

MEDVISTAGYM: A Scalable Training Environment for *Thinking with Medical Images* via Tool-Integrated Reinforcement Learning

Anonymous ACL submission

Abstract

Vision language models (VLMs) achieve strong performance on general image understanding but struggle to *think with medical images*, especially when performing multi-step reasoning through iterative visual interaction. Medical VLMs often rely on static visual embeddings and single-pass inference, preventing models from re-examining, verifying, or refining visual evidence during reasoning. While tool-integrated reasoning offers a promising path forward, open-source VLMs lack the training infrastructure to learn effective tool selection, invocation, and coordination in multimodal medical reasoning. We introduce MEDVISTAGYM, a scalable and interactive training environment that incentivizes tool-integrated visual reasoning for medical image analysis. MEDVISTAGYM equips VLMs to determine when and which tools to invoke, localize task-relevant image regions, and integrate single or multiple sub-image evidence into interleaved multimodal reasoning within a unified, executable interface for agentic training. Using MEDVISTAGYM, we train MEDVISTAR1 to interleave tool use with agentic reasoning through trajectory sampling and end-to-end reinforcement learning. Across six medical VQA benchmarks, MEDVISTA-R1-8B exceeds comparably sized tool-augmented baselines by 19.10 to 24.21%, demonstrating that structured agentic training—not tool access alone—unlocks effective tool-integrated reasoning for medical image analysis.

1 Introduction

Vision language models (VLMs) have achieved remarkable progress on medical image understanding, demonstrating strong performance across visual question answering (VQA) (Chen et al., 2024b), disease diagnosis (Liu and Song, 2025), diagnostic report generation (Goswami et al., 2025; Xia et al., 2025), and agentic medical visual analysis (Guo et al., 2025b). Much of this

progress can be attributed to text-based reasoning paradigms such as Chain-of-Thought (CoT) (Wei et al., 2022) and reinforcement learning (RL) (Guo et al., 2025a), which decompose complex problems into intermediate reasoning steps and improve reasoning quality through outcome-based optimization (Liu et al., 2025; Huang et al., 2025; Zhang et al., 2025). Together, these advances have moved medical VLMs beyond direct prediction toward more structured, step-by-step clinical reasoning.

Despite these advances, current medical VLMs remain limited in interactive visual reasoning, i.e., the ability to iteratively acquire, verify, and refine visual evidence during multi-step inference. Existing approaches largely rely on static visual embeddings and shallow cross-modal alignment, causing models to attend to irrelevant anatomy while overlooking fine-grained diagnostic cues such as blurred lesion boundaries, low-contrast abnormalities, and subtle tissue textures (Zheng et al., 2025; Fan et al., 2025). These limitations stem not from model capacity but from static, single-pass reasoning frameworks that lack mechanisms for continuous image interaction. As perception and decision-making occur in one forward pass, models cannot re-observe, verify, or refine visual evidence during reasoning, preventing dynamic visual exploration. Consequently, existing approaches fall short of true “*thinking-with-images*” (Su et al., 2025), where reasoning is tightly coupled with iterative visual perception in complex clinical scenarios.

To bridge this gap, recent work has explored *tool-integrated reasoning* (TIR) (Guo et al., 2025c; Zheng et al., 2025; Lu et al., 2025; Xu et al., 2025), which augments VLMs with external tools to support fine-grained visual reasoning. However, current open-source VLMs still struggle to leverage these tools effectively for medical reasoning tasks, largely due to the absence of a training environment that allows agents to learn dynamic tool selection and coordinated multi-step interaction through ex-

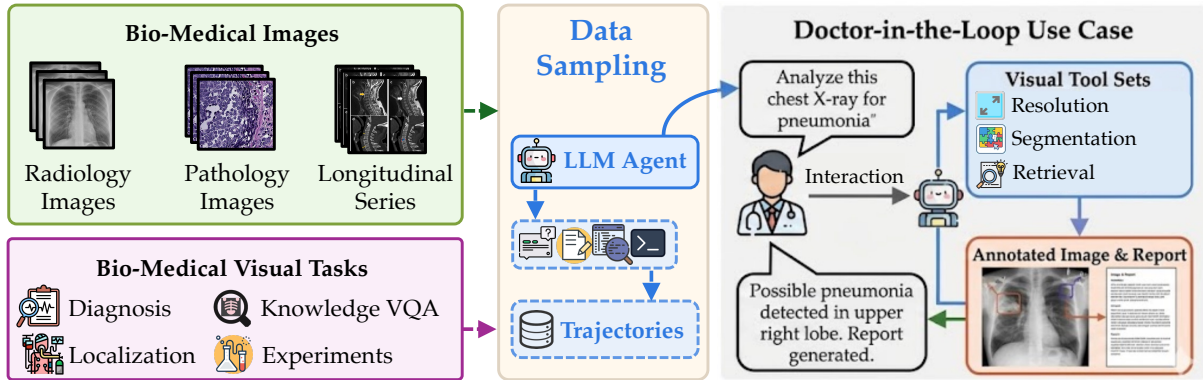


Figure 1: Overview of MEDVISTAGYM, which contains a comprehensive suite of reasoning-intensive medical image analysis tasks and tools in an interactive execution LLM environment, scaling visual-centric tool-integrated agentic reinforcement learning for VLM agents.

085 perience. Although some studies explore tool use
 086 in specific medical settings (Li et al., 2024a; Fathi
 087 et al., 2025), systematic simulation and training of
 088 tool-integrated thinking for medical visual reason-
 089 ing remain largely unexplored.

090 Motivated by these challenges, we introduce
 091 MEDVISTAGYM (**M**edical **V**isual-centric **T**ool-
 092 integrated **A**gentic training environment), a scal-
 093 able and interactive training environment designed
 094 to operationalize tool-integrated thinking for med-
 095 ical visual reasoning. MEDVISTAGYM encapsu-
 096 lates diverse visual tool operations and provides
 097 structured textual feedback, enabling closed-loop
 098 interaction between reasoning and perception (Fig-
 099 ure 1). It supports (i) multimodal medical reason-
 100 ing across *six* public datasets, (ii) a unified
 101 and extensible interface to *fifteen* visual and med-
 102 ical knowledge tools, and (iii) scalable infras-
 103 tructure for efficient agent training and evalua-
 104 tion. Building upon MEDVISTAGYM, we develop
 105 MEDVISTA-R1, a VLM-based agent for robust
 106 tool-augmented medical image reasoning. Across
 107 three in-domain and three out-of-domain med-
 108 ical VQA benchmarks, MEDVISTA-R1-8B deliv-
 109 ers consistent gains over comparably sized open-
 110 source baselines under both tool-enabled and tool-
 111 free settings, validating MEDVISTAGYM as a scal-
 112 able environment for training agentic VLMs.

113 2 MEDVISTAGYM

114 2.1 Problem Formulation

115 We formulate tool-integrated medical image anal-
 116 ysis as a Partially Observable Markov Decision
 117 Process (POMDP). Given a medical question with
 118 associated image(s), an agent interacts with an exe-
 119 cutable environment under instruction space \mathcal{I} and

120 receives partial observations \mathcal{O} . Here \mathcal{S} denotes
 121 latent environment states, \mathcal{A} the set of executable
 122 tool actions, and $\mathcal{P} : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$ the deter-
 123 ministic state transition function. At each time step,
 124 the agent generates an explicit reasoning step g_t
 125 to guide decision-making, then selects a tool ac-
 126 tion $a_t \in \mathcal{A}$. The environment executes the ac-
 127 tion and returns a new observation as external evi-
 128 dence. This process yields an interaction trajectory
 129 $U = (g_0, a_0, \dots, g_T, \hat{y})$, where \hat{y} is the final
 130 answer after T interaction steps.

131 2.2 Environment Design

132 Following the formulation in §2.1, we instantiate
 133 the agent-environment interaction as an executable
 134 system. MEDVISTAGYM is a scalable environment
 135 for multi-turn, tool-integrated medical visual
 136 reasoning. It defines the agent’s operating world, i.e.,
 137 tasks, tools, interaction interfaces, and execution
 138 infrastructure, independent of any specific training
 139 algorithm. A key property of MEDVISTAGYM is
 140 its support for *online and adaptive tool invocation*:
 141 each tool call at time t returns a real observation
 142 that conditions the agent’s next action at time $t + 1$.
 143 As a result, tool execution emerges dynamically
 144 from agent-environment interaction, rather than be-
 145 ing scripted in advance.

146 2.3 Tasks and Tool Sets

147 **Medical VQA Tasks.** MEDVISTAGYM com-
 148 prises a set of verifiable medical VQA tasks that
 149 demand grounded, multi-step reasoning over visual
 150 inputs and intermediate evidence. These tasks span
 151 diverse diagnostic scenarios, including clinical per-
 152 ception, lesion-level evidence localization, subtle
 153 abnormality detection, and diagnosis-oriented ev-

idence aggregation, where generating reliable answers requires calling external tool support. The training data in MEDVISTAGYM is organized along two complementary axes. (1) *Radiology VQA*, covering cross-sectional and projection imaging, includes **VQA-RAD** (Lau et al., 2018), which focuses on anatomy and finding recognition in X-ray, CT, and MRI images, and **SLAKE** (Liu et al., 2021), a knowledge-aware dataset with clinically grounded questions over diverse radiology images. (2) *Pathology VQA*, covering microscopy and histopathology, includes **PathVQA** (He et al., 2020), which emphasizes cellular morphology and tissue patterns. More details are provided in Appendix A.2.

Visual-Centric Tools Sets. To support grounded medical image verifiable analysis, MEDVISTAGYM exposes a standardized and extensible suite of executable tools that enable agents to offload perception, localization, segmentation, and knowledge retrieval. These tools return structured outputs that can be directly used as intermediate evidence during reasoning. The tools used in MEDVISTAGYM are organized into four complementary families: (1) *Resolution and Region Refinement*, which enable focused inspection of image regions (e.g., agent4k, zoom-in); (2) *Medical Localization and Segmentation* support the detection and delineation of anatomical or pathological regions (e.g., groundingdino, medsam2); (3) *Medical Visual Understanding and Parsing*, providing a structured interpretation of medical images (e.g., biomedclip, biomedparse); (4) *External Biomedical Knowledge Retrieval* enables access to curated medical knowledge sources (e.g., GoogleSearch, DrugBank, PubMed). See Appendix E for tool details.

2.4 Training Infrastructure

Agent-Env Interface. MEDVISTAGYM is an executable training environment built upon a Gym-style interaction protocol for medical image analysis. It defines a flexible API, an explicit and verifiable action space, and a structured observation space, enabling agents to perform multi-step reasoning through continuous interaction with the environment. Specifically, MEDVISTAGYM provides an executable interface to initialize interaction episodes and return observations while ensuring that all agent actions correspond to well-defined, executable, and verifiable tool invocations

that support evidence-grounded medical reasoning and decision-making. Please refer to Appendix H.1 for more details.

Scalable Execution Infrastructure. To support large-scale, multi-turn medical visual reasoning, MEDVISTAGYM defines a scalable execution *capability interface* that enables high-frequency tool invocation within the agent–environment interaction loop. All medical tools, including compute-intensive foundation models, are exposed through a unified execution interface as independently executable services.

Asynchronous Tool Execution. The execution interface supports asynchronous and batched tool invocation across interaction episodes, enabling efficient multi-turn rollouts while preserving reliable tool-mediated evidence acquisition.

