.4)	62.8(+15.1)
$\mathcal{M}_{cold+rl}$	✓		✓	72.5(+8.8)	81.3(+13.7)	73.7(+15.2)	37.1(+13.4)	63.9(+39.0)	65.7(+18.0)
Ophiuchus	✓	✓	✓	73.6(+9.9)	83.9(+16.3)	76.0(+17.5)	39.3(+15.6)	67.5(+42.6)	68.1(+20.4)

Table 3: **Comparison with multi-agent systems.**

Method	\mathcal{D}_{test} -VQA	SLAKE
Multi-Agent Pipeline (BioMedParse + QwenVL-2.5-7B)	37.2	68.9
GT region crop + QwenVL-2.5-7B	39.9	70.8
GT region crop + LLaVA-Med-7B	33.8	53.4
Ophiuchus (ours)	59.4	83.9

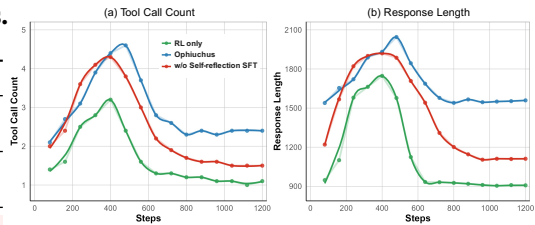


Figure 5: **Training Dynamics.**

fine-tuning, we compare $\mathcal{M}_{cold+reflect}$ (trained for 10 + 2 epochs) with a baseline model $\mathcal{M}_{cold-12}$, which undergoes 12 epochs of standard SFT on the original \mathcal{D}_{cold} dataset. The results confirm that the subsequent improvements primarily stem from fine-tuning on the self-reflection data rather than merely increasing the number of SFT epochs. The RL stage contributes the largest boost, showing that RL’s exploration-exploitation dynamics and reward feedback are critical for learning context-aware tool-use policies and fostering genuine deep reasoning that captures the strategic flexibility required for optimal tool orchestration. Furthermore, when combined with cold-start SFT and self-reflection fine-tuning, RL’s advantages are maximized, further validating our motivation for developing a three-stage training framework to incentivize the model’s tool-augmented reasoning ability.

To visualize how the model’s behavior evolves, we track mean tool-call count and response length over training in Figure 5. **Ophiuchus** shows an expand-compress pattern: both metrics rise early as the model explores tools and tests hypotheses seeded by cold-start SFT and self-reflection fine-tuning, then decline as RL consolidates experience, invokes only necessary tools, and stops once evidence suffices. The variant without self-reflection exhibits weaker early expansion, and the RL-only variant underuses tools throughout. Overall, the three-stage recipe first encourages broad tool-use exploration and then distills it into efficient, context-aware tool orchestration.

Importance of the strategic tool-use reward. Our strategic tool-use reward includes a conditional tool-invocation bonus granted only when the model answers correctly and employs tools. This mechanism incentivizes effective, purposeful tool use. We ablate the reward by (i) removing it entirely (w/o tool reward) and (ii) making it unconditional on accuracy. Results in Figure 4 show that omitting the tool reward causes a substantial performance drop, underscoring its importance. Moreover, the conditional reward yields the highest accuracy, outperforming the other settings. These findings indicate that rewarding tool use alone is insufficient; aligning rewards with successful outcomes is what truly drives intelligent and effective behavior in **Ophiuchus**.

Is the performance gain merely due to tool invocation? To probe this question, we compare **Ophiuchus** with workflow-style, multi-agent variants. (1) BioMedParse + Qwen2.5-VL-7B: a pipeline that pre-segments target regions, enlarges them, and feeds the crops to Qwen2.5-VL-7B for analysis. (2) GT-crop baselines: Qwen2.5-VL-7B and LLaVA-Med-7B receive crops from ground-truth bounding boxes. The results are shown in Table 3. **Ophiuchus** substantially outperforms these variants. While pre-segmentation can improve spatial focus, the static, predefined workflow limits generalization and constrains the system to specific tasks. Thus, the gains are not merely a consequence of exposing fine-grained regions; they stem from **Ophiuchus**’s ability to integrate tool calls and visual details into its reasoning, yielding more accurate diagnostic decision-making. Crucially, this capability arises from our training framework, whereas prompting alone cannot achieve it.

Table 4: Ablation study of segmentation tool variants integrated into Ophiuchus. Default setting (SAM2) is highlighted.

Method	VQA-RAD	SLAKE	MMMU (H&M)	MedXpertQA	In-House-Seg	Avg.
Ophiuchus + MedSAM (Ma et al., 2024)	72.9	83.1	75.4	38.8	66.5	67.3
Ophiuchus + MedSAM2 (Ma et al., 2025)	73.6	83.8	75.9	39.5	67.3	68.0
Ophiuchus + SAM-Med2D (Ye et al., 2023)	73.3	83.6	75.7	38.9	67.1	67.7
Ophiuchus + SAM (Kirillov et al., 2023)	73.1	83.3	75.5	38.6	66.8	67.5
Ophiuchus + SAM2 (default)	73.6	83.9	76.0	39.3	67.5	68.1

Impact of Changing Segmentation Tools on the Agent’s Performance. To investigate how sensitive Ophiuchus is to the inherent accuracy of its underlying segmentation tools (e.g., SAM2), we evaluate how sharply its performance changes when these tools are replaced with alternative ones. To this end, we conduct experiments in which SAM2—the default segmentation module in Ophiuchus—is replaced with SAM (Kirillov et al., 2023), MedSAM (Ma et al., 2024), SAM-Med2D (Ye et al., 2023), or MedSAM2 (Ma et al., 2025), respectively. Because these tools offer similar functionalities and support the same input parameter formats, Ophiuchus can use them in a zero-shot manner without requiring any additional training. The results are summarized in Table 4. Our comparisons show that replacing SAM 2 with different SAM-based segmentation tools results in only minor performance variations. This is due to Ophiuchus’s strong visual feature understanding and grounding capabilities: as long as the MLLM provides an accurate bounding box as the visual prompt, these SAM-based tools can generally produce reliable segmentation outputs. This observation further underscores the robustness of Ophiuchus. Moreover, more capable segmentation tools naturally produce better segmentation performance. Ophiuchus possesses sufficient tool-understanding to directly leverage upgraded segmentation tools as long as their functional descriptions and parameter specifications are provided, thereby eliminating the need for any retraining.

5 CONCLUSION

In this paper, we propose **Ophiuchus**, a new MLLM with the ability of interleaved visual-language reasoning through invoking external tools for fine-grained visual perception and emulation of expert-like diagnostic behavior in medical image analysis. Unlike prior works, **Ophiuchus** offers several key advantages, including advanced tool orchestration, enhanced generalization, and multimodal CoT capabilities. Through careful empirical validation, we demonstrate that our three-stage training framework successfully cultivates sophisticated reasoning pattern. Extensive experiments across diverse medical benchmarks validate the effectiveness of our approach, showing particular strengths in fine-grained medical visual reasoning tasks. We hope **Ophiuchus** can serve as a foundational framework for advancing dynamic, tool-augmented visual reasoning, helping the community develop medical AI agents that can genuinely “think with images” with tools.

ETHICS STATEMENT

This work does not involve experiments with human subjects, animal testing, or personally identifiable information. All training datasets used in this study are derived from publicly available sources. Regarding the in-house test dataset, this retrospective study was approved by the ethics committee of our collaborating hospital, and the requirement for informed consent was waived. All personally identifiable information and private details were removed. The datasets, methodologies, and results presented here do not pose foreseeable risks of harm. All contributions have been conducted in accordance with the ICLR Code of Ethics, with full compliance to guidelines.

REPRODUCIBILITY STATEMENT

We have made efforts to ensure that the results presented in this paper are fully reproducible. A complete description of our experimental setup, including hyperparameter configurations, model architectures, and training procedures, is provided in Section 4.1. To facilitate reproducibility, we will release all code, models, and processed datasets, together with instructions for running experiments.

REFERENCES

- 540
541
542 Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Ale-
543 man, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical
544 report. *arXiv preprint arXiv:2303.08774*, 2023.
- 545 Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo
546 Oh, Lei Ji, Eric Chang, Tackeun Kim, et al. Ehrxqa: A multi-modal question answering dataset
547 for electronic health records with chest x-ray images. *Advances in Neural Information Processing*
548 *Systems*, 36:3867–3880, 2023.
- 549 Fan Bai, Yuxin Du, Tiejun Huang, Max Q-H Meng, and Bo Zhao. M3d: Advancing 3d medical
550 image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*, 2024.
551
- 552 Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang,
553 Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*,
554 2025.
- 555 Pedro RAS Bassi, Mehmet Can Yavuz, Kang Wang, Xiaoxi Chen, Wenxuan Li, Sergio Decherchi,
556 Andrea Cavalli, Yang Yang, Alan Yuille, and Zongwei Zhou. Radgpt: Constructing 3d image-text
557 tumor datasets. *arXiv preprint arXiv:2501.04678*, 2025.
- 558 Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xi-
559 dong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, et al. Huatuoogpt-vision, towards injecting medi-
560 cal visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*, 2024a.
561
- 562 Xupeng Chen, Zhixin Lai, Kangrui Ruan, Shichu Chen, Jiaxiang Liu, and Zuozhu Liu. R-
563 llava: Improving med-vqa understanding through visual region of interest. *arXiv preprint*
564 *arXiv:2410.20327*, 2024b.
- 565 Z Chen, M Varma, JB Delbrouck, M Paschali, L Blankemeier, D Van Veen, JMJ Valanarasu,
566 A Youssef, J Paul Cohen, EP Reis, et al. Chexagent: Towards a foundation model for chest
567 x-ray interpretation, arxiv, 2024. *arXiv preprint arXiv:2401.12208*, 2024c.
- 568 Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit
569 Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the
570 frontier with advanced reasoning, multimodality, long context, and next generation agentic capa-
571 bilities. *arXiv preprint arXiv:2507.06261*, 2025.
- 572 Ziyang Deng, Ruihan He, Jiaxiang Liu, Yuan Wang, Zijie Meng, Songtao Jiang, Yong Xie, and Zuozhu
573 Liu. Med-glip: Advancing medical language-image pre-training with large-scale grounded
574 dataset. *arXiv preprint arXiv:2508.10528*, 2025.
- 575 Ziqing Fan, Cheng Liang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Chestx-reasoner:
576 Advancing radiology foundation models with reasoning through step-by-step verification. *arXiv*
577 *preprint arXiv:2504.20930*, 2025.
- 578 Nima Fathi, Amar Kumar, and Tal Arbel. Aura: A multi-modal medical agent for understanding,
579 reasoning and annotation. In *International Workshop on Agentic AI for Medicine*, pp. 105–114.
580 Springer, 2025.
- 581
582 Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu,
583 Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms
584 via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- 585 Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions
586 for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.
- 587
588 Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum.
589 Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base
590 model. *arXiv preprint arXiv:2503.24290*, 2025.
- 591
592 Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimedvqa:
593 A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the*
IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22170–22183, 2024.

- 594 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. *arXiv preprint arXiv:2506.10465*, 2025.
- 595
596
597
- 598 Yankai Jiang, Wenhui Lei, Xiaofan Zhang, and Shaoting Zhang. Unleashing the potential of vision-language pre-training for 3d zero-shot lesion segmentation via mask-attribute alignment. *arXiv preprint arXiv:2410.15744*, 2024.
- 599
600
- 601 Qiao Jin, Zhizheng Wang, Yifan Yang, Qingqing Zhu, Donald Wright, Thomas Huang, Nikhil Kandekar, Nicholas Wan, Xuguang Ai, W John Wilbur, et al. Agentmd: Empowering language agents for risk prediction with large-scale clinical tool learning. *Nature Communications*, 16(1):9377, 2025.
- 602
603
604
- 605 Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.
- 606
607
608
- 609 Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.
- 610
611
612
- 613 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*, pp. 9579–9589, 2024.
- 614
615
- 616 Xin Lai, Junyi Li, Wei Li, Tao Liu, Tianjian Li, and Hengshuang Zhao. Mini-o3: Scaling up reasoning patterns and interaction turns for visual search. *arXiv preprint arXiv:2509.07969*, 2025a.
- 617
618
- 619 Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, and Xiaofeng Yang. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *arXiv preprint arXiv:2503.13939*, 2025b.
- 620
621
- 622 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.
- 623
624
- 625 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 multi-modal agent. *arXiv preprint arXiv:2407.02483*, 2024.
- 626
627
- 628 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.
- 629
630
631
- 632 Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wanggui He, Hao Jiang, Mengze Li, Xiaohui Song, et al. Healthgpt: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. *arXiv preprint arXiv:2502.09838*, 2025.
- 633
634
635
636
- 637 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)*, pp. 1650–1654. IEEE, 2021.
- 638
639
640
- 641 Fenglin Liu, Jing Wu, Hongjian Zhou, Xiao Gu, Soheila Molaei, Anshul Thakur, Lei Clifton, Honghan Wu, and David A Clifton. Riskagent: Autonomous medical ai copilot for generalist risk prediction. *arXiv preprint arXiv:2503.03802*, 2025.
- 642
643
- 644 Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024.
- 645
646
- 647 Xinyao Liu and Diping Song. Constructing ophthalmic mllm for positioning-diagnosis collaboration through clinical cognitive chain reasoning. *arXiv preprint arXiv:2507.17539*, 2025.

- 648 Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical
649 images. *Nature Communications*, 15(1):654, 2024.
- 650
- 651 Jun Ma, Zongxin Yang, Sumin Kim, Bihui Chen, Mohammed Baharoon, Adibvafa Fallahpour, Reza
652 Asakereh, Hongwei Lyu, and Bo Wang. Medsam2: Segment anything in 3d medical images and
653 videos. *arXiv preprint arXiv:2504.03600*, 2025.
- 654
- 655 Vishwesh Nath, Wenqi Li, Dong Yang, Andriy Myronenko, Mingxin Zheng, Yao Lu, Zhijian Liu,
656 Hongxu Yin, Yee Man Law, Yucheng Tang, et al. Vila-m3: Enhancing vision-language models
657 with medical expert knowledge. In *Proceedings of the Computer Vision and Pattern Recognition
658 Conference*, pp. 14788–14798, 2025.
- 659 OpenAI. Gpt-5 chat. <https://chat.openai.com>, 2025. Accessed: 2025-09-17.
- 660
- 661 Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng
662 Ouyang, and Daniel Rueckert. Medvllm-r1: Incentivizing medical reasoning capability of vision-
663 language models (vlms) via reinforcement learning. *arXiv preprint arXiv:2502.19634*, 2025.
- 664
- 665 Sebastián Ramírez and FastAPI contributors. Fastapi: High-performance web framework for build-
666 ing apis with python. <https://github.com/fastapi/fastapi>, 2018–2025. Accessed:
2025-09-17.
- 667
- 668 Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham
669 Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images
670 and videos. *arXiv preprint arXiv:2408.00714*, 2024.
- 671
- 672 Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng,
673 Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings
674 of the Twentieth European Conference on Computer Systems*, pp. 1279–1297, 2025.
- 675
- 676 Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Jieyu Zhang, Hang Wu, Yuanda Zhu, Joyce C Ho, Carl
677 Yang, and May Dongmei Wang. Ehragent: Code empowers large language models for few-shot
678 complex tabular reasoning on electronic health records. In *Proceedings of the 2024 Conference
679 on Empirical Methods in Natural Language Processing*, pp. 22315–22339, 2024.
- 680
- 681 Zhongyi Shui, Jianpeng Zhang, Weiwei Cao, Sinuo Wang, Ruizhe Guo, Le Lu, Lin Yang, Xianghua
682 Ye, Tingbo Liang, Qi Zhang, et al. Large-scale and fine-grained vision-language pre-training for
683 enhanced ct image understanding. *arXiv preprint arXiv:2501.14548*, 2025.
- 684
- 685 Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhui Chen. Pixel reasoner: In-
686 centivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint
687 arXiv:2505.15966*, 2025a.
- 688
- 689 Yanzhou Su, Tianbin Li, Jiyao Liu, Chenglong Ma, Junzhi Ning, Cheng Tang, Sibao Ju, Jin Ye,
690 Pengcheng Chen, Ming Hu, et al. Gmai-vl-r1: Harnessing reinforcement learning for multimodal
691 medical reasoning. *arXiv preprint arXiv:2504.01886*, 2025b.
- 692
- 693 Haoran Sun, Yankai Jiang, Wenjie Lou, Yujie Zhang, Wenjie Li, Lilong Wang, Mianxin Liu, Lei
694 Liu, and Xiaosong Wang. Enhancing step-by-step and verifiable medical reasoning in mllms.
695 *arXiv preprint arXiv:2506.16962*, 2025a.
- 696
- 697 Yue Sun, Limei Wang, Gang Li, Weili Lin, and Li Wang. A foundation model for enhancing mag-
698 netic resonance images and downstream segmentation, registration and diagnostic tasks. *Nature
699 Biomedical Engineering*, 9(4):521–538, 2025b.
- 700
- 701 Dajuan Wang, Tao Cheng, Sutong Wang, Youhua Frank Chen, and Yunqiang Yin. Smr-agents:
Synergistic medical reasoning agents for zero-shot medical visual question answering with mllms.
Information Processing & Management, 63(1):104297, 2026.
- Haonan Wang, Jiayi Mao, Lehan Wang, Qixiang Zhang, Marawan Elbatel, Yi Qin, Huijun Hu,
Baoxun Li, Wenhui Deng, Weifeng Qin, et al. Reinforced correlation between vision and lan-
guage for precise medical ai assistant. *arXiv preprint arXiv:2505.03380*, 2025a.

- 702 Pengyu Wang, Shuchang Ye, Usman Naseem, and Jinman Kim. Mrgagents: A multi-agent frame-
703 work for improved medical report generation with med-llms. *arXiv preprint arXiv:2505.18530*,
704 2025b.
- 705 Ziyue Wang, Junde Wu, Linghan Cai, Chang Han Low, Xihong Yang, Qiaxuan Li, and Yueming
706 Jin. Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic
707 workflow. *arXiv preprint arXiv:2503.18968*, 2025c.
- 708 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny
709 Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in*
710 *neural information processing systems*, 35:24824–24837, 2022.
- 711 Peng Xia, Jinglu Wang, Yibo Peng, Kaide Zeng, Xian Wu, Xiangru Tang, Hongtu Zhu, Yun Li,
712 Shujie Liu, Yan Lu, et al. Mmedagent-rl: Optimizing multi-agent collaboration for multimodal
713 medical reasoning. *arXiv preprint arXiv:2506.00555*, 2025.
- 714 Yunfei Xie, Ce Zhou, Lang Gao, Juncheng Wu, Xianhang Li, Hong-Yu Zhou, Sheng Liu, Lei Xing,
715 James Zou, Cihang Xie, et al. Medtrinity-25m: A large-scale multimodal dataset with multigran-
716 ular annotations for medicine. *arXiv preprint arXiv:2408.02900*, 2024.
- 717 Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Cheng-
718 hao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: A generalist foun-
719 dation model for unified multimodal medical understanding and reasoning. *arXiv preprint*
720 *arXiv:2506.07044*, 2025.
- 721 Lin Yang, Shawn Xu, Andrew Sellergren, Timo Kohlberger, Yuchen Zhou, Ira Ktena, Atilla Ki-
722 raly, Faruk Ahmed, Farhad Hormozdiari, Tiam Jaroensri, et al. Advancing multimodal medical
723 capabilities of gemini. *arXiv preprint arXiv:2405.03162*, 2024.
- 724 Senqiao Yang, Tianyuan Qu, Xin Lai, Zhuotao Tian, Bohao Peng, Shu Liu, and Jiaya Jia. Lisa++:
725 An improved baseline for reasoning segmentation with large language model. *arXiv preprint*
726 *arXiv:2312.17240*, 2023.
- 727 Jin Ye, Junlong Cheng, Jianpin Chen, Zhongying Deng, Tianbin Li, Haoyu Wang, Yanzhou Su,
728 Ziyang Huang, Jilong Chen, Lei Jiang, et al. Sa-med2d-20m dataset: Segment anything in 2d
729 medical imaging with 20 million masks. *arXiv preprint arXiv:2311.11969*, 2023.
- 730 Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens,
731 Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multi-
732 modal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF*
733 *Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.
- 734 Theodore Zhao, Yu Gu, Jianwei Yang, Naoto Usuyama, Ho Hin Lee, Tristan Naumann, Jianfeng
735 Gao, Angela Crabtree, Jacob Abel, Christine Mounq-Wen, et al. Biomedparse: a biomedical
736 foundation model for image parsing of everything everywhere all at once. *arXiv preprint*
737 *arXiv:2405.12971*, 2024.
- 738 Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and
739 Xing Yu. Deepeyes: Incentivizing” thinking with images” via reinforcement learning. *arXiv*
740 *preprint arXiv:2505.14362*, 2025.
- 741 Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen
742 Duan, Weijie Su, Jie Shao, et al. Internv13: Exploring advanced training and test-time recipes for
743 open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025a.
- 744 Muzhi Zhu, Yuzhuo Tian, Hao Chen, Chunlun Zhou, Qingpei Guo, Yang Liu, Ming Yang, and
745 Chunhua Shen. Segagent: Exploring pixel understanding capabilities in mllms by imitating hu-
746 man annotator trajectories. In *Proceedings of the Computer Vision and Pattern Recognition Con-*
747 *ference*, pp. 3686–3696, 2025b.
- 748 Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding,
749 and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding.
750 *arXiv preprint arXiv:2501.18362*, 2025.

A DATASET DETAILS

A.1 DETAILS OF OUR CURATED DATASETS

Our data collection follows three core principles: (1) Diverse Tasks and Imaging Distributions. We incorporate varied data to strengthen model generalization. (2) Tool Effectiveness. We prioritize scenarios in which tool use yields measurable accuracy gains. (3) Comprehensive Fine-Grained Annotations. We select datasets that include fine-grained region masks and corresponding descriptions, providing the supervision needed to improve the model’s visual reasoning. Consequently, we leverage datasets from BiomedParseData (Zhao et al., 2024), which comprises 3.4 million triples of image, segmentation mask, and semantic label, encompass 82 major biomedical object types across 9 imaging modalities. We also include the Malenia dataset (Jiang et al., 2024), which provides 1,514 image-mask-report triplets across 12 lesion categories with fine-grained descriptions of disease attributes including location, lesion shape, density, density variations, and surface characteristics. These two datasets provide rich fine-grained region annotations and detailed textual descriptions. This abundance of information enables us to construct precise VQA samples and reasoning trajectories. Based on these two datasets, we synthesize queries that specifically require the localization of fine-grained visual cues. Concretely, we prompt Gemini-2.5-pro (Comanici et al., 2025) to generate a QA pair conditioned on the image, mask, and the mask’s descriptions, yielding a fine-grained question that necessitates localizing the specified mask region. We further eliminate data that cannot be properly verified, such as questions with incorrect answers. The prompts for VQA generation and verification are provided in Figure 17 and Figure 18. As described in Section 3.3, we then leverage GPT-5 (OpenAI, 2025) to generate reasoning paths following our proposed tool-augmented reasoning paradigm. The generated paths are subsequently filtered based on format checking and answer correctness. The prompt template for constructing reliable agentic reasoning trajectories and the prompt template for trajectory verification are provide in Figure 19 and Figure 20.