Extensible Tool Infrastructure. MEDVISTAGYM provides a unified BaseTool abstraction that serves as a stable capability interface for plug-and-play integration of new medical perception and knowledge tools with minimal engineering overhead. Implementation details are provided in Appendix H.2.

3 MEDVISTA-R1

3.1 Training Data Collection

Data Collection and QA Generation. Our data collection follows three core principles: (1) coverage of diverse medical tasks and imaging modalities; (2) settings where tool usage leads to measurable performance gains; and (3) tasks that require multi-tool interaction. Consequently, we curate data from multiple established medical VQA benchmarks, including PathVQA, SLAKE, and VQA-RAD, which cover diverse medical imaging types and clinically grounded questions. To ensure that external tools are genuinely necessary for resolving vision-language queries, we employ GPT-5 (OpenAI, 2025) to perform tool-augmented reasoning and verify whether tool usage provides clear benefits for each instance. We further remove samples that cannot be reliably verified, such as those with incorrect or ambiguous answers.

Reasoning Trajectory Generation and Data Selection. We initialize the policy with behavioral cloning (BC) on synthesized tool-augmented medical reasoning trajectories that interleave thoughts

and tool calls, providing a stable prior over tool syntax, selection, ordering, and grounded reasoning. To build the supervision set, we use GPT-5 (OpenAI, 2025) to generate candidate tool-executing trajectories under the same task and tool constraints, and keep only those whose final answers match the ground truth (outcome-based filtering). We further use GPT-5 as an external trajectory judge: for each retained trajectory, it assigns an ordinal quality score on a 4-point scale based on overall reasoning-guided tool use. These scores are used to further filter and weight demonstrations, favoring disciplined and medically grounded tool orchestration. Please see Appendix A.1 for more details.

3.2 Training Framework

We train the agent in two stages to acquire tool-integrated medical reasoning capabilities in MEDVISTAGYM: (1) a cold-start SFT to establish basic multi-turn tool-invocation and visual-interaction capabilities, including self-verification for reassessing tool outputs and revising decisions; and (2) an agentic tool-based online RL stage, where the agent leverages environmental feedback to refine tool orchestration and evidence-grounded reasoning across tasks.

Stage I: Cold-Start Supervised Bootstrapping.

Direct prompting of vanilla VLMs falls short of the reliability required for multi-turn clinical reasoning and precise tool execution. We therefore initialize the agent’s tool-integrated reasoning capability via supervised fine-tuning on a curated cold-start dataset D_{cold} , which contains multi-turn reasoning trajectories interleaving explicit thoughts and executable tool calls. We train the model using behavioral cloning by minimizing the negative log-likelihood over all reasoning and tool-call tokens in the trajectory. The objective encourages the agent to generate syntactically valid tool invocations, select appropriate tools in the correct order, and maintain coherent reasoning across extended interaction horizons.

Stage II: Agentic Tool-Integrated Online RL.

Building on cold-start supervision, we train the agent via online reinforcement learning within an executable environment, enabling it to jointly improve multi-turn reasoning and tool orchestration through direct interaction. At each interaction step, the agent generates a reasoning segment followed by a tool invocation; the environment executes the action and returns structured observations that are

appended to the interaction context, supporting continued reasoning and the learning of multi-step, compositional tool-use behaviors beyond static supervision.

Rollout Formulation.

MEDVISTAGYM extends conventional text-only reinforcement learning by explicitly modeling tool invocations and the execution feedback they produce. Tool calls are executed as external functions, whose outputs are treated as environmental observations—rather than model-generated tokens—and appended to the interaction context. Given a user query Q and image I , a multi-turn reasoning trajectory up to step k is denoted as: $R_k=(r_1, t_1, o_1; r_2, t_2, o_2; \dots; r_k, t_k, o_k)$, where r_i represents the reasoning text at step i , t_i the corresponding tool invocation, and o_i the observation returned by the tool execution. At step $k + 1$, the policy jointly generates the next reasoning segment r_{k+1} and selects a tool invocation t_{k+1} conditioned on the image I , query Q , and prior trajectory R_k : $(r_{k+1}, t_{k+1}) \sim \pi_\theta(\cdot | I, Q, R_k)$. The rollout alternates between reasoning and tool execution until a final answer is produced or a predefined tool-call limit is reached. To prevent inefficient or cyclic behavior, rollouts are terminated early if a tool invocation repeats a previously executed action. The model follows a strict output format, `<think></think><tool_call></tool_call><think></think><answer></answer>`, to explicitly distinguish reasoning steps, tool calls, and the final answer. Tool calls are automatically parsed into executable functions using model-predicted parameters, and the resulting outputs are inserted into the `<obs>` field and appended to the ongoing trajectory. Observation tokens are treated as a whole and excluded from loss computation. Prompt details are provided in Appendix G.

Reward Design.

To provide dense feedback for diverse tool-use scenarios and multi-turn interactions, we design a rule-based reward function decomposed into fine-grained signals. The final reward consists of three components: a reasoning-format reward, a final-answer accuracy reward, and an answer-conditioned tool-use reward.

- *Format reward* R_{format} . the format reward evaluates the structural validity of a trajectory by checking whether the model output contains all required special tokens in the correct order.
- *Final-answer accuracy reward* R_{acc} . this reward evaluates the accuracy of the final prediction \hat{y} extracted from the `<answer>...</answer>` span. To

ensure a low-noise and format-aware training signal, accuracy is rewarded only when the output is well-formed, free of repetitive generations, and the predicted answer matches the ground-truth label:

$$R_{\text{acc}}(U) = \mathbb{I}\{\hat{y} = y\}$$

• *Answer-conditioned tool-use reward* R_{tool} . Beyond format and final-answer accuracy, we further design an answer-conditioned tool-use reward as a *conditional reward term*. This reward is assigned only to trajectories that both invoke at least one external tool and produce a correct final answer. By conditioning the tool-use reward on task success, we encourage the agent to learn when and how tool invocation genuinely contributes to solving the task, rather than invoking tools arbitrarily.

• *Final Reward*. The rollout-level reward is defined as the sum of the three components: $R(U) = R_{\text{format}}(U) + R_{\text{acc}}(U) + R_{\text{tool}}(U)$. This format-aware, outcome-conditioned reward provides positive feedback only to structurally valid trajectories that yield correct answers and use tools in a task-relevant way *conditioned on task success*, encouraging the policy to internalize the *think*→*tool_call*→*answer* protocol. Please refer to Appendix C for more details.

Optimization. Based on above defined rollout formulation and reward, we optimize the policy using Group Relative Policy Optimization (GRPO) (Guo et al., 2025a), which normalizes advantages across groups of sampled trajectories to stabilize policy updates and encourage relative preference among higher-quality reasoning paths. Detailed formulations are provided in Appendix D.

4 Experiments

4.1 Experiment Setups

Datasets. We evaluate on several representative public benchmarks: (1) in-distribution: PathVQA, SLAKE, VQA-RAD; (2) out-of-distribution: MMMU (H&M) (Yue et al., 2024), PMC-VQA (Zhang et al., 2024), and MicroVQA (Burgess et al., 2025). MicroVQA decomposes scientific visual reasoning into three core capabilities, including *Expert Visual Understanding* (V), *Hypothesis Generation* (H), and *Experiment Proposal* (E).

Evaluation Metrics. For evaluation metrics, we use *accuracy* for multiple-choice VQA benchmarks. We report per-category accuracy on MicroVQA to better disentangle model capabilities

across perception, scientific reasoning, and experimental planning, which are not distinguished in conventional medical VQA benchmarks.

Implementation Details. We implement MEDVISTA-R1 on InternVL3 (Zhu et al., 2025) with 2B and 8B variants, and train the model using VerI-Tool (Jiang et al., 2025) on a cluster of 10 NVIDIA A100 GPUs. During SFT, we train for 5 epochs with a learning rate of 1×10^{-5} and a total batch size of 256. For RL, we use a batch size of 256, sample 8 candidate reasoning trajectories per question with up to 6 tool calls, and adopt a constant learning rate of 1×10^{-6} with a maximum context length of 2.6K tokens. During inference, we deploy external tools (e.g., BiomedParse and 4kAgent) as FastAPI (Ramírez and Contributors, 2018) services to accelerate tool invocation.

Baselines. We compare MEDVISTA-R1 with three categories of baselines: (1) General-purpose VLMs with visual reasoning, including (i) API-based VLMs: GPT-5 (OpenAI, 2025), GPT-5-mini (OpenAI, 2025), GPT-o4-mini (OpenAI, 2024), Gemini-2.5-Pro (Google, 2025b), Gemini-2.5-Flash (Google, 2025a), Claude-4.5-Sonnet (Anthropic, 2025a), and Claude-4.5-Haiku (Anthropic, 2025b); (ii) Open-source VLMs: InternVL3, Qwen2.5-VL (Bai et al., 2025b), Qwen3-VL (Bai et al., 2025a), DeepEyes-7B (Zheng et al., 2025), Mini-o3-7B (Lai et al., 2025), and PixelReasoner (Wang et al., 2025a). (2) Medical-specific VLMs: LLaVA-Med (Li et al., 2023), HuatuoGPT-Vision (Chen et al., 2024a), Chiron-o1 (Sun et al., 2025), and Lingshu (Team et al., 2025). (3) Medical agent frameworks, including MMedAgent (Li et al., 2024a), VILA-M3 (Nath et al., 2025), and MMedAgent-RL. All proprietary baselines are evaluated *without tool access*.

4.2 Main Results

Table 1 shows the main results of MEDVISTA-R1 trained in MEDVISTAGYM against competitive baselines on three in-domain and three out-of-domain benchmarks. From the results we make the following key observations: **(i) MEDVISTA-R1 achieves strong performance over comparable baselines.** Specifically, MEDVISTA-R1-8B outperforms vanilla backbone models of similar size by 8.43%–13.58% without tools. When compared against the same backbones with tool access but without RL, MEDVISTA-R1-8B further achieves an additional 19.10%–24.21% improvement. **(ii)**