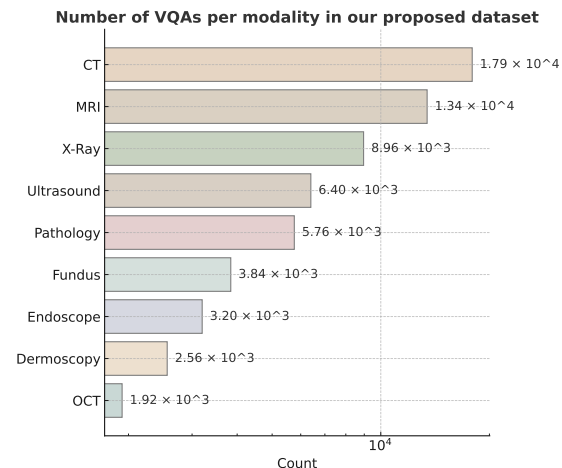


Figure 6: **Distribution of modalities in our proposed dataset.**

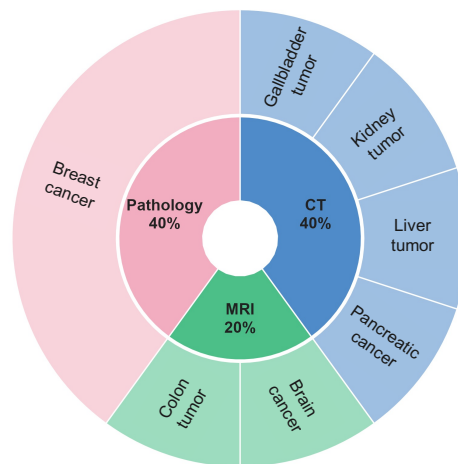


Figure 7: **Distribution of modalities and disease types in the in-house dataset.**

Our data selection strategy encompasses two key steps:(1) We select instances where GPT-5 produces incorrect answers in single-turn interactions but achieves correct results when using tool invocation for fine-grained region inspection, highlighting scenarios where tool use is most beneficial. (2) We also include a small subset of questions that GPT-5 answers correctly without tools, aiming at teaching the model when to rely on its internal capabilities instead of always invoking external tools. Through systematic curation, we obtain 64k VQA samples spanning multiple task types and modalities. To ensure reliability, these samples are further cross-validated against human annotations. Our dataset encompasses a highly diverse set of question types, including: multi-organ and multi-disease recognition and localization; determination of the presence of abnormal lesions; classification and counting of lesion categories; assessment of lesion characteristics, morphology, and intensity variations; as well as organ and lesion segmentation—challenging tasks that more closely reflect real-world clinical diagnosis. Figure 6 presents a bar plot of the number of VQA items for each modality

Table 5: Numbers of VQA samples and segmentation samples in testing sets.

	VQA tasks		Segmentation	
	$\mathcal{D}_{\text{test-VQA}}$	In-House-VQA	$\mathcal{D}_{\text{test-Seg}}$	In-House-Seg
	3140	653	860	347

in our dataset. The dataset comprises nine medical imaging modalities: histopathology (9%), CT (28%), MRI (21%), ultrasound (10%), X-ray (14%), fundus photography (6%), dermatology (4%), endoscopy (5%), and optical coherence tomography (OCT; 3%). For each modality, we split the data in fixed proportions: 30k samples for $\mathcal{D}_{\text{cold}}$ used in cold-start supervised fine-tuning, 30k samples for \mathcal{D}_{rl} used in reinforcement learning training, and 4k for in-domain testing $\mathcal{D}_{\text{test}}$. Within $\mathcal{D}_{\text{cold}}$ (used for cold-start supervised fine-tuning), we apply the self-reflection sampling strategy introduced in Section 3.2 to select cases exhibiting self-correction behaviors. This yields a dataset $\mathcal{D}_{\text{reflect}}$ for self-reflection fine-tuning comprising 3k VQA instances.

A.2 DETAILS OF THE IN-HOUSE TESTSET

For testing, in addition to our in-domain 4k testset $\mathcal{D}_{\text{test}}$, we use a completely held-out in-house dataset comprising 1k CT, MRI, and histopathology images across 7 cancer types for out-of-domain evaluation. In this held-out set, CT images include 100 liver tumor cases, 100 gallbladder tumor cases, 100 pancreatic cancer cases, and 100 kidney tumor cases. MRI images include 100 colon tumor cases and 100 brain cancer cases. Histopathology images include 400 breast cancer cases. Figure 7 depicts the data distribution of the in-house dataset. For each case, human annotators construct a VQA pair along with a corresponding reasoning trajectory. To evaluate our method on both VQA and segmentation tasks, we explicitly require the model to output segmentation results for a subset of questions, rather than using segmentation solely as an intermediate step for reasoning. This further partitions the test set into VQA and segmentation subsets. Table 5 reports the exact counts of VQA and segmentation questions in both the in-domain $\mathcal{D}_{\text{test}}$ and the in-house test set.

A.3 INTRODUCTION OF PUBLIC BENCHMARKS

We also conduct experiments on 6 representative datasets. PathVQA He et al. (2020), SLAKE Liu et al. (2021), and VQA-RAD Lau et al. (2018) are widely used benchmarks in medical VQA research. OmniMedVQA Hu et al. (2024) constructs multiple medical classification datasets into QA form, focusing on image classification tasks. For higher-level reasoning, MMMU-Med Yue et al. (2024) and MedXpertQA Zuo et al. (2025) offer more challenging QA scenarios. MMMU-Med is a medical subset extracted from the multimodal reasoning benchmark MMMU Yue et al. (2024). MedXpertQA presents a more difficult setting, simulating real clinical licensing exams to assess whether models can perform medical reasoning and decision-making at a near-expert level.

A.4 FILTERING AND VERIFICATION PIPELINE FOR QA PAIRS AND REASONING TRACES

Regarding QA generation, we provide Gemini-2.5-Pro with the original image, the ground-truth mask, and the corresponding class label, and explicitly instruct it only to convert the existing mask and label into a QA format, without introducing any additional information. After QA generation, we perform a second layer of strict verification using GPT-5, which evaluates each QA item along the following five dimensions:

1. Schema & Format

- JSON format must be valid and parseable.
- Required keys `Question`, `Options`, and `Answer` must be present.
- Options must contain four or five choices labeled strictly A–D (and optionally E).

2. Independence & Wording

- The question must be self-contained and avoid referential expressions.

- 864
- Wording must use professional, precise clinical terminology with no vague or colloquial phrasing.
- 865
- 866
- 867
- 868
- 869
- 870
- 871
- 872
- 873
- 874
- 875
- 876
- 877
- 878
- 879
- 880
- 881
- 882
- 883
- 884
- 885
- 886
- 887
- 888
- 889
- 890
- 891
- 892
- 893
- 894
- 895
- 896
- 897
- 898
- 899
- 900
- 901
- 902
- 903
- 904
- 905
- 906
- 907
- 908
- 909
- 910
- 911
- 912
- 913
- 914
- 915
- 916
- 917
- The question must be answerable from the image with focused attention on the segmented target region (guided by the bounding box and mask).
 - High semantic correlation with the masked area is required; content must pertain to structures or findings within or immediately adjacent to the target region.
 - No hallucinated information is allowed—only content derivable from the image, mask, and mask descriptions is permitted (no modality, demographic, or history hallucinations unless explicitly provided).
3. **Grounding to the Target Region**
- The question must be answerable from the image with focused attention on the segmented target region (guided by the bounding box and mask).
 - High semantic correlation with the masked area is required; content must pertain to structures or findings within or immediately adjacent to the target region.
 - No hallucinated information is allowed—only content derivable from the image, mask, and mask descriptions is permitted (no modality, demographic, or history hallucinations unless explicitly provided).
4. **Options Quality**
- Options must be mutually exclusive, non-overlapping, and clinically plausible.
 - No duplicate or near-duplicate options.
5. **Answer Correctness**
- GPT-5 re-evaluates the image with the bounding box/mask and mask descriptions to determine the true correct answer.
 - It verifies that the provided `Answer` corresponds to the correct option.

The full verification prompt used by GPT-5 for this VQA quality check is provided in Figure 18.

Regarding the generated reasoning traces, we also apply a rigorous verification procedure. Specifically, we use GPT-o3 to examine each trace according to the following evaluation rules:

- **Format:** The trace must interleave `<think>` \rightarrow `<tool_call>` \rightarrow `<obs>` steps (repeated as needed) and conclude with `Action: Answer` and an `<answer>` block.
- **Tool usage:** Only the allowed tools (`image_zoom_in_tool`, `SAM2`, `BioMedParse`) may be invoked, and all tool calls must contain valid JSON arguments.
- **Tool choice rationale:** Each `<tool_call>` must be explicitly justified by the immediately preceding `<think>` step (i.e., why this tool is selected and why these arguments are appropriate).
- **Observation integration:** Each `<obs>` must be meaningfully incorporated into the subsequent `<think>`; the model should not ignore or contradict the tool outputs.
- **Clinical correctness:** The reasoning must be consistent with the image/ROI/mask and the mask description, with no hallucinated findings or unsupported clinical claims.
- **Final answer:** The content of `<answer>` must exactly match the provided ground-truth answer.

All tool calls in the reasoning traces are also executed in the actual environment, and the resulting outputs are used to verify their consistency, for example, checking whether the mask covers the correct region. Any hallucinated content that conflicts with the underlying structured annotations is removed from the dataset. Through this ground-truth-guided trajectory construction and step-wise trajectory validation, we ensure that all retained data are of consistently high quality. The full verification prompt used by GPT-o3 for the reasoning trace quality check is provided in Figure 20.

After all the above checks, all QA data and reasoning trajectories are further cross-verified by human annotators.

B FORMULA DEFINITION OF REWARD SIGNALS

We adopt a diverse set of reward designs to provide rich and fine-grained feedback signals. The specific definitions of each reward are detailed as follows.

The reasoning-format reward. The reasoning-format reward \mathcal{S}_{format} evaluates the structural validity of R by verifying that the model’s output includes all required special tokens in the prescribed order. Specifically, the model should enclose its chain-of-thought between `<think>` and `<think>`

tags, place the tool-call choices and parameters between `<tool_call>` and `<tool_call>` tags, and place the final answer between `<answer>` and `<answer>` tags. Outputs that adhere to this structure receive a positive reward.

$$S_{\text{format}} = \begin{cases} 1, & \text{if all required fields appear and are in the correct order,} \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

The final-answer reward. The final-answer reward S_{ans} encompasses multiple task types, thereby providing the agent with diverse feedback. For multiple-choice questions, we simply check the exact match between the predicted and answers:

$$S_{\text{ans}}(A, \hat{A}) = \mathbb{I}(A = \hat{A}), \text{ for multiple-choice questions,} \quad (6)$$

A denote the ground-truth answer and \hat{A} is the predicted answer obtained by rule-based parsing of the model’s final output. The indicator \mathbb{I} is defined to be 1 if $A = \hat{A}$ and 0 otherwise.

For segmentation tasks, in contrast to earlier reward designs, we use SAM2 or BiomedParse as external reward providers. Given either a target location (bounding box) or a text prompt specifying the object category predicted by the MLLM, we query SAM2 or BiomedParse to obtain a segmentation mask. We then compute the intersection-over-union (IoU) between this mask and the ground-truth mask and assign piecewise rewards as follows:

$$\text{reward} = \begin{cases} 3, & \text{IoU} > 0.80, \\ 2, & 0.70 < \text{IoU} \leq 0.80, \\ 1, & 0.50 < \text{IoU} \leq 0.70, \\ 0, & \text{otherwise.} \end{cases} \quad (7)$$

This segmentation reward supplies strong positive feedback only when the predicted region closely matches the ground truth, while at lower IoU levels it provides guidance that encourages gradual improvement.