Models(↓)/Datasets(→)	In-Distribution				Out-of-Distribution						All	
	SLAKE	VQA-RAD	PathVQA	Avg.	PMC-VQA	MMMU (H&M)	MicroVQA			Avg.		Avg.
	Overall	Overall	Overall		Overall	Overall	Overall	V	H	E		
<i>Commercial VLMs (vanilla prompt)</i>												
GPT-5	70.28	67.53	64.4	67.40	62.00	77.86	60.80	62.00	61.20	58.30	66.89	67.15
GPT-5-mini	64.59	73.31	68.65	68.85	60.00	74.28	57.10	58.70	57.61	53.51	63.79	66.32
GPT-o4-mini	76.27	63.21	66.00	68.49	56.00	78.75	65.00	68.75	67.78	54.35	66.52	67.51
Claude-4.5-haiku	77.00	62.50	60.00	66.50	59.00	72.14	53.14	52.81	51.21	57.00	61.43	63.96
Claude-4.5-sonnet	75.00	63.39	63.00	67.13	59.00	76.73	56.33	55.60	57.91	54.83	64.02	65.58
Gemini-2.5-pro	85.00	71.43	71.00	75.81	62.00	71.29	60.90	61.20	62.60	57.40	64.73	70.27
Gemini-2.5-flash	72.00	62.50	36.08	56.86	35.00	46.43	47.00	50.00	44.50	46.50	42.81	49.84
<i>Base Size: 7-13B parameters</i>												
Qwen2.5vl-7B	42.11	64.14	62.40	56.22	49.00	46.43	33.00	34.38	31.11	34.78	42.81	49.51
LLaVA-Med-7B	61.97	56.60	59.00	59.19	27.00	37.71	16.51	12.20	17.60	21.70	27.07	41.13
Qwen3vl-8B	52.63	73.71	62.90	63.08	52.50	52.86	35.90	38.80	31.90	38.30	47.09	55.08
MEDVISTAGYM (Qwen3vl-8B)	27.00	51.89	56.40	45.10	49.50	42.86	36.28	40.62	35.61	32.63	42.88	43.98
InternVL3-8B	43.06	72.91	68.10	61.36	56.50	55.89	31.50	31.90	26.70	39.60	46.76	54.66
MEDVISTAGYM (InternVL3-8B)	28.50	31.70	46.00	35.40	43.50	48.57	35.00	39.06	32.20	34.80	42.36	38.88
GRPO w/o Tools	35.00	70.75	70.50	58.75	51.50	45.30	43.00	40.63	44.44	43.48	46.60	52.68
Direct GRPO w/o cold-start	30.00	72.64	68.00	56.88	49.00	43.80	41.00	43.75	38.89	41.30	44.61	50.74
Cold-start w/o Tools	35.00	70.75	72.00	59.25	51.00	55.00	42.50	43.75	43.33	39.10	49.50	54.38
Cold-start w/o Reasoning	61.00	66.04	66.00	64.35	56.00	52.14	39.00	39.10	37.80	41.30	49.05	56.70
MEDVISTA-R1	81.36	70.75	69.00	73.70	58.00	56.43	43.00	42.20	37.80	54.41	52.48	63.09
<i>Base Size: < 7B parameters</i>												
Internvl3-2B	42.25	41.04	41.80	41.71	44.50	43.57	30.48	30.00	28.57	34.86	39.52	40.61
MEDVISTAGYM (InternVL3-2B)	19.00	18.87	31.11	22.99	40.00	36.43	31.00	34.40	25.60	37.00	35.81	29.40
GRPO w/o Tools	29.00	58.49	63.00	50.16	48.50	37.50	37.00	34.38	37.78	39.13	41.00	45.58
Direct GRPO w/o cold-start	25.00	64.15	66.00	51.72	47.00	42.00	38.50	35.90	36.67	43.48	42.50	47.11
Cold-start w/o Tools	31.00	59.43	61.00	50.48	46.50	44.29	35.50	39.10	41.11	34.78	42.10	46.29
Cold-start w/o Reasoning	45.00	52.83	57.00	51.61	47.50	45.71	37.00	35.90	38.90	34.80	43.40	47.51
MEDVISTA-R1	60.99	58.49	55.00	58.16	49.50	47.43	39.50	37.50	37.78	45.65	45.48	51.82

Table 1: Main results (Acc.%) on three in-distribution and out-of-distribution medical VQA benchmarks. *w/o Tools* removes tool access during both training and inference; *w/o Reasoning* removes the RL reasoning stage. *w/o cold-start* removes the SFT stage. MEDVISTAGYM (*Backbone*) denotes the vanilla model operating within the MEDVISTAGYM environment with tool access enabled, without additional training.

RL is critical for boosting TIR in VLMs. Simply augmenting VLMs with tools without explicit reasoning supervision degrades performance, whereas RL yields substantial gains, indicating that RL unlocks effective tool use in medical visual reasoning. (iii) **MEDVISTA-R1 has a strong parameter efficiency.** MEDVISTA-R1-2B achieves competitive or even better performance with 8B baselines, while MEDVISTA-R1-8B performs comparably to baselines with some proprietary VLMs. The superiority of MEDVISTA-R1 demonstrates that MEDVISTAGYM provides a scalable training ground for tool-integrated agentic RL, enabling robust visual reasoning in open-source VLMs. Comparison with additional baselines is provided in Appendix B.

4.3 Ablation Studies

Effect of Tool-Integrated Reasoning To assess the impact of MEDVISTAGYM, we isolate tool use, reasoning, and reinforcement learning through controlled variants in Table 2. We observe that naive tool access is not inherently beneficial for either open-source or proprietary models. For InternVL3-8B, enabling tools without structured learning signals causes a substantial performance drop (54.66% → 38.88%). Similarly, directly enabling tool in-

Variant	Tools	Reason	RL	ID Avg.	OOD Avg.	Overall
GPT5	×	×	×	67.40	66.89	67.15
GPT5 (Direct Tool Access)	✓	×	×	65.81	62.36	64.08
MEDVISTAGYM (GPT5)	✓	×	×	76.33	66.91	71.62
InternVL3-8B	×	×	×	61.36	40.76	54.66
MEDVISTAGYM (InternVL3-8B)	✓	×	×	35.40	42.36	38.88
Cold-start w/o Tools	×	×	×	59.25	49.50	54.38
Cold-start w/o Reasoning	×	×	×	64.35	49.05	56.70
GRPO w/o Tools	×	×	✓	58.75	46.60	52.68
Direct GRPO w/o Cold-start	✓	✓	✓	56.88	44.61	50.74
MEDVISTA-R1	✓	✓	✓	73.70	52.48	63.09

Table 2: Effect of tool-integrated reasoning.

vocation for GPT-5 degrades performance relative to its vanilla counterpart, whereas deployment within MEDVISTAGYM improves results (67.15% → 71.62%). These findings show that tool integration is not plug-and-play: without disciplined interaction and learning signals, tools often act as distractors rather than facilitators of medical visual reasoning. Further ablations show that reasoning or RL alone is also suboptimal—Cold-start w/o Tools (54.38%), GRPO w/o Tools (52.68%), Cold-start w/o Reasoning (56.70%), and Direct GRPO w/o cold-start (50.74%). The full MEDVISTA-R1 achieves the best performance (63.09%), demonstrating that gains arise from coupling reasoning-guided tool invocation with online RL in MEDVISTAGYM.

Effect of Training Stages. Figure 2 (a) shows that SFT provides a crucial warm-up by align-

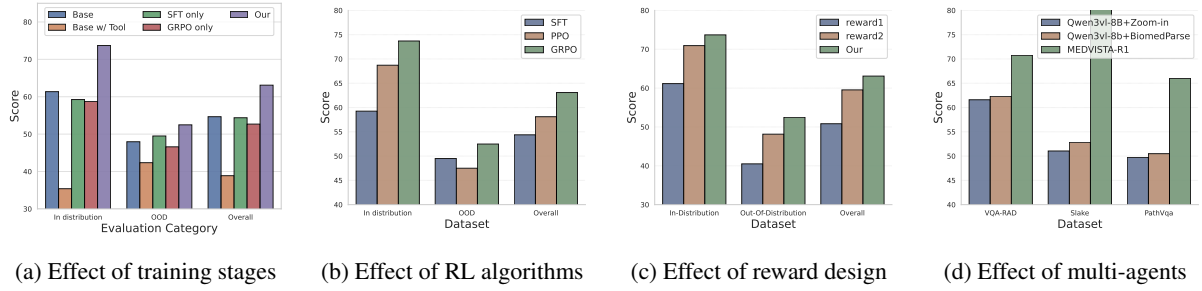


Figure 2: Ablation study on training configurations.

ing the model with tool-use formats and syntax (+2.32%), while RL builds on this foundation to deliver larger gains (+8.72%). Together, SFT establishes reliable tool-use priors, and RL enables deeper interleaved reasoning for solving complex tasks.

Effect of RL Algorithm. We evaluate PPO as an alternative to GRPO; Figure 2 (b) shows that GRPO yields more robust performance. By retaining all rollouts and using group-normalized advantages, GRPO provides lower-variance, difficulty-adaptive credit assignment, resulting in more stable gains.

Effect of Reward Design. Figure 2 (c) compares three reward designs: (1) reward-1, a base reward $R_{\text{sparse}} = R_{\text{format}} \cdot R_{\text{correct}}$ that evaluates only format compliance and answer correctness; (2) reward-2, an extra tool use reward added to the base reward; and (3) ours, a conditional tool use reward that is granted only when tool use leads to a correct answer. The results show that removing the tool reward leads to a substantial performance drop, highlighting its importance. Among all settings, the answer-conditioned tool-use reward achieves the highest accuracy. These findings suggest that rewarding tool use alone is insufficient; instead, aligning tool use rewards with successful outcomes is crucial for inducing intelligent and effective behavior in MEDVISTA-R1.

Exploration of Multiagent Workflow Variants. To evaluate whether performance gains arise solely from tool usage, we compare MEDVISTA-R1 with tool-based and multi-agent workflow baselines of comparable model sizes (Figure 2d). These baselines isolate the effect of tool invocation from reasoning-driven interaction. We observe that (1) *Qwen3-VL-8B + Zoom-in* applies a predefined zoom-in operation as a static preprocessing step, decoupled from the model’s reasoning; (2) *Qwen3-VL-8B + BioMedParse* uses BioMedParse to pro-

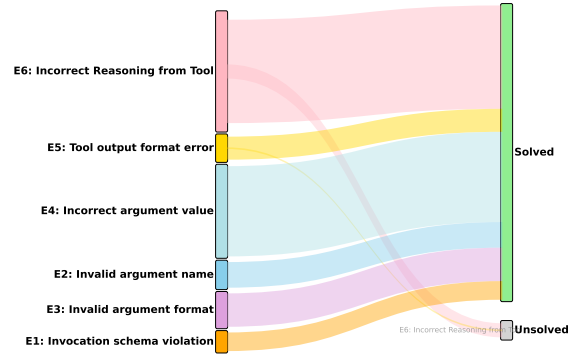


Figure 3: Error analysis.

vide segmented regions and cropped inputs, but follows a fixed, non-adaptive workflow. Across these variants, simply incorporating tools yields limited gains. In contrast, MEDVISTA-R1 dynamically coordinates tool invocation with reasoning, enabling selective and task-dependent use of visual evidence.

4.4 Error Analysis and Case Study

Error Analysis. To better understand the reason vanilla open-source VLMs struggle with to effectively invoke external tools (Table 2), we manually analyze 100 error cases per benchmark for GPT-5 and InternVL3-8B. We identify several tool-related failure modes, including mistimed or inappropriate tool invocation, incorrect tool selection, malformed arguments, and faulty reasoning over tool outputs (Table 7). These patterns suggest that open-source VLMs lack the interaction priors required for disciplined multi-step tool use, highlighting a gap between proprietary and open-source models in tool-augmented medical image analysis. We then re-evaluate the same samples using MEDVISTA-R1-8B (InternVL3). As shown in Figure 3, most tool-related and reasoning errors are substantially reduced, with only a small fraction remaining. This

improvement demonstrates that MEDVISTAGYM enhances both tool invocation discipline and reasoning over tool outputs. More details are provided in Appendix I.

Case Study. We summarize the capabilities acquired through agentic training in MEDVISTAGYM into four categories, illustrated with representative case studies: (1) schema- and argument-correct tool invocation (Figure 10); (2) effective post-tool reasoning grounded in evidence from tool outputs (Figure 11); (3) coordinated multi-tool usage for complex problem solving (Figure 12); and (4) robustness to imperfect tool outputs through multi-turn reasoning (Figure 13). Additionally, we analyze representative *hard cases* that reveal the boundaries of tool-integrated medical reasoning, where failures occur despite correct tool usage and coherent reasoning (Table 8). See Appendix J for details.