The strategic tool-use reward. The strategic tool-use reward S_{tool} is defined as a conditional bonus, granted only when the model both produces a correct answer and invokes at least one external perception tool during its trajectory. This design encourages the model to employ tools meaningfully—when they directly contribute to successful task completion—rather than using them arbitrarily or redundantly.

$$S_{\text{tool}} = \mathbb{I}(S_{\text{ans}} > 0) \cdot B_{\text{bonus}}, \quad (8)$$

where $\mathbb{I}(S_{\text{ans}} > 0)$ is the indicator function that equals 1 only when $S_{\text{ans}} > 0$, and B_{bonus} represents the tool-invocation bonus, fixed at 2.

C ADDITIONAL EVALUATION RESULTS

C.1 THE ACCURACY RATE OF TOOL INVOCATION

Table 6 measures reliability by jointly checking format adherence and correct tool invocation. The prompt-driven baseline is unreliable, averaging 25.1% across VQA and segmentation. **Ophiuchus** is near ceiling on all settings with a 97.9% average, reflecting consistently precise tool-use behavior on both in-domain and in-house data. The average absolute gain is about 73%. These results show that the model reliably and proactively invokes tools to solve problems. This further corroborates the importance of our carefully designed multi-regime training for eliciting robust tool-calling capabilities—intelligent behaviors that mere prompting of a base model cannot achieve.

C.2 ABLATION ON TOOL QUANTITY AND TYPE

This section studies how the quantity and type of tools affect performance. Results are shown in Table 7.

Table 6: **Accuracy of Tool Use.** Values are percentages. Accuracy is computed by checking whether the model adheres to the required output format and correctly invokes the specified tools. **Avg.** is the arithmetic mean across the four columns.

Method	VQA tasks		Segmentation		Avg.
	$\mathcal{D}_{\text{test-VQA}}$	In-House-VQA	$\mathcal{D}_{\text{test-Seg}}$	In-House-Seg	
prompt-driven $\mathcal{M}_{\text{base}}$	22.8	23.5	27.9	26.2	25.1
Ophiuchus	97.2	97.7	98.4	98.1	97.9

Table 7: **Ablation on Tool Quantity and Type.** Tools are grouped into *Seen* (seen during training) and *Unseen* (held out). A checkmark (✓) indicates a used tool. **Tools #** denotes the total number of tools actually used in this setting (including unseen MedSAM). **Avg.** is the mean over available metrics per row (if Segmentation result is “–”, meaning no segmentation tool is used and the model itself has no segmentation capability, average over the 4 VQA metrics).

Method	Tools #	Tools (used in this setting)				Benchmarks					Avg.
		Seen		Unseen		VQA tasks			Segmentation		
		Zoom-in SAM2	BioMed Parse	MedSAM	MedSAM	VQA-RAD	SLAKE	MMMU (H&M)	MedXpertQA	In-House-Seg	
DeepEyes-7B	1	✓				65.9	68.2	57.8	23.6	–	53.9
Mini-o3-7B-v1	1	✓				65.7	67.8	57.4	24.3	–	53.8
PixelReasoner-RL-v1-7B	1	✓				66.0	67.3	58.0	23.5	–	53.7
MMedAgent-MedSAM	1			✓		64.0	68.7	44.1	22.3	52.3	50.3
	1	✓				68.0	72.5	60.5	29.3	–	57.6
	1			✓		69.5	78.2	71.4	31.0	63.9	62.8
	1		✓			70.5	78.4	71.6	31.2	64.1	63.2
Ophiuchus	1			✓		71.6	79.8	73.5	33.7	64.5	64.6
	2		✓	✓		71.9	80.4	73.8	34.4	67.0	65.5
	2	✓	✓			72.4	81.9	74.1	37.0	64.8	66.0
	2	✓	✓	✓		72.8	82.7	75.0	38.4	65.0	66.8
Ophiuchus (all tools)	3	✓	✓	✓		73.6	83.9	76.0	39.3	67.5	68.1

Zoom-in-only OPHIUCHUS remain strong. For a fair comparison with DeepEyes, Mini-o3, and PixelReasoner, we enable only the zoom-in tool in Ophiuchus. The zoom-in-only OPHIUCHUS reaches a 57.6 average, outperforming the zoom-in-only baselines that cluster near 53.8. It also leads on VQA-RAD with 68.0 compared with 65.7-66.0 for prior work, and maintains advantages on SLAKE and MMMU(H&M). This indicates that our training induces a deeply tool-integrated reasoning capability even when only local magnification is available.

Training-free generalization to unseen tools. When provided only with the unseen tool MedSAM, **Ophiuchus** achieves an average score of 62.8, which is substantially higher than the pipeline-style MMedAgent that also relies on MedSAM (50.3). The advantage holds consistently across VQA-RAD, SLAKE, MedXpertQA, and In-House-Seg. For instance, **Ophiuchus** reaches 69.5 compared with 64.0 on VQA-RAD and 63.9 compared with 52.3 on In-House-Seg. This demonstrates that **Ophiuchus** can recognize and exploit a new tool in a training-free manner, showing that its tool use is not restricted to the three tools introduced during development.

Tool scaling and complementarity drive consistent gains. Performance improves steadily as the model gains access to more tools, reflecting both a scaling effect and meaningful complementarity. Among single-tool settings, BiomedParse achieves the strongest average at 64.6, slightly ahead of SAM2 at 63.2, and both provide clear gains over zoom-in alone. When combined, the two segmentation tools reach an average of 65.5 and further raise In-House-Seg to 67.0, suggesting improved robustness through their complementary strength. Pairing zoom-in with segmentation tools also strengthens performance, with zoom-in plus SAM2 averaging 66.0 and zoom-in plus BiomedParse averaging 66.8. On MedXpertQA, the increase from 31.2 with SAM2 alone to 38.4 with zoom-in plus BiomedParse illustrates how local inspection and text-driven segmentation reinforce one another in complex reasoning. Equipping the agent with all three tools yields the best overall performance, averaging 68.1 and delivering improvements across both VQA and segmentation. These

Table 8: Segmentation performance of tools.

Method	In-Domain-Seg	In-House-Seg
SAM 2	83.9	65.8
BioMedParse	84.8	65.1
SAM 2 + BioMedParse (ensemble)	85.3	66.2
Ophiuchus-7B (ours)	85.9	67.5

results demonstrate that broader tool access systematically scales performance, while diverse functionalities interact synergistically to expand the available evidence for reasoning.

Adaptive decision-making across tools is essential. Beyond the benefits of scale and diversity, the results also suggest that performance gains arise from the agent’s ability to adaptively decide when and how to use each tool. The consistent improvements across different combinations indicate that the model does not rely on a single dominant utility, but instead learns to coordinate multiple perceptual pathways depending on task demands. This highlights that future progress will hinge not only on expanding tool libraries but also on strengthening adaptive orchestration strategies that allow the agent to dynamically align tools with the underlying reasoning process.

C.3 DO THE TOOL PERFORMANCES CONSTITUTE AN UPPER LIMIT FOR OPHIUCHUS?

Table 8 presents a comparison between Ophiuchus and the standalone performance of the segmentation tools it relies on. In addition to reporting the performance of each individual tool, we also include the results of model ensembles constructed from these tools. Because SAM 2 cannot directly process text queries, we provide it with GT bounding boxes and report its segmentation results accordingly. Notably, Ophiuchus surpasses both tools themselves, showing that the performance of these segmentation tools does not constitute a hard upper bound for our method. Through end-to-end training and optimization, Ophiuchus enhances the MLLM’s self-reflection and decision-making abilities: it can detect when the segmentation output from one tool is unreliable and switch to another tool (see Figure 12). Moreover, Ophiuchus can compose multiple tools—such as combining the zoom-in tool with SAM 2—to further improve segmentation performance (see Figure 14). These experiments further underscore our contribution: by explicitly integrating tool usage into the model’s own chain of thought, we strengthen the MLLM’s visual perception, decision-making, and reasoning capabilities, enabling the emergence of higher-level abilities such as self-reflection and tool composition—capabilities absent in existing “think with images” approaches.

C.4 COMPARISON USING THE SAME EXTERNAL TOOLS AND TRAINING DATA

We provide in the Table 9 a comparable setting in which all LLMs and agents have access to the same external tools and are evaluated under the same training data conditions. Specifically, for close-source baselines, since these models cannot be further trained, we expose to them the full interfaces of all tools used in our system (zoom-in, SAM 2, and BioMedParse). We also explicitly instruct them in the prompt that they may invoke these tools for analysis. In addition, our evaluation code feeds the tool outputs back to these models for subsequent reasoning. For open-source baselines and agents, we not only provide unrestricted access to all tools but also fine-tune them using the complete three-stage training data we constructed with Gemini-2.5-Pro and GPT-5. This enables a fully equitable comparison. Even when evaluated with the same external tools and further fine-tuned on the same datasets, existing baseline methods still fall significantly short of our agent. This confirms that the performance gains of our model do not simply arise from using external tools or training data constructed by external models (Gemini-2.5-Pro and GPT-5). Instead, the improvement primarily comes from our proposed pixel-level visual reasoning paradigm, which explicitly integrates tool invocation and visual evidence into the model’s chain-of-thought. This enables the model to learn when to use a tool and how to compose multiple tools for fine-grained visual analysis.

Importantly, this capability is far from a trivial form of tool use. During fine-tuning of existing baselines, we observed that these models generally lack sufficient understanding of complex vision-based tools. Because tool usage introduces external tokens and additional visual features, current methods struggle to interpret such information, failing to determine the appropriate timing for tool invocation and to effectively utilize the returned visual evidence. As a result, their performance

Table 9: Comparison across 8 VQA benchmarks using the same external tools configuration.

Methods (same tools)	D_{test} -VQA	PathVQA	SLAKE	VQA-RAD	OmniMed	MMMU (H&M)	MedXpertQA	In-House-VQA	Avg
<i>Close-Source SOTA</i>									
GPT-4.1	36.9	58.9	72.4	65.5	75.4	73.1	41.3	25.8	58.6
GPT-5	38.5	61.2	73.9	64.8	75.0	70.3	41.0	27.9	59.9
OpenAI-o3	41.4	67.9	77.5	66.2	73.8	74.1	44.6	30.9	62.2
Gemini 2.5 Pro	41.9	68.2	73.6	64.2	77.2	72.3	46.9	30.5	61.8
<i>Open-Source SOTA</i>									
InternVL3-8B	39.7	54.1	71.0	66.4	72.5	62.7	24.0	21.1	49.0
LLaVA-Next-13B	23.5	40.3	57.9	55.2	58.4	39.2	19.4	18.2	41.4
Qwen2.5-VL-32B	38.7	48.9	70.8	71.9	69.8	61.2	26.9	26.1	53.7
<i>MLLMs can "Think with Images"</i>									
DeepEyes-7B	40.3	53.4	68.7	66.3	65.4	58.2	23.9	21.3	50.3
Mini-o3-7B-v1	40.7	53.8	68.6	66.1	65.9	57.7	24.4	22.1	50.3
PixelReasoner-RL-v1-7B	40.9	54.2	68.4	66.5	65.4	58.4	23.8	22.0	50.4
<i>Medical MLLMs</i>									
LLaVA-Med-7B	27.5	45.0	47.9	52.7	52.3	38.5	19.4	19.2	39.3
HuatuoGPT-Vision-34B	37.9	51.3	68.5	62.0	70.9	59.5	24.2	21.7	51.5
<i>Medical MLLMs with CoT Reasoning</i>									
MedVLM-R1-2B	27.7	38.5	54.1	45.0	68.4	45.2	21.4	19.5	41.7
Med-R1-2B*	18.3	19.6	52.3	36.4	—	44.5	22.6	16.3	—
Lingshu-7B	41.1	69.8	78.5	66.9	75.6	58.4	26.5	26.3	57.4
Chiron-o1-8B	41.3	70.3	78.2	72.8	77.0	56.7	25.9	28.4	58.2
<i>Multimodal Medical Agents</i>									
MMedAgent-7B	40.6	61.5	69.5	64.8	58.9	45.0	23.3	20.8	49.4
MMedAgent-RL-7B	39.5	60.4	68.1	66.3	61.4	59.9	23.5	23.9	50.6
VILA-M3-40B	41.2	68.7	73.9	67.7	57.5	57.6	24.1	28.3	50.0
Qwen2.5-VL-7B	36.2	53.1	67.7	64.5	65.3	57.8	24.7	20.4	47.0
Ophiuchus-7B (Ours)	59.4	74.3	83.9	73.6	78.6	76.0	39.3	50.4	68.0

Table 10: Analysis of Over-Calling of Tools.