5 Related Work

Medical Visual Language Model Agents. Recent advances in medical VLMs extend foundational VLMs toward agent-like reasoning for medical visual question answering and radiology report generation. MedRAX (Fallahpour et al., 2025) introduces an agent for chest X-ray interpretation that integrates multiple perception modules for structured clinical reasoning and report generation. CXR-Agent (Sharma, 2024) proposes an agentic vision-language framework for generating structured radiology reports and diagnostic predictions, while CheXagent (Chen et al., 2024c) presents a unified medical VLM supporting chest X-ray understanding across reasoning tasks.

RL and Tool-Integrated Reasoning for VLMs. Augmenting VLMs for medical image analysis with external tools has emerged as an effective way to extend their capabilities beyond standalone reasoning by leveraging specialized functions or expert models. Existing approaches broadly fall into two categories. The first focuses on instruction-tuned tool invocation, where models are trained to call predefined medical tools for specific sub-tasks as exemplified by MMedAgent (Li et al., 2024a), which curates an instruction-tuning corpus over multiple medical tools; VILA-M3 (Nath et al., 2025), which triggers expert models for perception tasks; and AURA (Fathi et al., 2025), which integrates heterogeneous tools for medical VQA.

While effective, these methods rely on fixed perception behaviors and fragmented tool usage, limiting coherent reasoning and holistic planning. The second category explores multi-turn, multi-step reasoning frameworks for long-horizon decision-making and iterative refinement. For example, Med-VRAgent (Guo et al., 2025b) employs visual guidance and Monte Carlo Tree Search to enable ROI-grounded, multi-step medical reasoning, while MedAgent-Pro (Wang et al., 2025d) which employs hierarchical multi-agent workflows with retrieval and specialized tools for disease diagnosis.

Agentic Training Environments for Medical VLMs. Several environments have been proposed for agentic training and evaluation of VLMs. General-purpose frameworks such as AgentGym-RL (Xi et al., 2025), Collaborative Gym (Shao et al., 2024), and RAGEN (Wang et al., 2025c) support multi-step reasoning and decision-making, but operate primarily in text-only settings without multimodal perception. For VLMs, only a few environments, such as VAGEN (Wang et al., 2025b), enable multimodal agent training, but they are designed for non-medical domains. Although these environments support sequential tool composition, they are not tailored to agentic medical reasoning that requires multimodal grounding, adaptive tool selection, and iterative verification.

6 Conclusion

In this work, we introduce MEDVISTAGYM, a scalable agentic training environment for tool-integrated medical image analysis in VLMs, along with MEDVISTA-RL, an agent trained to interleave multi-turn reasoning with structured tool use. Our findings indicate that effective tool-integrated medical reasoning does not arise from tool access alone, but instead requires learning disciplined interaction policies. In particular, successful training relies on (i) explicitly interleaving reasoning with tool invocation, (ii) a two-stage paradigm that combines supervised warm-up with online reinforcement learning to refine multi-turn tool reasoning, and (iii) broad action-space coverage through diverse medical imaging tasks and tools to promote generalizable interaction behaviors. By providing unified interfaces, executable feedback, and efficient trajectory logging, MEDVISTAGYM serves as a learnable training substrate that supports systematic progress in medical image analysis.

656 Limitations

657 Although MEDVISTA-R1 achieve substantial im-
658 provements in tool-integrated medical visual rea-
659 soning using MEDVISTAGYM, several limitations
660 remain. Despite optimization, multi-turn agentic
661 training with frequent tool invocation is computa-
662 tionally expensive, and rollout depth and action-
663 space size impose practical constraints on exhaust-
664 ive exploration of all possible reasoning paths. The
665 current environment and evaluation focus primar-
666 ily on medical VQA tasks, and extending MED-
667 VISTAGYM to other clinical reasoning paradigms
668 or non-medical domains may require additional
669 task and tool adaptation. Furthermore, tool effec-
670 tiveness can be limited when tool outputs are noisy
671 or when images exhibit low quality, subtle abnor-
672 malities, or highly complex visual patterns, which
673 may still lead to incorrect reasoning.

674 Ethical Considerations

675 Ethical considerations are integral to this work. All
676 experiments are conducted using publicly available
677 datasets, and we employ open-source or widely
678 adopted models without accessing any private pa-
679 tient data. Our study focuses on methodological ad-
680 vances in tool-integrated medical visual reasoning
681 rather than clinical deployment, and model outputs
682 should not be interpreted as medical advice. We
683 emphasize transparency throughout the research
684 process and advocate responsible use of these tech-
685 niques with appropriate human oversight to ensure
686 safe and beneficial application.

687 References

688 Anthropic. 2025a. Claude sonnet 4.5. <https://www.anthropic.com/claude/sonnet>. Accessed: 2026-01-04.
689
690
691 Anthropic. 2025b. Introducing claude haiku
692 4.5. <https://www.anthropic.com/news/claude-haiku-4-5>. Accessed: 2026-01-04.
693
694 Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen,
695 Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei
696 Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhi-
697 fang Guo, Qidong Huang, Jie Huang, Fei Huang,
698 Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng
699 Li, and 45 others. 2025a. *Qwen3-vl technical report*.
700 *Preprint*, arXiv:2511.21631.

701 Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wen-
702 bin Ge, Sibao Song, Kai Dang, Peng Wang, Shi-
703 jie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu,
704 Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei

Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 oth-
ers. 2025b. *Qwen2.5-vl technical report*. *Preprint*,
arXiv:2502.13923.

James Burgess, Jeffrey J Nirschl, Laura Bravo-Sánchez,
Alejandro Lozano, Sanket Rajan Gupte, Jesus G.
Galaz-Montoya, Yuhui Zhang, Yuchang Su, Disha
Bhowmik, Zachary Coman, Sarina M. Hasan,
Alexandra Johannesson, William D. Leineweber,
Malvika G Nair, Ridhi Yarlagadda, Connor Zuraski,
Wah Chiu, Sarah Cohen, Jan N. Hansen, and 4
others. 2025. *Microvqa: A multimodal reasoning
benchmark for microscopy-based scientific research*.
Preprint, arXiv:2503.13399.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wan-
long Liu, Rongsheng Wang, Jianye Hou, and Benyou
Wang. 2024a. *Huatuogpt-o1: Towards medical com-
plex reasoning with large language models*. *arXiv
preprint arXiv:2412.18925*.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe
Gao, Shunian Chen, Guiming Hardy Chen, Xi-
dong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji,
Guangjun Yu, Xiang Wan, and Benyou Wang. 2024b.
*Huatuogpt-vision, towards injecting medical visual
knowledge into multimodal llms at scale*. *Preprint*,
arXiv:2406.19280.

Zhihong Chen, Maya Varma, Justin Xu, Magdalini
Paschali, Dave Van Veen, Andrew Johnston, Alaa
Youssef, Louis Blankemeier, Christian Bluethgen,
Stephan Altmayer, Jeya Maria Jose Valanarasu, Mo-
hamed Siddig Eltayeb Muneer, Eduardo Pontes Reis,
Joseph Paul Cohen, Cameron Olsen, Tanishq Mathew
Abraham, Emily B. Tsai, Christopher F. Beaulieu, Je-
nia Jitsev, and 4 others. 2024c. *A vision-language
foundation model to enhance efficiency of chest x-ray
interpretation*. *Preprint*, arXiv:2401.12208.

Adibvafa Fallahpour, Jun Ma, Alif Munim, Hong-
wei Lyu, and Bo Wang. 2025. *Medrax: Med-
ical reasoning agent for chest x-ray*. *Preprint*,
arXiv:2502.02673.

Yue Fan, Xuehai He, Dijia Yang, Kaizhi Zheng, Ching-
Chen Kuo, Yuting Zheng, Sravana Jyothi Naraya-
naraju, Xinze Guan, and Xin Eric Wang. 2025. *Grit:
Teaching mllms to think with images*. *arXiv preprint
arXiv:2505.15879*.

Nima Fathi, Amar Kumar, and Tal Arbel. 2025. *Aura:
A multi-modal medical agent for understanding, rea-
soning & annotation*. *Preprint*, arXiv:2507.16940.

Google. 2025a. Gemini-2.5-flash-image model —
google ai studio. [https://aistudio.google.com/
models/gemini-2-5-flash-image](https://aistudio.google.com/models/gemini-2-5-flash-image). Accessed:
2026-01-04.

Google. 2025b. Gemini 2.5 pro. [https:
//aistudio.google.com/app/prompts/new_
chat?model=gemini-2.5-pro](https://aistudio.google.com/app/prompts/new_chat?model=gemini-2.5-pro). Accessed: 2026-
01-04.

760	Debanjan Goswami, Ronast Subedi, and Shayok Chakraborty. 2025. MediVLM: A vision language model for radiology report generation from medical images . In <i>Findings of the Association for Computational Linguistics: EMNLP 2025</i> , pages 10287–10304, Suzhou, China. Association for Computational Linguistics.	817
761		818
762		819
763		820
764		821
765		822
766		
767	Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025a. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. <i>Nature</i> , 645(8081):633–638.	823
768		824
769		825
770		826
771		
772	Guangfu Guo, Xiaoqian Lu, and Yue Feng. 2025b. Medvragent: A framework for medical visual reasoning-enhanced agents . <i>Preprint</i> , arXiv:2510.18424.	827
773		828
774		829
775	Xingang Guo, Utkarsh Tyagi, Advait Gosai, Paula Vergara, Ernesto Gabriel Hernández Montoya, Chen Bo Calvin Zhang, Bin Hu, Yunzhong He, Bing Liu, and Rakshith Sharma Srinivasa. 2025c. Beyond seeing: Evaluating multimodal llms on tool-enabled image perception, transformation, and reasoning. <i>arXiv preprint arXiv:2510.12712</i> .	830
776		831
777		832
778		833
779		834
780		
781		
782	Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. Pathvqa: 30000+ questions for medical visual question answering . <i>Preprint</i> , arXiv:2003.10286.	835
783		836
784		837
785		838
786	Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. 2025. Vision-r1: Incentivizing reasoning capability in multimodal large language models. <i>arXiv preprint arXiv:2503.06749</i> .	839
787		840
788		841
789		
790		
791	Dongfu Jiang, Yi Lu, Zhuofeng Li, Zhiheng Lyu, Ping Nie, Haozhe Wang, Alex Su, Hui Chen, Kai Zou, Chao Du, Tianyu Pang, and Wenhui Chen. 2025. Verl-tool: Towards holistic agentic reinforcement learning with tool use . <i>Preprint</i> , arXiv:2509.01055.	842
792		843
793		844
794		845
795		846
796	Xin Lai, Junyi Li, Wei Li, Tao Liu, Tianjian Li, and Hengshuang Zhao. 2025. Mini-o3: Scaling up reasoning patterns and interaction turns for visual search . <i>Preprint</i> , arXiv:2509.07969.	847
797		848
798		849
799		
800	James Lau, Sam Gayen, Asma Ben Abacha, and 1 others. 2018. A dataset of clinically generated visual questions and answers about radiology images . <i>Scientific Data</i> , 5:180251.	850
801		851
802		852
803		853
804	Binxu Li, Tiankai Yan, Yuanting Pan, Jie Luo, Ruiyang Ji, Jiayuan Ding, Zhe Xu, Shilong Liu, Haoyu Dong, Zihao Lin, and Yixin Wang. 2024a. Mmedagent: Learning to use medical tools with multi-modal agent . <i>Preprint</i> , arXiv:2407.02483.	854
805		855
806		856
807		
808		
809	Binxu Li, Tiankai Yan, Yuanting Pan, Jie Luo, Ruiyang Ji, Jiayuan Ding, Zhe Xu, Shilong Liu, Haoyu Dong, Zihao Lin, and Yixin Wang. 2024b. MMedAgent: Learning to use medical tools with multi-modal agent . In <i>Findings of the Association for Computational Linguistics: EMNLP 2024</i> , pages 8745–8760, Miami, Florida, USA. Association for Computational Linguistics.	857
810		858
811		859
812		860
813		
814		
815		
816		
	Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day . <i>Preprint</i> , arXiv:2306.00890.	861
		862
		863
		864
	Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering . <i>Preprint</i> , arXiv:2102.09542.	865
		866
		867
		868
	Xinyao Liu and Diping Song. 2025. Constructing ophthalmic mllm for positioning-diagnosis collaboration through clinical cognitive chain reasoning . <i>Preprint</i> , arXiv:2507.17539.	869
		870
	Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. 2025. Visual-rft: Visual reinforcement fine-tuning . <i>arXiv preprint arXiv:2503.01785</i> .	
	Meng Lu, Ran Xu, Yi Fang, Wenxuan Zhang, Yue Yu, Gaurav Srivastava, Yuchen Zhuang, Mohamed Elhoseiny, Charles Fleming, Carl Yang, Zhengzhong Tu, Yang Xie, Guanghua Xiao, Hanrui Wang, Di Jin, Wenqi Shi, and Xuan Wang. 2025. Scaling agentic reinforcement learning for tool-integrated reasoning in vlms . <i>Preprint</i> , arXiv:2511.19773.	
	Vishwesh Nath, Wenqi Li, Dong Yang, Andriy Myronenko, Mingxin Zheng, Yao Lu, Zhijian Liu, Hongxu Yin, Yucheng Tang, Pengfei Guo, Can Zhao, Ziyue Xu, Yufan He, Greg Heinrich, Yee Man Law, Benjamin Simon, Stephanie Harmon, Stephen Aylward, Marc Edgar, and 6 others. 2025. Vila-m3: Enhancing vision-language models with medical expert knowledge . <i>Preprint</i> , arXiv:2411.12915.	
	OpenAI. 2024. Gpt-4o mini: Advancing cost-efficient intelligence . https://openai.com/zh-Hans-CN/index/gpt-4o-mini-advancing-cost-efficient-intelligence/ . Accessed: 2026-01-04.	
	OpenAI. 2025. Gpt-5 . https://openai.com/zh-Hans-CN/gpt-5/ . 2025-12-23.	
	Sebastián Ramírez and FastAPI Contributors. 2018. Fastapi: High-performance web framework for building apis with python . https://github.com/fastapi/fastapi . Accessed: 2025-09-17.	
	Yijia Shao, Vinay Samuel, Yucheng Jiang, John Yang, and Diyi Yang. 2024. Collaborative gym: A framework for enabling and evaluating human-agent collaboration . <i>arXiv preprint arXiv:2412.15701</i> .	
	Naman Sharma. 2024. Cxr-agent: Vision-language models for chest x-ray interpretation with uncertainty aware radiology reporting . <i>Preprint</i> , arXiv:2407.08811.	
	Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide	

871	Zeng, Zhengyuan Yang, and 1 others. 2025. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. <i>arXiv preprint arXiv:2506.23918</i> .	
872		
873		
874		
875	Haoran Sun, Yankai Jiang, Wenjie Lou, Yujie Zhang, Wenjie Li, Lilong Wang, Mianxin Liu, Lei Liu, and Xiaosong Wang. 2025. Chiron-01: Igniting multimodal large language models towards generalizable medical reasoning via mentor-intern collaborative search . <i>Preprint</i> , arXiv:2506.16962.	
876		
877		
878		
879		
880		
881	LASA Team, Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, Yu Sun, Junao Shen, Chaojun Wang, Jie Tan, Deli Zhao, Tingyang Xu, Hao Zhang, and Yu Rong. 2025. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning . <i>Preprint</i> , arXiv:2506.07044.	
882		
883		
884		
885		
886		
887		
888		
889	Haozhe Wang, Alex Su, Weiming Ren, Fangzhen Lin, and Wenhui Chen. 2025a. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning . <i>Preprint</i> , arXiv:2505.15966.	
890		
891		
892		
893	Kangrui Wang, Pingyue Zhang, Zihan Wang, Yan-ang Gao, Linjie Li, Qineng Wang, Hanyang Chen, Chi Wan, Yiping Lu, Zhengyuan Yang, and 1 others. 2025b. Vagen: Reinforcing world model reasoning for multi-turn vlm agents . <i>arXiv preprint arXiv:2510.16907</i> .	
894		
895		
896		
897		
898		
899	Zihan Wang, Kangrui Wang, Qineng Wang, Pingyue Zhang, Linjie Li, Zhengyuan Yang, Xing Jin, Kefan Yu, Minh Nhat Nguyen, Licheng Liu, and 1 others. 2025c. Ragen: Understanding self-evolution in llm agents via multi-turn reinforcement learning . <i>arXiv preprint arXiv:2504.20073</i> .	
900		
901		
902		
903		
904		
905	Ziyue Wang, Junde Wu, Linghan Cai, Chang Han Low, Xihong Yang, Qiaxuan Li, and Yueming Jin. 2025d. Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow . <i>Preprint</i> , arXiv:2503.18968.	
906		
907		
908		
909		
910	Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models . In <i>Advances in Neural Information Processing Systems</i> .	
911		
912		
913		
914		
915	Zhiheng Xi, Jixuan Huang, Chenyang Liao, Baodai Huang, Honglin Guo, Jiaqi Liu, Rui Zheng, Junjie Ye, Jiazheng Zhang, Wenxiang Chen, and 1 others. 2025. Agentgym-rl: Training llm agents for long-horizon decision making through multi-turn reinforcement learning . <i>arXiv preprint arXiv:2509.08755</i> .	
916		
917		
918		
919		
920		
921	Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. 2025. Mmed-rag: Versatile multimodal rag system for medical vision language models . <i>Preprint</i> , arXiv:2410.13085.	
922		
923		
924		
925		
	Ran Xu, Yuchen Zhuang, Yishan Zhong, Yue Yu, Xiangyu Tang, Hang Wu, May D Wang, Peifeng Ruan, Donghan Yang, Tao Wang, and 1 others. 2025. Medagentgym: Training llm agents for code-based medical reasoning at scale . <i>arXiv preprint arXiv:2506.04405</i> .	926
		927
		928
		929
		930
	Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi . <i>Preprint</i> , arXiv:2311.16502.	931
		932
		933
		934
		935
		936
		937
		938
	Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Pmc-vqa: Visual instruction tuning for medical visual question answering . <i>Preprint</i> , arXiv:2305.10415.	939
		940
		941
		942
	Xintong Zhang, Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaowen Zhang, Yang Liu, Tao Yuan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, and 1 others. 2025. Chain-of-focus: Adaptive visual search and zooming for multimodal reasoning via rl . <i>arXiv preprint arXiv:2505.15436</i> .	943
		944
		945
		946
		947
		948
	Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. 2025. Deepeyes: Incentivizing "thinking with images" via reinforcement learning . <i>arXiv preprint arXiv:2505.14362</i> .	949
		950
		951
		952
		953
	Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models . <i>arXiv preprint arXiv:2504.10479</i> .	954
		955
		956
		957
		958
		959

A Dataset Details

A.1 Details of Curated Datasets

We curate our data following three core principles. (1) Diverse tasks and imaging distributions: we incorporate heterogeneous datasets to enhance model generalization. (2) Tool effectiveness: we prioritize scenarios where tool usage leads to measurable accuracy improvements. In generating reasoning trajectories, unlike prior work that first plans tool usage and then executes multiple tools in a pre-determined order, we allow GPT-5 to operate directly within the MEDVISTAGYM environment and generate interactive reasoning trajectories through multiple rounds of real tool invocations. We retain only trajectories that produce correct final answers and further apply GPT-5 for self-verification. We believe this approach more faithfully captures the practical reasoning process involved in multi-tool collaboration. These trajectories are filtered through format validation and answer-correctness checks. The prompt template for constructing agentic reasoning trajectories and the prompt for trajectory verification are provided in Figure 5 and Figure 6, respectively.

Datasets	VQA-RAD	Slake	PathVqa	PMCVQA	MMMU(H&M)	MicroVQA
Train	3500	3500	3500	0	0	0
Test	400	400	400	400	200	300

Table 3: Data statistics.

A.2 Dataset Details

We conduct experiments on six representative datasets. SLAKE (Liu et al., 2021), PathVQA (He et al., 2020) and VQA-RAD (Lau et al., 2018) are widely used benchmarks in medical VQA research. For higher-level medical reasoning, we further evaluate PMCVQA (Zhang et al., 2024), a generative medical visual question answering benchmark that requires models to perform fine-grained visual understanding and domain-specific reasoning beyond fixed answer classification, MMMU(H&M) (Yue et al., 2024) which focuses on expert-level multimodal understanding and reasoning over complex medical imagery, and MicroVQA (Burgess et al., 2025), which emphasizes fine-grained, region-level visual reasoning. The statistics of the datasets are shown in Figure 3. For evaluation metrics, we use answer accuracy for multiple-choice VQA benchmarks. MMMU(H&M) (Yue et al., 2024) is a subset extracted from the multimodal reasoning benchmark MMMU (Yue et al., 2024).

B Additional Experimental Results Analysis

B.1 Effect of Tool-Integrated Reasoning

Table 1 further disentangles the contributions of reasoning and tool use by comparing the following controlled variants: (1) *GRPO w/o Tools* (33.29%), which retains multi-step reasoning through reinforcement learning while disabling all external tool access during both training and inference; (2) *Cold-start w/o Reasoning* (37.42%), which exposes the model to external tools using a fixed, predefined invocation format learned during cold-start supervision, but removes the reinforcement learning stage that induces adaptive reasoning over tool usage; (3) *Cold-start w/o Tools* (33.00%), which performs only supervised fine-tuning without tool access or reinforcement learning; (4) *Direct GRPO w/o Cold-start* (31.33%), which initializes reinforcement learning directly from the base backbone without cold-start supervision, allowing tool access but without prior exposure to tool syntax or invocation patterns, leading to unstable exploration in the large action space.

B.2 Additional Baselines

As shown in Table 4, we obtain the following observations:

Compared with recent SOTA general-purpose VLMs, MEDVISTA-R1 outperforms DeepEyes-7B (Zheng et al., 2025) with tools (e.g., zoom-in), Mini-o3-7B (Lai et al., 2025), and PixelReasoner (Wang et al., 2025a) by an average of 11.50%, demonstrating superior performance and highlighting the effectiveness of environment-driven tool interactions.

Compared with SOTA medical-specific reasoning VLMs, MEDVISTA-R1 outperforms LLaVamed-7B (Li et al., 2023) 25.43%, HuatuoGTP-vision-34B (Chen et al., 2024a) 13.47%, Chiron-o1-8B (Sun et al., 2025) 0.80%, and Lingshu-7B (Team et al., 2025) 2.84%, demonstrating that integrating external tools enables MLLMs to acquire and utilize fine-grained visual cues for medical image reasoning more effectively.