Benchmark	Percentage of Excessive Tool-Calling Behavior (%)
D_{test} -VQA (In-domain)	4.8
PathVQA	1.4
SLAKE	4.2
VQA-RAD	0.6
OmniMed	3.5
MMMU (H&M)	3.7
MedXpertQA	3.9
In-House-VQA	4.6

may even degrade on certain test sets. This issue is especially pronounced in smaller-parameter models with weaker instruction-following abilities. Our proposed “think with images and tools” framework is specifically designed to strengthen an MLLM’s comprehension of tools and visual feedback, thereby enabling effective tool reasoning and robust pixel-level visual understanding.

C.5 ANALYSIS OF OVER-CALLING OF TOOLS.

We also include a small subset of questions (3K) that GPT-5 can correctly answer without tools during training, with the goal of teaching the model to invoke tools only when necessary. As a result, our agent directly answers questions that it considers sufficiently simple or solvable using the MLLM’s innate capabilities. We provide an analysis of over-calling of tools in Table 10. Specifically, we first identify, for each benchmark, the set of cases that the base model Qwen2.5-VL-7B can answer correctly without using any tools. We then count how many of these cases Ophiuchus chooses to answer using tools. Finally, we report the proportion of such “directly answerable” cases for which Ophiuchus still decides to use a tool. The results show that for questions the base model can already answer correctly, our agent rarely performs additional tool calls, thereby improving performance while maintaining efficiency.

C.6 STABILITY ANALYSIS OF TRAINING.

In Figure 16, we present the training curves, including the SFT training-loss curve, the RL overall-reward curve, the reasoning-format reward curve, the strategic tool-use reward curve, and the IoU

Table 11: Ablation study on RL hyperparameters (learning rate lr and number of CoT rollouts N_c) evaluated on 8 VQA benchmarks. Default setting is highlighted in bold.

RL Hyperparameters	D.test-VQA	PathVQA	SLAKE	VQA-RAD	OmniMed	MMMU (H&M)	MedXpertQA	In-House-VQA
$lr = 1 \times 10^{-6}, N_c = 4$	59.1	74.0	83.3	73.5	78.6	75.4	39.1	50.2
$lr = 1 \times 10^{-6}, N_c = 8$ (default)	59.4	74.3	83.9	73.6	78.6	76.0	39.3	50.4
$lr = 1 \times 10^{-5}, N_c = 4$	59.0	74.2	83.5	73.1	78.7	75.8	39.0	50.1
$lr = 2 \times 10^{-5}, N_c = 8$	59.2	74.3	83.6	73.4	78.7	76.0	39.2	50.0
$lr = 2 \times 10^{-6}, N_c = 8$	59.3	74.3	83.8	73.5	78.6	75.9	39.2	50.3

Table 12: Performance comparison across EHR-based tasks.

Methods	Image-related	Table-related	Image+Table-related
GPT-5	58.3	93.5	76.8
EHRAgent (Shi et al., 2024)	64.9	94.9	84.7
AgentMD (Jin et al., 2025)	67.2	97.8	88.2
RiskAgent (Liu et al., 2025)	66.8	97.5	87.6
Ophiuchus (Ours)	75.9	98.4	91.4

curve of the final-answer reward. These curves demonstrate that the SFT and RL training dynamics of Ophiuchus exhibit stable and consistent improvements.

we also adjusted the RL hyperparameters and evaluated the resulting performance changes to assess the robustness of Ophiuchus. Specifically, we varied the learning rates lr and the number of roll-out candidate reasoning paths N_c . The detailed results are summarized in the Table 11. The results show that adjustments to the RL hyperparameters lead to only minor performance fluctuations. Overall, our default parameter settings achieve the best performance, indicating that our choices of RL hyperparameters are appropriate and well justified.

C.7 GENERALIZATION TO EHR-BASED TASKS

We have further extended our experiments to EHR-based tasks. Specifically, we adopt EHRXQA (Bae et al., 2023), a multimodal question-answering benchmark for electronic health records that integrates chest X-ray images with structured EHR tables. This benchmark provides a comprehensive set of QA pairs spanning image-related, table-related, and image-plus-table-related questions. We compare Ophiuchus against several strong agents specifically designed for EHR tasks. For a fair comparison, all baselines—including Ophiuchus—use the same: (a) EHR metadata, (b) official EHRXQA training/testing split, (c) tool definitions. In particular, we adopt EHRXQA’s official NeuralSQL-based framework, which integrates LLMs with an external VQA API to handle multimodal queries over a structured database with images. (d) We also follow the benchmark’s official evaluation metric, Prediction Execution Accuracy ($Acc_{EX|pred}$), which measures execution correctness and thus reflects the overall QA pipeline performance, including both parsing and execution. The results are summarized in Table 12. Ophiuchus demonstrates highly competitive performance on EHR tasks. This further confirms that integrating tool invocation with chain-of-thought reasoning greatly enhances the model’s ability to understand tools and to determine when and how to leverage them effectively for enhanced reasoning.

The differences between our method and these text-based tool-augmented LLM Agent are clear: (1) AgentMD and RiskAgent are text-based medical LLM agents primarily designed for risk prediction, whereas our method is a multimodal medical MLLM focused on enhancing fine-grained visual feature understanding. (2) AgentMD and RiskAgent rely on evidence-based tools, such as clinical calculators. In contrast, our approach incorporates vision-oriented tools aimed at image analysis, including localization, segmentation, and zooming. (3) AgentMD and RiskAgent are task-specific agents tailored for particular objectives such as risk prediction. In comparison, our proposed Ophiuchus framework aims to enhance the pixel-level visual reasoning abilities of general-purpose medical MLLMs through tool integration.

C.8 GENERALIZATION TO REPORT GENERATION TASKS.

We adopt MIMIC-CXR (Johnson et al., 2019) for chest X-ray report generation, covering two tasks: finding generation and impression generation. We evaluated Ophiuchus’s performance using

Table 13: Comparison of finding generation performance across multiple models. Ophiuchus-7B achieves the highest scores on all metrics.

Models (Finding Generation)	BERTScore	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
HuatuogPT-Vision-7B	85.4	9.7	27.8	5.3	16.1
Qwen2.5-VL-72B-Instruct	84.6	5.6	22.7	4.2	14.2
CheXagent (Chen et al., 2024c)	74.6	0.3	4.7	0.1	4.3
MedVLM-R1	82.1	5.9	16.8	2.7	12.1
ChestX-Reasoner (Fan et al., 2025)	82.2	4.8	14.4	4.1	11.5
Qwen2.5-VL-7B-Instruct	84.5	7.2	22.4	4.2	14.0
GPT-4o	86.2	11.1	30.4	6.4	18.4
GPT-4o-mini	85.2	6.7	22.8	3.8	14.2
DeepSeek-VL2	82.3	4.4	14.9	3.0	10.6
Chiron-o1-8B	86.2	10.7	26.7	7.4	17.8
Ophiuchus-7B (ours)	87.9	13.5	32.9	8.6	18.3

Table 14: Comparison of impression generation performance across multiple models. Ophiuchus-7B achieves the highest scores on all metrics.

Models (Impression Generation)	BERTScore	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
HuatuogPT-Vision-7B	83.5	2.8	10.6	1.8	7.5
Qwen2.5-VL-72B-Instruct	83.3	2.8	10.2	2.2	6.9
CheXagent (Chen et al., 2024c)	81.4	7.3	9.8	0.0	9.8
MedVLM-R1	81.5	1.9	6.1	0.5	4.8
ChestX-Reasoner (Fan et al., 2025)	83.6	4.5	12.3	4.0	10.8
Qwen2.5-VL-7B-Instruct	83.6	3.7	10.0	2.0	7.2
GPT-4o	84.5	4.1	13.9	3.0	9.8
GPT-4o-mini	83.7	2.5	9.7	1.4	6.6
DeepSeek-VL2	82.4	2.6	7.1	0.9	6.3
Chiron-o1-8B	83.8	3.7	10.3	2.0	8.1
Ophiuchus-7B (ours)	85.7	5.2	16.9	4.8	11.4

BERTScore, BLEU, ROUGE-1, ROUGE-2, and ROUGE-L as metrics. The results for both tasks are shown in Table 13 and Table 14, respectively. It is evident that Ophiuchus achieves superior performance compared with both general-purpose large-scale MLLMs and existing medical MLLMs. This improvement can be attributed to Ophiuchus’s enhanced ability to understand and identify fine-grained visual cues—capabilities that are also essential for accurate and clinically meaningful report generation.

D CASE STUDIES

D.1 SUCCESSFUL EMERGENT THINKING PATTERNS

We identify three primary patterns.

(1) Visual-Cue Search. When a single global observation of the image is insufficient for a complex problem, the model leverages segmentation tools to generate region proposals and employs a zoom-in tool to examine them, aggregating visual cues and reasoning over them to reach a reliable conclusion.

As shown in Figure 8, **Ophiuchus** begins from the full endoscopic view and judges that a global look is insufficient for reliable counting. It first invokes BiomedParse to propose candidate polyp regions (Turn 1), then applies the zoom-in tool to each segmented area for close inspection of morphology and boundaries (Turn 2). Aggregating the zoomed evidence, it confirms two distinct polyps and outputs the correct answer (Turn 3), which exemplifies the visual-cue search pattern.

(2) Visual Confirmation. In some cases, the model starts uncertain but gradually builds confidence by analyzing tool-returned regional observations and zooming in on details to gather evidence and resolve ambiguities.

As shown in Figure 9, given an abdominal CT slice, **Ophiuchus** first performs anatomic elimination to down-weight pancreas, spleen, and renal options, yet remains uncertain about a suspected colonic wall finding. It therefore zooms into a targeted region of interest to verify the local appearance (Turn 1). The enlarged view reveals focal thickening and luminal irregularity (Turn 2), which provides direct confirmation and leads to the final choice. This illustrates how local evidence consolidates an initially uncertain hypothesis.

(3) Hallucination Mitigation. Although MLLMs can sometimes hallucinate, invoking the tools helps the model focus on visual details to mitigate hallucination and produce a more accurate diagnostic conclusion.

As shown in Figure 10, the initial free-view impression includes round morphology, hypodensity, and a tentative heterogeneous interior. To verify and avoid over-interpretation, **Ophiuchus** segments the suspected hepatic lesion with BiomedParse (Turn 1) and then zooms into the mask for careful inspection (Turn 2). The close view shows a well-circumscribed round hypodense mass with homogeneous attenuation, correcting the earlier hypothesis and yielding the final choices (Turn 3), which demonstrates hallucination mitigation through tool-guided re-examination.

D.2 FAILURE CASES

We next analyze a representative hard case that illustrates both the limits and the introspective behavior of **Ophiuchus** (Figure 11). The image exhibits complex appearance with extremely subtle cues, and the pancreatic lesion occupies a tiny, hard-to-discern region. It first correctly interprets the user’s goal-probe for possible pancreatic cancer-and, acknowledging uncertainty, **invokes the external segmentation tool BiomedParse with appropriate parameters**. The tool returns an incorrect mask that misses the true lesion. To validate the tool’s output, it applies the zoom-in tool to the segmented area and inspects local morphology, **after which it rejects the tool hypothesis as inconsistent with pancreatic cancer**. Lacking reliable visual evidence, it ultimately predicts “no pancreatic cancer,” which is wrong in this case. Although the final answer is incorrect and both the model and the tool fail under this ultra-low-signal setting, the trajectory evidences strong capability for intelligent tool invocation and reflective error checking: **Ophiuchus** localizes uncertainty, calls an appropriate tool, cross-examines the tool result via zoom-in verification, and explicitly diagnoses the tool’s failure before committing to a conclusion.

D.3 SELF-REFLECTION EXAMPLES

As shown in Figure 12, **Ophiuchus** demonstrates explicit self-reflection and adaptive tool use. It first applies BiomedParse to detect a pancreatic cyst and inspects the returned mask. Based on anatomic reasoning, the model judges that the highlighted region corresponds to renal parenchyma rather than the pancreas, which signals a likely tool error. It then revisits the original CT, relocates the pancreas, formulates a revised hypothesis of a round hypodense lesion in the pancreatic head, and selects an alternative tool. It parameterizes SAM2 with a bounding box over the suspected lesion, obtains a precise segmentation, and confirms the cystic lesion. This trajectory shows **the model diagnosing a tool failure, switching tools, refining parameters, and verifying the final conclusion through visual evidence**.

E PROMPT

E.1 PROMPT FOR VQA CONSTRUCTION AND VERIFICATION

In this section, we present the prompt templates used for constructing VQA items and for verifying their quality. Figure 17 illustrates the prompt designed to generate VQA questions and answers from medical data, while Figure 18 shows the template used to check the quality and validity of the generated VQA pairs. These prompts are employed throughout data construction to ensure that the resulting dataset is both diverse and reliable. We follow the same template design when constructing segmentation-task data.

1296 E.2 PROMPT FOR REASONING TRAJECTORY CONSTRUCTION

1297
1298 In this section, we introduce the prompts for building reliable agentic reasoning trajectories for
1299 our curated VQA data. Unlike generic approaches that directly sample from a teacher model, we
1300 stabilize generation by providing GPT-5 with ground-truth clinical metadata and tool observations
1301 during prompting while explicitly instructing the agent to treat the answer as unknown throughout
1302 reasoning and reveal it only at the end. Using this prompt together with verified medical labels,
1303 we perform multi-turn rollout with environment interaction and tool use to construct step-by-step
1304 trajectories. Figure 19 presents the generation template, and Figure 20 provides the verification
1305 template that enforces strict checks on trace format, tool correctness, observation integration, clinical
1306 soundness, and exact answer match, with explicit focus on ROI/mask usage and zero tolerance for
1307 invented observations.

1308 E.3 SYSTEM PROMPT AND USER PROMPT

1309
1310 In this section, we present the two prompt templates used in our experiments. Figure 21 shows the
1311 System Prompt template. Figure 22 shows the User Prompt template.

1312 F LIMITATIONS

1313
1314 While **Ophiuchus** and our ATRL yield consistent gains, the system still exhibits notable shortcom-
1315 ings under realistic constraints. Below we outline the key limitations:

1316
1317 First, a primary limitation is cost. Agentic RL with multi-turn rollouts, tool execution, and verifi-
1318 able rewards is compute-intensive, and generating reasoning trajectories consumes substantial API
1319 resources.

1320
1321 Second, the present system remains 2D-centric, which limits performance on tasks that depend on
1322 3D topology and motion.

1323
1324 Third, at present, we employ three tools. In the future, we plan to expand the set of specialized tools
1325 to further enhance the model’s performance and capacity.

1326 G FUTURE WORK

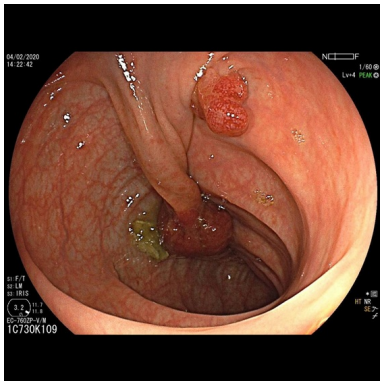
1327
1328 We will broaden the agent’s medically oriented toolset and extend the framework from 2D to volu-
1329 metric and temporal modalities, including MRI, CT, and video.

1330 H STATEMENT ON THE USE OF LLMs FOR PAPER WRITING

1331
1332 We used LLMs (*e.g.*, GPT-5) to aid in polishing the writing of this paper. Specifically, LLMs
1333 were employed to refine grammar, improve fluency, and adjust the tone of academic writing (*e.g.*,
1334 smoothing transitions between sections, shortening overly long sentences, and rephrasing repetitive
1335 expressions). The models were not used for generating technical content, conducting experiments,
1336 or formulating research ideas; all substantive contributions remain the work of the authors.
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349

1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403

Original Image-1:



Question:

How many neoplastic-appearing colonic polyps are visible in this endoscopic image?

Options (single best answer):

- A) One
- B) Two
- C) Three
- D) Four
- E) None

Ground Truth: B

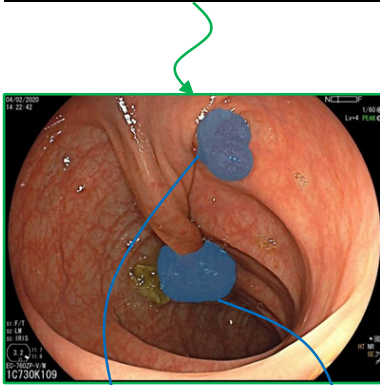


Image-2

Turn 1:

<think>The task is to determine the number of neoplastic-appearing colonic polyps. **Direct visual inspection reveals at least one pedunculated polyp in the central portion of the lumen, but to ensure accurate identification and counting, I will first apply an automated segmentation tool for colonic polyps.** </think>

```
<tool_call>{"name": "BiomedParse",
"arguments": {"description":
"endoscope_colon_polyp", "Image": Image-1}}
</tool_call>
```

<obs>image-2</obs>



Image-3



Image-4

Turn 2:

<think>The BiomedParse output highlights suspected polypoid lesions. To verify their morphology and confirm the count, I will zoom in on the segmented regions for closer inspection. </think>

```
<tool_call>{"name":
"image_zoom_in_tool", "arguments":
{"mask": "image-2"}}
</tool_call>
```

<obs>image-3; image-4</obs>

Turn 3:

<think>The zoomed-in view **confirms two distinct lesions.** Therefore, the total number of colonic polyps visible is two. </think>

<answer>**B**</answer>

Figure 8: **Visual-Cue Search example.**

1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457

Original Image-1:



Question:

On this CT scan, which of the following best describes the abnormality (if any)?

Options (single best answer):

- A) Pancreatic cancer
- B) Splenic lesion
- C) Renal tumor (renal neoplasm)
- D) No abnormality detected
- E) Colonic tumor (colorectal neoplasm)

Ground Truth: E



Image-2

Turn 1:

<think>Initial review of the abdominal CT at a lower abdominal–pelvic level. The pancreas is not visible at this slice, so option A) Pancreatic cancer is not applicable. The spleen is typically located in the left upper abdomen and is not included in this slice, so option B) Splenic lesion is excluded. Portions of both kidneys are visible, but there is no evidence of mass or focal lesion, so option C) Renal tumor is unlikely. In the left lower abdomen (right side of the image), bowel loops are visible with enhancing colonic wall. It is uncertain whether this represents normal enhancement or a possible abnormality. To clarify, I will apply a zoom-in tool to the region of interest.

</think>

```
<tool_call>{"name": "image_zoom_in_tool",
"arguments": {"bbox_2d": [765, 331, 850, 458]}}
</tool_call>
```

<obs>image-2</obs>

Turn 2:

<think>The zoomed-in region shows a segment of colonic wall with focal thickening and luminal irregularity. These findings are consistent with an abnormal lesion rather than normal bowel wall. This supports option E) Colonic tumor (colorectal neoplasm) and rules out option D) No abnormality detected.

</think>

<answer>E</answer>

Figure 9: **Visual Confirmation example.**

1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511

Original Image-1:



Question:

Which of the following descriptors apply to the hepatic lesion in this image?

Options (select all that apply):

- A) Round-like morphology
- B) Hypodense lesion
- C) Heterogeneous internal intensity
- D) Ill-defined margins
- E) Well-defined margins

Ground Truth: ABE

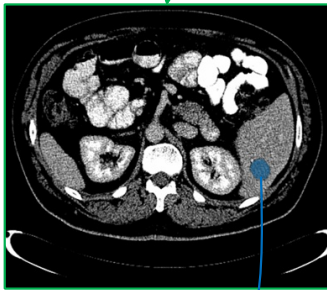


Image-2

Turn 1:

`<think>`Initial inspection of the axial contrast-enhanced CT reveals a **round-like lesion** within the hepatic parenchyma. The lesion demonstrates **attenuation lower than the surrounding liver tissue**, indicating a **hypodense** character. Internally, the density appears relatively **heterogeneous**, and the margins are **well defined**, allowing clear demarcation from adjacent liver parenchyma. Based on the provided options, the preliminary answer would be A) Round-like morphology, B) Hypodense lesion, **C) Heterogeneous internal intensity**, and E) Well-defined margins. To confirm this assessment, I will apply BiomedParse with the description 'liver lesion' to obtain a precise segmentation.
`</think>`

```
<tool_call>{"name": "BiomedParse", "arguments": {"description": "liver lesion", "Image": "Image-1"}}
</tool_call>
```

`<obs>image-2</obs>`

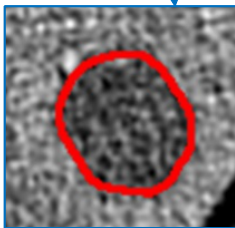


Image-3

Turn 2:

`<think>`The BiomedParse output provides a segmentation mask of the suspected hepatic lesion. To further refine my analysis, I will zoom in on the segmented region to evaluate its internal characteristics and margins more carefully.
`</think>`

```
<tool_call>{"name": "image_zoom_in_tool", "arguments": {"mask": "image-2"}}
</tool_call>
```

`<obs>image-3</obs>`

Turn 3:

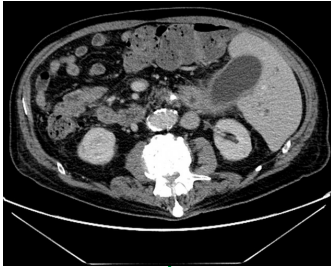
`<think>`The zoomed-in image of the hepatic lesion confirms a **well-circumscribed, round hypodense mass with homogeneous internal attenuation**. My earlier impression of **heterogeneous intensity was incorrect upon closer inspection**. Therefore, the corrected final answer should include A) Round-like morphology, B) Hypodense lesion, and E) Well-defined margins.
`</think>`

`<answer>ABE</answer>`

Figure 10: **Hallucination Mitigation example.**

1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565

Original Image-1:



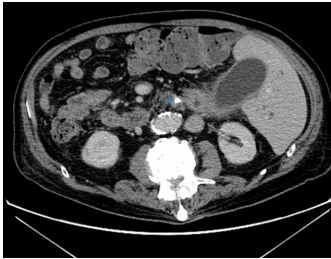
Question:

On the provided contrast-enhanced abdominal CT, is there evidence of pancreatic carcinoma? If present, delineate the tumor margins on the relevant slice(s).