Compared with prior medical agent frameworks, including MMedAgent-7B (Li et al., 2024a), VILA-M3-40B (Nath et al., 2025) and MMedAgent-RL-7B (Li et al., 2024b), MEDVISTA-R1 outperform by average of 8.46%, which demonstrates a substantial margin of improvement. Unlike approaches that treat tool usage as isolated function calls, our

Models	Tool	PathVQA	Slake	VQARAD	Avg.
<i>MLLMs can think with image</i>					
DeepEyes-7B	✓	52.9	68.2	65.9	62.33
Mini-o3-7B-vl	✓	53.4	67.8	65.7	62.30
PixelReasoner-RL-vl-7B	✓	52.6	67.3	66.0	61.97
<i>Opensource SOTA</i>					
Llava-Next-13B	✗	39.8	57.1	54.8	50.57
Qwen2.5vl-32B	✗	47.4	70.1	71.7	63.07
<i>Medical MLLMs</i>					
LlaVa-med-7B	✗	44.6	47.7	52.5	48.27
HuatuoGPT-Vision-34B	✗	50.7	68.3	61.7	60.23
<i>Medical MLLMs with CoT Reasoning</i>					
MedVLM-R1-2B	✗	38.3	54.3	45.2	45.93
Med-R1-2B	✗	19.2	52.1	36.5	35.93
Lingshu-7B	✗	68.4	77.8	66.4	70.87
Chiron-01-8B	✗	68.8	77.4	72.5	72.90
<i>Multimodal medical agents</i>					
MMedAgent-7B	✓	59.47	68.7	64.0	64.06
AURA	✓	59.8	68.4	64.5	64.23
SMR-Agents	✓	38.2	53.5	46.9	46.20
MedAgent-Pro	✓	58.5	69.4	63.3	63.73
MMedAgentRL-7B	✗	58.5	67.9	66.1	64.17
VILA-M3-40B	✓	66.4	71.4	65.7	67.83
MEDVISTA-R1	✓	69.00	81.36	70.75	73.70

Table 4: Performance (%) SOTA models comparison on medical VQA benchmarks.

agent–environment framework enables structured, multi-turn interaction with external medical tools, producing verifiable reasoning trajectories that support coherent multi-tool composition and more faithful visual grounding.

C Reward Details

We design the reward function to provide rich and fine-grained feedback signals. The individual reward components are described below.

The reasoning-action format reward The format reward S_{format} evaluates the structural validity of R by verifying that the model output contains all required special tokens in the prescribed order: `<think></think><tool_call></tool_call><think></think><answer></answer>`. Specifically, before invoking a tool, the model must enclose its tool-selection reasoning within `<think></think>` tags and place the tool-call json between `<tool_call></tool_call>` tags. After tool execution, instead of directly producing the final answer, the model is required to include an additional reasoning step enclosed by `<think></think>` and then output the final answer within `<answer></answer>`. Outputs that strictly follow this structured reasoning–action format receive a positive reward.

The final-answer accuracy reward The final-answer accuracy reward S_{ans} evaluates whether the predicted answer matches the ground-truth answer.

The answer-conditioned tool-use reward To align with the interaction paradigm in MEDVISTAGYM, where correct reasoning trajectories are generated

through multiple rounds of tool invocation, we introduce a conditional reward mechanism that grants an additional reward only when the model both appropriately invokes external tools during the trajectory and produces a correct final answer. This design encourages meaningful tool usage when tools substantively contribute to successful task completion, rather than arbitrary or redundant tool calls.

D Optimization Details

We optimize the agent in Stage II (Section 3.2) using Group Relative Policy Optimization (GRPO) (Guo et al., 2025a), which performs policy updates based on relative performance among a group of sampled rollouts.

Multi-turn Rollouts Given an input (I, Q) , we sample a group of G multi-turn interaction trajectories $\{\tau_1, \dots, \tau_G\}$ from the current policy $\pi_{\theta_{\text{old}}}$. Each trajectory $\tau_i = (u_{i,1}, \dots, u_{i,T_i})$ consists of interleaved reasoning tokens, tool-call tokens, and the final answer, following the rollout formulation defined in Section 3.2. Each trajectory is assigned a scalar rollout-level reward $R(\tau_i)$ according to the reward design in Section 3.2.

Group-Normalized Advantage To emphasize relative quality among sampled reasoning paths, we compute a group-normalized advantage for each trajectory:

$$A_i = \frac{R(\tau_i) - \text{mean}(\{R(\tau_1), \dots, R(\tau_G)\})}{\text{std}(\{R(\tau_1), \dots, R(\tau_G)\})}. \quad (1)$$

This normalization encourages the policy to prefer trajectories that perform better than others within the same rollout group, rather than relying on absolute reward magnitudes.

Token-Level GRPO Objective Policy optimization is performed at the token level. For the k -th token in trajectory τ_i , we define the importance ratio:

$$r_{i,k}(\theta) = \frac{\pi_{\theta}(\tau_{i,k} | \tau_{i,<k})}{\pi_{\theta_{\text{old}}}(\tau_{i,k} | \tau_{i,<k})}. \quad (2)$$

The GRPO objective is then defined as:

$$\mathcal{L}_{\text{GRPO}}(\theta) = \frac{1}{G} \sum_{i=1}^G \frac{1}{|\tau_i|} \sum_{k=1}^{|\tau_i|} \min \left(r_{i,k}(\theta) \cdot A_i, \text{clip}(r_{i,k}(\theta), 1 - \epsilon, 1 + \epsilon) \cdot A_i \right). \quad (3)$$

where $|\tau_i|$ denotes the number of trainable tokens in trajectory τ_i , excluding tool-executed observations.

This objective updates the policy to increase the likelihood of tokens belonging to higher-quality reasoning trajectories while maintaining stable updates through clipping.

E Tools Details

In MEDVISTAGYM, we enable access to 15 tools organized into four complementary families.

Resolution and Region Refinement This family enables focused inspection and quality enhancement of image regions by improving visual fidelity and local details.

4KAgent is an agentic image enhancement system that supports super-resolution upscaling with configurable scale factors ($2\times/4\times/8\times/16\times$) using models such as HAT-PSNR, DiffBIR, and OSediff. It further provides dehazing (DehazeFormer, RIDCP, MAXIM), denoising (NAFNet, Restormer, SwinIR), and image brightening (CLAHE, Gamma Correction, FourierDiff), allowing agents to recover fine-grained visual cues from degraded or low-quality medical images.

Medical Localization and Segmentation This family supports the detection and delineation of anatomical and pathological regions, providing region-level evidence for downstream reasoning.

GroundingDINO performs open-set, text-conditioned object detection, localizing anatomical structures or pathological regions described by natural-language queries.

SAM2 is a promptable foundation model for image and video segmentation that enables zero-shot segmentation using points, bounding boxes, or masks. *BiomedParse* is a unified biomedical image parsing model that segments organs and pathological regions across diverse medical imaging modalities, supporting organ-level, lesion-level, comprehensive, and text-prompted segmentation.

MedSAM2 is a medical-domain adaptation of the Segment Anything Model, optimized for high-precision anatomical segmentation with improved boundary accuracy using bounding-box prompts and organ names.

Medical Visual Understanding and Parsing

This family provides higher-level semantic interpretation of medical images.

BiomedCLIP is a vision-language foundation model pre-trained on large-scale biomedical image-text pairs. It enables zero-shot medical image

classification across multiple label types (e.g., abnormality, modality, organ) by computing vision-language similarity scores, producing structured semantic signals for reasoning.

External Biomedical Knowledge Retrieval In MEDVISTAGYM, agents may retrieve external medical knowledge from resources such as *DrugBank* and *PubMed*, allowing integration of visual findings with domain-specific medical knowledge when required.

Category	Main Function	Specific tools	Key Capability
Image Enhancement	4KAgent	super_resolution	Super-resolution upscaling ($2\times-16\times$)
		dehazing denoising brightening	Haze and fog removal Noise artifact elimination Low-light enhancement
Perception	GroundingDINO SAM2	grounding_dino sam2	Open-set object detection Universal promptable segmentation
		BiomedCLIP	Zero-shot medical classification
Medical Analysis	BiomedParse	biomedparse_organ	Anatomical structure segmentation
		biomedparse_lesion	Pathological region detection
		biomedparse_all	Comprehensive target segmentation
		biomedparse_text	Custom text-prompted segmentation
	MedSAM2	medsam2	Prompt-based medical segmentation
Knowledge Retrieval	GoogleSearch	google_search	Real-time web search
	DrugBank	drugbank	Pharmaceutical knowledge lookup
	LongDocRAG	longdocrag	Long document QA

Table 5: Summary of tools in MEDVISTAGYM.

F Additional Ablation Analysis

In this section, we provide additional ablation studies to further analyze the behavior and generalization properties of our framework. Specifically, we examine: (1) whether the model can reliably invoke tools with correct formatting and accurate tool selection, and whether training effectively eliminates unstructured or erroneous tool calls (§F.1); (2) how the quantity and type of tools influence agent performance, including the scaling effects of tool composition (§F.2, §F.3); (3) whether the learned tool-use policy generalizes to unseen tools without additional training (§F.2); and (4) how increasing model size further amplifies the benefits of tool-integrated reasoning (§F.4).

F.1 The Improvement of Tool Invocation

Figure 3 further measures the reliability of tool invocation by jointly checking format adherence and correct tool calling, offering a complementary perspective to the error analysis discussed earlier. The vanilla model without tool-use training exhibits poor reliability, achieving only a 24.2% accuracy rate across evaluation benchmarks. In contrast, our trained model achieves near-ceiling performance with a 98.96% accuracy rate, demonstrating consistently precise and reliable tool-use behavior. The absolute improvement of approximately 75% highlights the effectiveness of our

Table 6: 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. Avg. is the mean over available metrics per row.

Method	Tools #	Tools (used in this setting)						Benchmarks		
		Seen					Unseen	VQA tasks		Avg.
		Zoom-in	MedSAM	BioMed Parse	Biomed CLIP	4KAgent	SAM2	VQA-RAD	SLAKE	
DeepEyes-7B	1	✓						65.9	68.2	67.1
Mini-o3-7B-v1	1	✓						65.7	67.8	66.8
PixelReasoner-RL-v1-7B	1	✓						66.0	67.3	66.7
MMedAgent-MedSAM	1		✓					64.0	68.7	66.4
MEDVISTA-R1	1	✓						65.1	69.5	67.3
	1						✓	64.5	72.8	68.7
	1		✓					65.1	72.5	68.8
	1			✓				66.9	73.5	70.2
	2		✓	✓				67.1	75.2	71.2
	2	✓	✓					67.6	76.0	71.8
	2	✓		✓				68.4	77.3	72.9
MEDVISTA-R1	5	✓	✓	✓	✓	✓		70.8	79.7	75.3

training paradigm. These results confirm that the model learns to reliably and proactively invoke appropriate tools when solving problems. This further corroborates the importance of our carefully designed training strategy for eliciting robust tool-calling capabilities—intelligent behaviors that mere prompting of a base model cannot achieve.

F.2 Impact of Tool Quantity and Type

We investigate the ability of our model to utilize out-of-domain (OOD) tools in a training-free setting, evaluated on both in-domain datasets and an OOD benchmark. For fair comparison with current SOTA baselines, we evaluate each agent under a single-tool setting. As shown in Table 6, when only the zoom-in tool is available, MEDVISTA-R1 maintains a clear performance advantage on the SLAKE dataset. This demonstrates that MEDVISTAGYM supports deeply tool-integrated reasoning even under severely constrained tool access. When provided with an unseen tool, SAM2, which is functionally similar to MedSAM, MEDVISTA-R1 achieves performance on SLAKE that is comparable to using seen tools, highlighting strong generalization across similar tool types.

F.3 Scaling of Tools Yields Consistent Gains

As shown in Table 6, overall performance improves steadily as the model gains access to a larger set of tools within the environment. Under single-tool settings, BiomedParse yields the best performance (73.74%), clearly surpassing zoom-in only (69.50%) and MedSAM only (72.51%). Moreover,

augmenting BiomedParse with MedSAM results in a further (+)1.78% performance gain, demonstrating that MEDVISTA-R1, after training in MEDVISTAGYM, is able to effectively coordinate multiple tool invocations and perform tool-integrated reasoning.

F.4 Scaling Model Sizes Yields Consistent Gains

We analyze the effect of model size on performance within MEDVISTAGYM. As shown in Table 1, scaling the backbone from 2B to 8B leads to consistent performance gains across benchmarks. In particular, MEDVISTA-R1-8B outperforms its 2B counterpart in average accuracy, with larger improvements on reasoning-intensive datasets that require multi-step visual evidence aggregation and tool-conditioned decision making. This indicates that larger models can more effectively support long-horizon, tool-integrated reasoning within MEDVISTAGYM.

Notably, despite its smaller parameter count, MEDVISTA-R1-2B matches or even surpasses several 8B baselines. This highlights the strong parameter efficiency of our framework and suggests that the observed gains stem not merely from increased model size, but from the synergy between agentic reasoning, structured tool use, and reinforcement learning. Scaling the backbone further amplifies these advantages, yielding reliable and consistent improvements rather than qualitatively different behaviors.

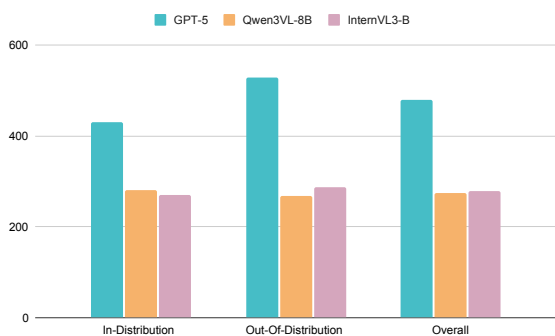


Figure 4: Effect of thinking trajectory length.

Overall, these results demonstrate that MED-VISTAGYM serves as a scalable training ground for medical VLM agents. Model scaling functions as a complementary factor that enhances agentic, tool-integrated reasoning, rather than acting as the primary driver of performance gains.

F.5 Effect of Thinking Trajectory Length

We analyze the relationship between thinking trajectory length and downstream performance under the vanilla with tool setting. As shown in Figure 4, we report the average reasoning trajectory length of GPT-5, Qwen3VL-8B, and InternVL3-8B across the In-Distribution, Out-of-Distribution, and Overall settings.

We observe that GPT-5 consistently produces substantially longer reasoning trajectories across all three settings, whereas Qwen3VL-8B and InternVL3-8B generate much shorter and comparable trajectories. Consistent with this trend, GPT-5 achieves markedly higher performance in all evaluation scenarios, with the most pronounced gains in the In-Distribution and Overall settings. In contrast, although Qwen3VL-8B and InternVL3-8B are also capable of invoking external tools, their shorter reasoning trajectories are associated with more limited performance improvements.

These results suggest that, in the absence of agentic reinforcement learning, tool availability alone is insufficient to guarantee performance gains. Instead, the ability to generate sufficient and well-structured intermediate reasoning trajectories plays a critical role in effectively leveraging tools for complex problem solving. This observation further validates our design choice of constructing explicit thinking trajectories using GPT-5 with tools in the MEDVISTAGYM environment as supervision for SFT, providing high-quality reasoning traces that are difficult to elicit from vanilla open-source mod-

els.

G Prompt Details

G.1 Prompt For Reasoning Trajectory Construction and Filtering

In this section, we introduce the prompt design for constructing reliable agentic reasoning trajectories over our curated VQA data. Unlike generic approaches that directly sample from a teacher model, and unlike methods that predefine a static tool set for GPT-5 or other large models, we provide GPT-5 with ground-truth clinical metadata and executable tool-calling interfaces during prompting, enabling real tool interaction within the MEDVISTAGYM environment. At each step, the model triggers a tool call based on the current action and determines the subsequent action according to the returned observation, with tool outputs fed back to the model immediately after each interaction.

Throughout the entire environment interaction and reasoning process, the agent is explicitly constrained to treat the final answer as unknown, which is revealed only during the verification stage. Combined with verified medical annotations, this prompt design enables multi-turn rollouts with environment interaction and tool use to construct step-by-step reasoning trajectories. Figure 9 illustrates the generation template, while Figure 5 presents the trajectory quality evaluation and filtering where GPT-5 assigns fine-grained scores to different aspects of the reasoning process, including tool-use justification, effective integration of observations, and exact answer matching.

G.2 System Prompt and User Prompt

We present the two prompt templates that we used in our experiments. Figure 7 shows the System Prompt template. Figure 8 shows the User Prompt template

H Interface and Infrastructure Details

H.1 Agent-Environment Interface Details

MEDVISTAGYM is an executable training environment built upon a Gym-style interaction protocol and specifically designed for medical image analysis. It provides a flexible *API interface*, an explicit and verifiable *action space*, and a structured *observation space*, enabling agents to perform multi-step reasoning through continuous interaction with the environment in medical settings.

System Prompt for Trajectory Evaluation:

You are an expert evaluator for tool-augmented medical VQA reasoning trajectories.

Context:

- The final answer is ALREADY CORRECT and FIXED. Do NOT evaluate answer correctness.
- All tool executions completed successfully (though outputs may or may not be useful).
- Your job is to evaluate the QUALITY OF REASONING in the generated think segments.

Evaluation Dimensions:

1. *Pre-tool Reasoning Quality* (for each tool call): Does it clearly explain WHY this specific tool is being called? Is the reasoning for tool selection appropriate? Does it avoid “hindsight bias”? Score: 1 (poor) - 5 (excellent)

2. *Tool Selection Appropriateness*: Is the chosen tool appropriate for the stated purpose? Are the tool arguments reasonable for the medical context? Score: 1 (poor) - 5 (excellent)

3. *Post-observation Reasoning Quality*: Does it correctly interpret what the tool returned? If the tool output was unhelpful, does the reasoning acknowledge this failure? Score: 1 (poor) - 5 (excellent)

4. *Final Reasoning to Answer*: Does it properly synthesize all gathered evidence? Is the logical chain from observations to conclusion clear? Score: 1 (poor) - 5 (excellent)

5. *Overall Coherence*: Is the reasoning flow natural and logical throughout? Is the medical terminology used correctly? Score: 1 (poor) - 5 (excellent)

Critical Issues to Flag: hindsight_bias, hallucination, tool_failure_ignored, logic_gap, inconsistent_reasoning

Output Format (strict JSON):

```
{
  "evaluation": {
    "pre_tool_reasoning": {
      "score": 1-5,
      "comments": "...",
      "tool_selection": {...},
      "post_observation_reasoning": {...},
      "final_reasoning": {...},
      "overall_coherence": {...},
      "issues_found": [...],
      "overall_score": 1-5,
      "quality_tier": "excellent/good/acceptable/needs_improvement/poor",
      "summary": "..."
    }
  }
}
```

IMPORTANT: Output ONLY valid JSON. No markdown, no code blocks, no extra text.

Figure 5: Prompt template for trajectory quality evaluation.

Executable Medical Interface. MEDVISTAGYM provides two core interface functions, `reset()` and `step()`. Calling `reset()` initializes a new interaction episode and returns the initial observation o_0 , which contains the current medical question along with the associated medical images. Each episode corresponds to a complete and independent agent–environment interaction instance.

Executable Medical Action Space The action space \mathcal{A} is strictly restricted to the set of executable medical tools defined in MEDVISTAGYM. Each action $a_t \in \mathcal{A}$ is formalized as a typed tuple that explicitly specifies the selected tool identifier and its corresponding arguments, which are passed to the appropriate tool interface for execution. This design ensures that all agent actions are explicitly defined, executable, and verifiable.

Medical Evidence Observation Space. After an action a_t is executed, the environment returns an observation $o_t \in \mathcal{O}$ containing structured tool outputs—such as region localizations, segmentation masks, quantitative measurements, or retrieved medical facts—as well as potential execution-time error messages. These observations serve as external evidence that supports subsequent medical reasoning and decision-making.

H.2 Scalable Execution Infrastructure Details

To enhance fine-grained medical visual perception and domain grounding, we incorporate compute-intensive medical foundation models as interactive tools, including high-resolution visual encoders and medical segmentation models.

To support large-scale, multi-turn medical visual reasoning, we design a scalable execution infrastructure that encapsulates computationally intensive medical foundation models as interactive tools. This design enhances fine-grained medical visual perception and domain grounding by enabling on-demand invocation of high-resolution visual encoders and medical image segmentation models, which provide reliable intermediate visual evidence during reasoning. All tools are deployed as independent services to accommodate the high-frequency invocation required by multi-turn agent–environment interaction. The system adopts a highly concurrent microservice architecture, where each tool is encapsulated as an HTTP service and organized into three functional layers: (1) a **FastAPI**-based interface layer that exposes asynchronous and batched RESTful endpoints; (2) a **Tool logic layer** that parses agent-issued tool-call instructions, retrieves the corresponding medical images from episode or trajectory metadata, and formats tool outputs into structured medical obser-

1414 vations; and (3) a **Ray Actor execution layer** that
1415 keeps model weights resident in GPU memory af-
1416 ter initialization, thereby avoiding repeated model
1417 loading under high-frequency tool invocations and
1418 significantly improving execution efficiency.

1419 **Asynchronous Tool-augmented Training** To sus-
1420 tain high throughput during RL rollouts, we em-
1421 ploy Ray to coordinate asynchronous execution
1422 between agents and tools. At each decision step,
1423 the policy first generates an explicit reasoning
1424 segment(between `<think>...</think>`), followed
1425 by a tool invocation (`<tool_call>...</tool_call>`).
1426 When the model emits the tool-call termina-
1427 tion token `</tool_call>`, decoding is temporarily
1428 paused, and the framework aggregates tool re-
1429 quests—containing trajectory identifiers and image
1430 paths—into batched HTTP calls. Ray manages re-
1431 quest queues and performs load balancing across
1432 different tool services. To improve resource effi-
1433 ciency, compute-intensive medical vision tools are
1434 pinned to dedicated GPUs, while lightweight utili-
1435 ties and knowledge-retrieval tools (e.g., DrugBank
1436 and PubMed) are multiplexed on shared CPU re-
1437 sources.

1438 **Extensible Tool Infrastructure** To facilitate rapid
1439 extension beyond the tools used in our experiments,
1440 MEDVISTAGYM provides a unified BaseTool ab-
1441 straction that enables *plug-and-play* integration of
1442 new medical perception or knowledge tools with
1443 minimal engineering overhead. This design sig-
1444 nificantly reduces the complexity of tool expan-
1445 sion and maintenance, supporting the continuous
1446 evolution of the environment’s capabilities. Sys-
1447 tem robustness is ensured through standardized
1448 health-check and monitoring endpoints (`/health`,
1449 `/metrics`), together with Ray’s automated failure
1450 recovery mechanism. In the event of tool or exe-
1451 cution failures, Ray transparently restarts affected
1452 actors without disrupting ongoing training, thereby
1453 maintaining training continuity and reliability.

1454 I Error Analysis Details

1455 I.1 Error Type Definitions

1456 We follow previous paper (Lu et al., 2025) that
1457 categorize tool-related failures into six error types
1458 based on inspection of model outputs and tool exe-
1459 cution traces. Note that a single sample may exhibit
1460 multiple error types.