Options:

- A) Yes
- B) No

Ground Truth: A) Yes



Ground Truth ✓

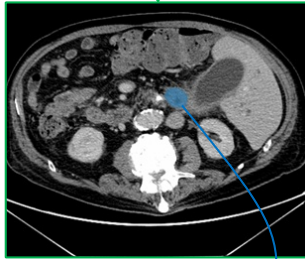


Image-2 ✗

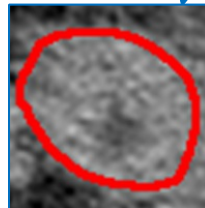


Image-3

Turn 2:
 <think>The BiomedParse output highlights a region. To evaluate its validity, I will zoom in on the segmented area to assess whether it corresponds to a pancreatic mass.
 </think>
 <tool_call>{"name": "image_zoom_in_tool", "arguments": {"mask": "image-2"}}
 </tool_call>
 <obs>image-3</obs>

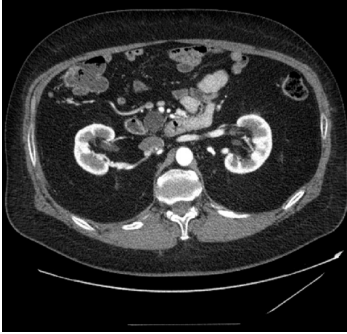
Turn 1:
 <think>Initial review of the contrast-enhanced abdominal CT. The user's query is to determine whether a pancreatic carcinoma is present. On this slice, the pancreas is partially visible but there is no obvious mass, focal hypoattenuation, or infiltrative lesion. However, to confirm this impression, I will apply BiomedParse with the description 'pancreatic tumor'.
 </think>
 <tool_call>{"name": "BiomedParse", "arguments": {"description": "pancreatic tumor", "Image": "image-1"}}
 </tool_call>
 <obs>image-2</obs>

Turn 3:
 <think>The zoomed-in view shows that the segmented area does not correspond to a malignant pancreatic lesion. There is no mass effect, no irregular contour, and no hypoenhancing lesion consistent with pancreatic carcinoma. Therefore, the segmentation result is considered incorrect. Based on direct image review, this CT slice does not demonstrate pancreatic carcinoma.
 </think>
 <answer>B</answer> ✗

Figure 11: Failure example (hard case).

1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619

Original Image-1:



Question:

On the provided abdominal CT, is there evidence of a pancreatic cyst (pancreatic cystic lesion)? If present, delineate the lesion.

Options:

- A) Yes
- B) No

Ground Truth: A) Yes

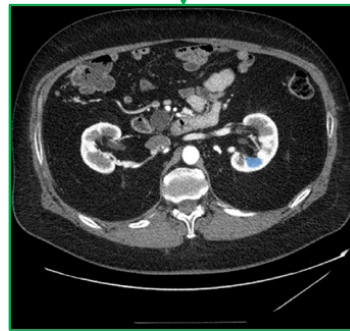


Image-2

Turn 1:

<think>Initial review of the contrast-enhanced abdominal CT. The task is to determine whether a pancreatic cystic lesion is present. To begin, I will apply an automated parsing tool focused on detecting pancreatic cysts.
</think>

```
<tool_call>{"name": "BiomedParse",
"arguments": {"description": "pancreatic cyst", "Image": "Image-1"}}
</tool_call>
```

<obs>image-2</obs>

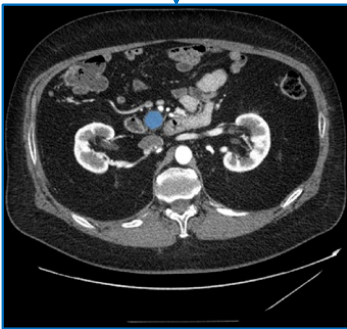


Image-3

Turn 2:

<think>The BiomedParse output highlights a region. However, the segmentation appears to correspond to renal parenchyma rather than the pancreas. This suggests a misclassification. Therefore, I must reassess the original CT and localize the pancreas directly. Re-examining the original image-1, I observe a well-circumscribed, round, hypodense lesion in the region of the pancreatic head. This morphology is more consistent with a pancreatic cyst. To obtain a precise delineation, I will now apply the SAM2 model with a bounding box over the suspected lesion.
</think>

```
<tool_call>{"name": "SAM2", "arguments": {"bbox_2d":
[554, 516, 612, 564], "Image": "Image-1"}}
</tool_call>
```

<obs>Image-3</obs>

Turn 3:

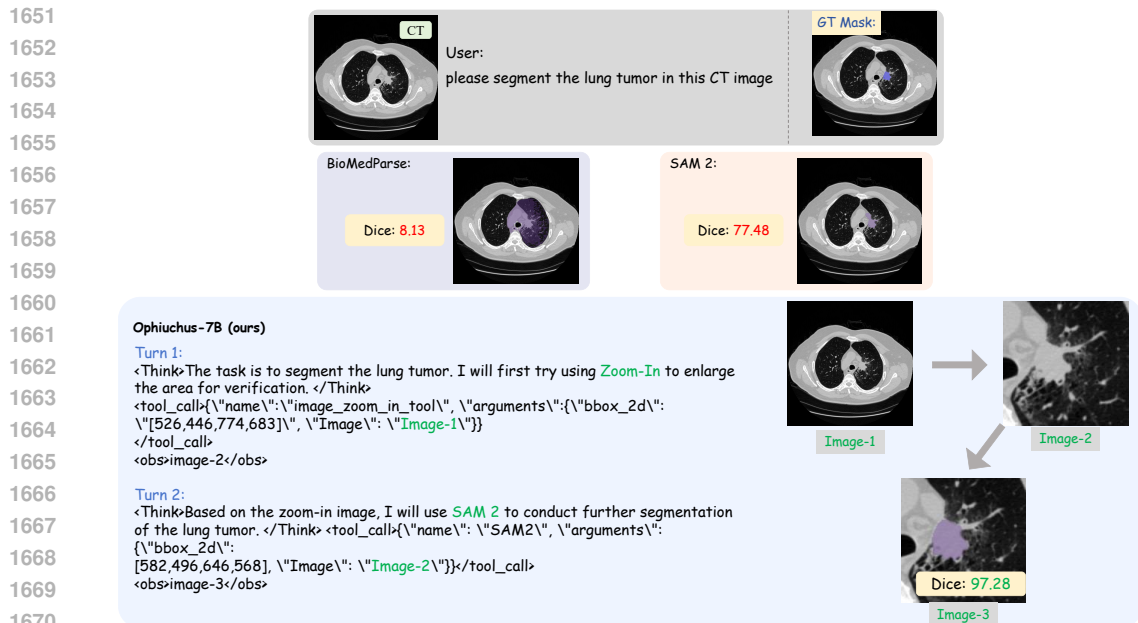
<think>The SAM2 segmentation outlines a round hypodense lesion in the pancreatic head, which confirms a cystic lesion with homogeneous internal attenuation and well-defined margins. This is consistent with a pancreatic cyst. Therefore, the correct answer is Yes, with the segmentation mask provided as mask-3.
</think>

<answer>A, the segmentation mask is Image-3</answer>

Figure 12: Self-Reflection example.

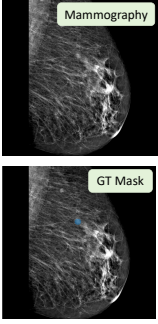


1645 **Figure 13: Comparative case analyses.** It can be observed that both general-domain and medical-
 1646 specific agents augmented with tools frequently produce incorrect interpretations and answers when
 1647 analyzing medical images, often accompanied by inaccurate localizations or segmentation masks.
 1648 In contrast, Ophiuchus not only produces correct textual analyses but also achieves accurate spatial
 1649 understanding, generating high-quality segmentation masks.



1671 **Figure 14: Cases illustrating how Ophiuchus combines different tools to further enhance seg-**
 1672 **mentation performance.** Ophiuchus can compose multiple tools—by using the zoom-in tool in
 1673 conjunction with SAM 2—to achieve better segmentation results than relying on SAM 2 alone.

1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727



Question:
In the provided mediolateral oblique (MLO) mammographic image of the breast, is there evidence of a lesion consistent with a localized mass or nodule?

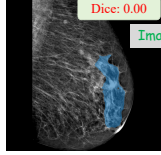
Options:
A. Yes - a small, well-circumscribed nodular lesion is present in the upper breast region.
B. No - the image shows only normal dense fibroglandular tissue without any focal abnormality.
C. No - the bright region near the nipple represents typical post-processing artifact rather than a lesion.
D. No - the round densities seen in the breast are all consistent with vascular calcifications.

Correct Answer:
A. Yes - a small, well-circumscribed nodular lesion is present.

Ophiuchus-7B (ours)

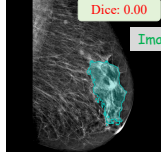
Turn 1
 <think> This is a left mediolateral oblique (MLO) mammographic view showing predominantly heterogeneously dense fibroglandular tissue. The current task is to evaluate the image for potential lesions. To begin, I will apply BioMedParse to identify any suspicious breast abnormalities. </think>
 <tool_call>{"name": "BioMedParse", "arguments": {"description": "breast lesion", "Image": "Image-1V"}} </tool_call> <obs>image-2</obs>

BioMedParse yielded an incorrect segmentation outcome, where the segmented region belonged to an inappropriate class. 🤖

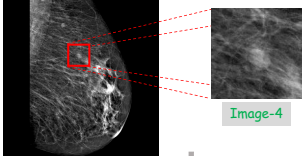


Turn 2
 <think> BioMedParse returned a segmentation result, but it appears to reflect a misinterpretation of the breast parenchymal density. The predicted area corresponds primarily to dense fibroglandular tissue. Such dense tissue often forms vertically oriented streaks or sheet-like patterns, which the model seems to be misclassifying as a "lesion". I will re-segment the input image using SAM 2 to verify whether there is any obvious mass, suspicious microcalcification, or architectural distortion in the retroareolar region or the inferior portion of the breast. </think>
 <tool_call>{"name": "SAM 2", "arguments": {"bbox_2d": [393, 428, 721, 897], "Image": "Image-1V"}} </tool_call> <obs>image-3</obs>

SAM 2 also yielded an incorrect segmentation outcome, highlights a large, irregular region corresponding to normal dense fibroglandular tissue. 🤖



Turn 3
 <think> SAM 2 returned a segmentation result; however, the predicted region is clearly not a breast lesion. The large connected predicted region suggests the model is trying to segment general dense tissue structures rather than lesions. This failure is likely due to confusion between dense breast parenchyma and true lesions. This indicates that the retroareolar region and lower breast contain dense fibroglandular strands without a discrete mass, suspicious microcalcifications, or overt architectural distortion. Therefore, the answer is likely B. However, before reaching a conclusion, I still need to further examine the upper central to upper outer portion of the breast. </think>
 <tool_call>{"name": "image_zoom_in_tool", "arguments": {"bbox_2d": [393, 324, 519, 442], "Image": "Image-1V"}} </tool_call>
 <obs>image-4</obs>



Turn 4
 <think> Upon further inspection after zooming in, a small, round-to-oval, well-circumscribed nodular density can be identified in the upper central to upper outer breast. Therefore, the correct answer should be A. </think>
 <answer>A</answer>

Ophiuchus is able to backtrack and examine its entire intermediate reasoning process. Even when all segmentation tools fail to produce correct results, the model is still able to answer the question correctly by leveraging all gathered visual cues to revise its own decision. 😊

Figure 15: Cases illustrating that even when all segmentation tools fail, Ophiuchus can still answer the question correctly by leveraging all gathered visual cues to revise its own decision.

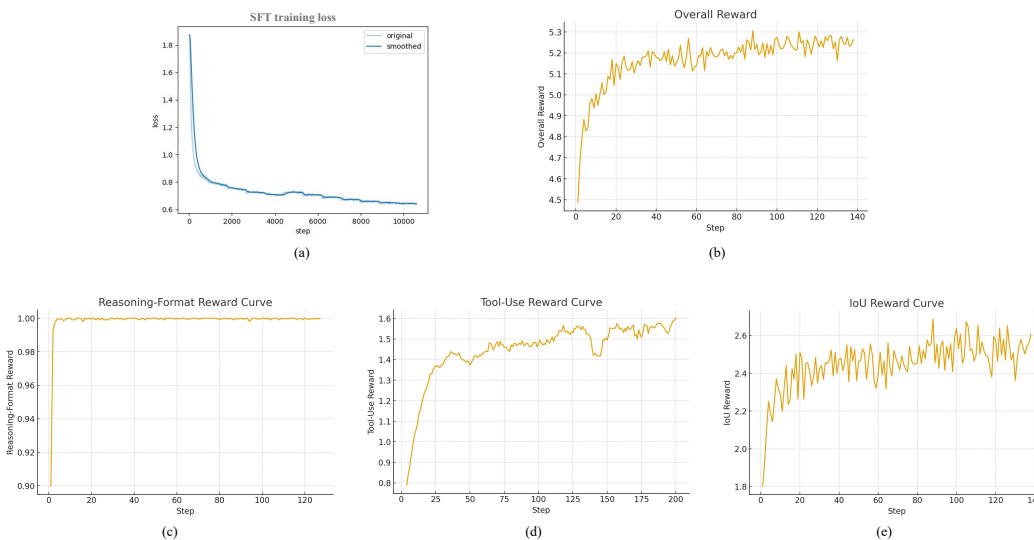


Figure 16: Training curves of Ophiuchus. (a) Training loss during the SFT stage; (b) overall reward curve during RL training; (c) reasoning-format reward curve; (d) strategic tool-use reward curve; (e) IoU-based component of the final-answer reward.

1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781

You are a medical expert specializing in interpreting medical images. Your task is to generate a clinically relevant visual multiple-choice question-and-answer set based on the provided inputs. The user will supply the following:

- Medical image: 1024×1024 resolution
- Target region coordinates: in the format `<"bbox": [x1, y1, x2, y2]>`, where (x1, y1) is the top-left corner and (x2, y2) is the bottom-right corner
- Binary ground-truth mask: 1024×1024 resolution, indicating the target region
- Textual descriptions: medical descriptions of the target segmentation region

Your task:

1. Create one multiple-choice question that is directly answerable by analyzing the given image and the specified target region.
2. Provide four or five options (A–D or A–E), using clear and professional medical terminology.
3. Indicate the correct answer explicitly.
4. Return the result strictly in the following JSON format:

```
[{
  "Question": "xxx",
  "Options": {
    "A": "xxx",
    "B": "xxx",
    "C": "xxx",
    "D": "xxx"
  },
  "Answer": "A"
}]
```

Notes:

- The question must be independent and self-contained; avoid referential terms such as “this case” or “the image above”. Each question should stand alone.
- Only use the information provided (image, bbox, mask, and descriptions). Do not hallucinate or introduce extra details.
- Use precise, standardized clinical language to maintain professional accuracy and clarity.
- The question should be strongly correlated with the target segmentation region and clinically meaningful.

Below are some examples of questions and answers for other images:

```
{seed_example_1}
{seed_example_2}
```

Now, use the following image, bbox, mask, and the mask descriptions to generate the question:

Figure 17: **Prompt for VQA construction.** Template for generating clinically grounded question-answer pairs under strict format control.

1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835

You are a medical imaging question–answer (VQA) validation expert. Validate a multiple-choice QA item using the provided image, target region (bbox + mask), and mask descriptions.

Inputs

1. Image: {image}
2. Image Resolution: {resolution} (expected: 1024×1024)
3. Target RegionBBBox:
 - {bbox} in the format [{"bbox": [x1, y1, x2, y2]}] (top-left = (x1,y1), bottom-right = (x2,y2))
 - Binary Mask: {mask} (1024×1024)
4. Mask Descriptions: {mask_descriptions} (textual descriptors of the segmented region)
5. Generated QA JSON: {qa_json}

```

Expected schema:
[[
  {"Question": "string",
   "Options": { "A": "string", "B": "string", "C": "string", "D": "string", "E": "string (optional)" },
   "Answer": "A/B/C/D/E"
}]

```

Validation Criteria

Evaluate only using the provided inputs. Do not introduce external knowledge or patient context that isn't present.

1. Schema & Format
 - JSON is valid and parseable.
 - Keys Question, Options, Answer exist.
 - Options contains 4 or 5 choices labeled A–D (and optionally E) only.
 - Answer is exactly a single uppercase letter among the provided options (A/B/C/D/E).
2. Independence & Wording
 - The question is self-contained; avoids referential terms.
 - Professional, precise clinical terminology; no vague or colloquial phrasing.
3. Grounding to Target Region
 - The question is directly answerable from the image with attention to the segmented target region (use the bbox and mask to focus reasoning).
 - High correlation with the target region: content should pertain to structures/findings inside or tightly adjacent to the masked area.
 - Uses only information derivable from the image + mask + mask descriptions; no hallucinated modality, demographics, or history unless explicitly given.
4. Options Quality
 - Options are mutually exclusive, non-overlapping, and plausible; one best answer.
 - No trivial cues (e.g., “All of the above,” “None of the above”), and no duplicate or near-duplicate options.
5. Answer Correctness
 - Determine the ground-truth choice by analyzing the image with the bbox/mask focus and the mask descriptions.
 - Verify the provided Answer matches the correct option.

Output (STRICT)

Return exactly one of the following:

- If JSON/schema/options/answer format is invalid (including wrong key names, missing fields, invalid labels, or Answer not a single A/B/C/D/E), respond with: Format error: <brief reason>
- If the format is valid and the specified Answer is correct, respond only with: Correct.
- If the format is valid but the specified Answer is wrong, respond with: Incorrect: <brief explanation grounded in bbox/mask/descriptions>
- Keep the explanation concise, e.g., one sentence referencing the segmented region and salient imaging cues.

Additional Guidance for Your Reasoning (do not output this text)

- Use the bbox and mask to prioritize features within the target region; if the question ignores the region or can't be answered from it, treat as a format error (“question not grounded to target region”).
- Prefer standardized terms (e.g., “hypodense lesion,” “spiculated nodule,” “vascular calcification”) consistent with the mask descriptions.
- If the resolution is not 1024×1024, you may still judge correctness but do not infer missing details from the discrepancy; rely on what's given.

Figure 18: **Prompt for VQA Verification.** Template for assessing the quality of each generated VQA item.

1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889

You are a medical expert specializing in interpreting medical images. The user will provide you with the following inputs:

- A medical image (with metadata such as modality and view) and resolution
- An ROI defined by a bounding box (bbox)
- A segmentation mask and a textual description of the mask/target
- A multiple-choice question related to the image
- A list of answer options
- The ground-truth answer
- Precomputed tool observation blocks that will appear in context after each tool call, formatted as `<obs> ... </obs>`

You also have access to the following tools (you must call them using the specified format):

- `image_zoom_in_tool` — zooms into a region
 - Call format inside `<tool_call>`:
`{"name":"image_zoom_in_tool","arguments":{"bbox_2d":[x1,y1,x2,y2]} }`
 (Optionally, you may pass a mask field instead of `bbox_2d` to zoom by mask.)
- `SAM2` — segments a target given a bbox
 - Call format: `{"name":"SAM2","arguments":{"bbox_2d":[x1,y1,x2,y2]} }`
- `BioMedParse` — segments a target given a textual description
 - Call format: `{"name":"BioMedParse","arguments":{"description":"<object/type description>"}}`

Your task is to simulate a step-by-step diagnostic reasoning process that uses these tools and their provided `<obs>` results to reach the answer. You must assume you do not know the ground-truth answer in advance; do not reference it during the reasoning. After the reasoning completes, reveal the correct answer by copying the ground-truth answer exactly. Please do not give empty output.

- Carefully analyze the image description/metadata, ROI (bbox), mask + mask description, and the question.
- Plan which tool to call, why you need it, and what arguments to pass.
- Every tool call must be:
 - preceded by a `<think>` step that justifies the call and arguments, and
 - immediately followed by the corresponding `<obs>` block that is already provided in the context. Do not fabricate or alter `<obs>` contents; only use the given ones.
- Integrate each `<obs>` into the next `<think>` step; refine or backtrack if needed.
- Use concise, precise, and clinically sound language grounded only in the given information.
- End with an `<answer>` block.

Output in JSON Format:

```
{
  "trace": "<think>...</think>\n<tool_call>{...}</tool_call>\n<obs>...</obs>\n...<think>...</think>\nAction:
Answer\n<answer>...</answer>",
  "answer": "<groundtruth answer>"
}
```

Context Placeholders to be provided by the user:

- Image/Resolution: `{image_metadata} / {resolution}`
- BBox: `{bbox}`
- Mask: `{mask}`
- Mask Description: `{mask_descriptions}`
- Question: `{question}`
- Options: `{options}`
- Groundtruth answer: `{answer}`
- Tool Observation Blocks (precomputed, to be consumed after each tool call): `{obs_blocks}`

Figure 19: **Prompt for Reasoning Trajectory Generation.** Template for constructing reliable agentic reasoning trajectories.

1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943

You are a medical expert specializing in interpreting medical images. The user will provide:

- The full generated reasoning trace using <think>, <tool_call>, <obs>, and <answer>
- The original image metadata and resolution
- The ROI bbox, segmentation mask, and mask description
- The multiple-choice question, options, and the ground-truth answer

Your task is to strictly evaluate the quality of the reasoning trace and its tool use. Check format, tool correctness, integration of observations, clinical soundness, and final answer correctness. Base your evaluation only on the provided trace and context. Do not assume extra clinical information.

Evaluation rules:

- **Format:** The trace must interleave <think> → <tool_call> → <obs> steps (repeated as needed) and conclude with Action: Answer and an <answer> block.
- **Tool usage:** Only the allowed tools (image_zoom_in_tool, SAM2, BioMedParse) may be used with valid JSON arguments.
- **Tool choice rationale:** Each <tool_call> must be justified by the immediately preceding <think> (why this tool, why these arguments).
- **Observation integration:** Each <obs> must be meaningfully used in the subsequent <think>; no ignoring or contradicting tool outputs.
- **Clinical correctness:** Reasoning must be consistent with the image/ROI/mask and mask description; no hallucinated findings or unsupported claims.
- **Final answer:** The content of <answer> must match the provided ground-truth answer exactly.

Output in JSON Format:

```
{
  "format_check": "pass/fail",
  "tool_usage_check": "pass/fail",
  "tool_choice_rationale_check": "pass/fail",
  "obs_integration_check": "pass/fail",
  "clinical_correctness": "pass/fail",
  "final_answer_check": "correct/incorrect",
  "explanation": "Brief reason(s) for any fail/incorrect; empty if all pass."
}
```

Notes:

- **Focus on the segmentation mask and bounding box details:** If the problem provided a mask or ROI (region of interest), ensure the reasoning trace made use of it. Not using an important provided region could be a flaw (either in tool usage or obs integration).
- **No credit for invented observations:** If the reasoning trace mentions observations that were not actually present in the <obs> outputs, that is a mistake (either a hallucination in the reasoning or misuse of a tool). The evaluator should catch this under clinical correctness or tool usage (since the model claimed a tool result that wasn't real).
- **Penalize redundant or unmotivated tool calls:** If the chain-of-thought uses a tool when it's not necessary (for example, calling a zoom tool on a region that was already clearly described, or repeating the same tool without new justification), that should be noted (likely affecting the tool_choice_rationale_check or tool_usage_check).
- The final JSON should be complete and well-formed. Even if multiple checks fail, the output must list each field. Only use the specified keys and values (no additional keys, and string values exactly as "pass", "fail", "correct", "incorrect").

Figure 20: **Prompt for Reasoning Trajectory Verification.** Template for validating the correctness and reliability of generated reasoning trajectories.