- 1461 • **E1: Invocation schema violation.** The model
1462 produces malformed function calls that vio-
1463 late the expected tool invocation schema (e.g.,

missing required fields or incorrect call struc- 1464
1465 ture).

- **E2: Argument name error.** The model speci- 1466
1467 fies incorrect or non-existent parameter names
1468 when invoking tools.
- **E3: Argument format error.** The model 1469
1470 provides arguments with invalid formats or
1471 data types (e.g., malformed bounding boxes
1472 or invalid coordinate values).
- **E4: Argument content error.** The model 1473
1474 supplies semantically incorrect argument val-
1475 ues despite using valid formats (e.g., selecting
1476 irrelevant regions or incorrect anatomical tar-
1477 gets).
- **E5: Tool output format error.** The model 1478
1479 fails to correctly parse or utilize tool outputs
1480 due to malformed responses or misinterpreta-
1481 tion of returned results.
- **E6: Tool-induced reasoning error.** The 1482
1483 model invokes tools correctly but performs
1484 incorrect reasoning after tool execution, lead-
1485 ing to erroneous conclusions.

1486 I.2 Errors in Vanilla Open-Source VLMs

1487 Table 7 reports the distribution of error types
1488 observed in 100 error samples for GPT-5 and
1489 InternVL3-8B. For the vanilla InternVL3-8B
1490 model, the majority of failures stem from higher-
1491 level interaction and reasoning issues rather than
1492 low-level invocation syntax.

1493 In particular, E4 (argument content errors) and
1494 E6 (tool-induced reasoning errors) dominate the
1495 error distribution, accounting for 56.7% and 73.8%
1496 of the inspected cases, respectively. These errors
1497 indicate that the model often selects inappropriate
1498 tool arguments or fails to reason correctly over tool
1499 outputs, even when tool invocation is syntactically
1500 valid. In contrast, lower-level schema and argu-
1501 ment format errors (E1–E3) occur less frequently,
1502 suggesting that the model can partially learn tool
1503 syntax from prompting or supervised data alone.

1504 J Case Study Details

1505 J.1 Successful Cases

1506 **Case-1: Targeted Visual Evidence Seeking for**
1507 **Modality Identification** When a single global
1508 inspection of the image is insufficient to resolve
1509 the task, the model learns to actively search for

ID	Error Type	GPT5	InternVL3-8B
E1	Invocation schema violation (malformed function call)	2.8	11.7
E2	Argument name error (incorrect parameter name)	0	16.1
E3	Argument format error (invalid value format)	19.2	21
E4	Argument content error (semantically incorrect value)	22.1	56.7
E5	Tool output format error (malformed tool response)	31.9	15.5
E6	Tool-induced reasoning error (incorrect reasoning after tool execution)	11.6	73.8

Table 7: Error pattern identification and distribution from 100 error samples(%). Note that one case may contain multiple error types.

1510 diagnostic visual cues by invoking spatially tar-
1511 geted tools. As shown in Figure 10, the agent
1512 identifies that determining the imaging modality
1513 requires reading projection-specific markers that
1514 are not reliably visible at full scale. Rather than
1515 reasoning solely from global appearance, it delib-
1516 erately zooms into a corner region where modality
1517 indicators (e.g., side markers and projection labels)
1518 are typically located. The retrieved local evidence
1519 directly grounds the final answer. This behavior
1520 demonstrates an emergent ability to translate ab-
1521 stract task requirements into concrete visual search
1522 strategies, selecting both the appropriate tool and
1523 the relevant region to inspect.

1524 **Case-2: Post-Tool Reflection and Corrective**
1525 **Evidence Alignment** Beyond executing tools,
1526 the model exhibits an ability to critically evalu-
1527 ate whether tool outputs truly support the question
1528 being asked. In Figure 11, the agent initially relies
1529 on a detection tool to localize brown-stained cells.
1530 However, upon inspecting the tool output, it recog-
1531 nizes a semantic mismatch: scattered detections do
1532 not align with the question’s emphasis on tubular
1533 lining. This triggers a corrective action, prompting
1534 the agent to zoom into the tubular regions instead.
1535 The subsequent observation reveals continuous lum-
1536 inal staining, which resolves the ambiguity and
1537 leads to the correct conclusion. This pattern reflects
1538 post-tool self-reflection, where the agent does not
1539 passively accept tool outputs but instead assesses
1540 their relevance and consistency with the problem
1541 semantics.

1542 **Case-3: Coordinated Multi-Tool Reasoning un-**
1543 **der Structural Ambiguity** For complex cases,
1544 the model learns to orchestrate multiple comple-
1545 mentary tools, assigning each a distinct epistemic
1546 role and synthesizing their outputs into a unified
1547 judgment. As illustrated in Figure 12, the agent
1548 sequentially employs a global classifier (Biomed-
1549 CLIP), a segmentation model (BiomedParse), and
1550 a detector (GroundingDINO) to assess whether cor-
1551 tical gyri are abnormal. Rather than treating each
1552 tool independently or relying on any single result,

Aspect	Case-1	Case-2
Failure Cause	Insufficient Visual Evidence	Knowledge Gap
Tool Usage	✓ Correct (zoom-in)	✓ Correct (all 3 tools)
Reasoning Process	✓ Sound	✓ Sound
Tool Output	Normal but limited	Correct and informative
Root Problem	Ultra-low visual signal	Incorrect medical knowledge
Implication	Tools cannot overcome visual limits	Tools cannot fix knowledge deficits

Table 8: Comparison of two representative failure cases. Both cases exhibit correct tool usage and sound reasoning, but fail due to fundamentally different limitations: perceptual boundaries (Case-1, Figure 14) versus knowledge gaps (Case-2, Figure 15).

1553 the agent integrates signals across tools—global
1554 priors, structural segmentation, and localized de-
1555 tection confidence. The final decision is supported
1556 by convergent evidence rather than majority voting
1557 or isolated observations. This behavior demon-
1558 strates an emergent capacity for tool-role differen-
1559 tiation and coordinated reasoning, enabling robust
1560 decision-making in ambiguous scenarios.

1561 **Case-4: Robust Reasoning under Imperfect**
1562 **Tool Outputs** Importantly, agentic training also
1563 equips the model with robustness to tool failures
1564 and unreliable outputs. In Figure 13, both segmen-
1565 tation and detection tools fail to localize the target
1566 subcellular structures due to inherent limitations.
1567 Rather than propagating these errors, the agent ex-
1568 plicitly recognizes the mismatch between tool out-
1569 puts and the biological target. It then falls back
1570 on direct visual inspection and domain knowledge,
1571 reasoning from the complete absence of PMP70
1572 staining in knockdown cells to infer impaired per-
1573 oxisomal membrane assembly. This pattern high-
1574 lights an emergent ability to estimate tool reliabil-
1575 ity and selectively override tool guidance, ensuring
1576 that reasoning remains grounded even when tool
1577 outputs are noisy or misleading.

1578 Together, these patterns reveal a progression
1579 from correct tool invocation, to reflective tool inter-
1580 pretation, to coordinated multi-tool reasoning, and
1581 finally to robust reasoning beyond tool limitations.
1582 Rather than merely augmenting perception, tools in
1583 MEDVISTAGYM become integrated components
1584 of a higher-level reasoning process, enabling the
1585 model to actively seek evidence, critique interme-
1586 diate results, and adapt its strategy across multiple
1587 interaction turns.

1588 J.2 Failure Cases

1589 **Case-1: Insufficient Visual Evidence** We first
1590 analyze a representative failure case to examine
1591 the limits of the proposed agentic framework under
1592 extremely subtle visual conditions. As shown in
1593 Figure 14, the task is to determine whether the

1594 patient’s left lung exhibits abnormal findings. The
1595 image contains only weak and ambiguous visual
1596 cues, making a reliable diagnosis challenging even
1597 after localized inspection.

1598 The agent first correctly interprets the diagnostic
1599 goal and recognizes that a global inspection of the
1600 radiograph is insufficient. To reduce uncertainty,
1601 it deliberately invokes the zoom-in tool to focus
1602 on the left lung field, selecting an anatomically ap-
1603 propriate region of interest. This targeted action
1604 reflects an explicit attempt to acquire localized evi-
1605 dence relevant to pulmonary abnormality detection.

1606 After inspecting the magnified region, the agent
1607 systematically evaluates common radiographic in-
1608 dicators of abnormality, including focal opacity,
1609 parenchymal consolidation, and costophrenic an-
1610 gle blunting. Based on the absence of clearly dis-
1611 cernible pathological signs in the zoomed view, it
1612 concludes that the left lung is normal and outputs
1613 a negative result. This prediction is incorrect, as
1614 the ground truth indicates a subtle abnormality that
1615 remains visually indistinguishable at the examined
1616 scale.

1617 Although the final conclusion is incorrect, this
1618 failure does not arise from improper tool usage or
1619 unreflective reasoning. Instead, it highlights an
1620 inherent limitation imposed by ultra-low-signal vi-
1621 sual evidence. The trajectory demonstrates that
1622 the agent (i) localizes uncertainty, (ii) selects and
1623 applies an appropriate verification tool, and (iii)
1624 grounds its decision in explicit visual criteria. This
1625 case thus illustrates a *principled failure mode*,
1626 where the reasoning process remains coherent and
1627 introspective, yet the available visual evidence is
1628 insufficient to support a correct diagnosis.

1629 **Case-2: Knowledge Gap Despite Successful Tool**
1630 **Execution** We present a second failure case that
1631 reveals a complementary limitation of current tool-
1632 integrated medical reasoning systems, and more
1633 importantly, highlights a promising direction for
1634 extending medical tool collaboration within MED-
1635 VISTAGYM.

1636 As shown in Figure 15, the task requires identi-
1637 fying the eponymous fracture type depicted in a
1638 wrist radiograph, with candidate options including
1639 Monteggia, Bennett, Jones, and Smith fractures.
1640 Throughout its reasoning trajectory, the agent ex-
1641 hibits systematic and appropriate tool orchestration,
1642 demonstrating strong perceptual competence.

1643 In Turn 1, the agent invokes BiomedParse to seg-
1644 ment the fracture region, successfully isolating the

1645 relevant anatomical structures at the distal radius.
1646 Recognizing that subtle visual cues may be diag-
1647 nostically important, it proceeds in Turn 2 to apply
1648 the Agent4K super-resolution tool, enhancing the
1649 image by a factor of $4\times$ to better resolve cortical
1650 bone features and fracture morphology. In Turn 3,
1651 the agent further employs GroundingDINO to pre-
1652 cisely localize the fracture site, obtaining bounding
1653 box coordinates that correctly identify the distal
1654 radius region.

1655 Despite these methodologically sound and suc-
1656 cessful visual operations, the agent arrives at an
1657 incorrect diagnosis, predicting a Jones fracture in-
1658 stead of the correct Smith fracture. Importantly,
1659 this error does not stem from tool malfunction or
1660 insufficient visual evidence. Rather, it arises from
1661 an incorrect mapping between visual findings and
1662 medical concepts: the agent incorrectly associates
1663 the observed distal radius fracture with the defini-
1664 tion of a Jones fracture, which in fact refers to a
1665 fracture at the base of the fifth metatarsal, whereas
1666 a Smith fracture denotes a volar-displaced distal
1667 radius fracture.

1668 This case illustrates that while visual enhance-
1669 ment tools such as segmentation, super-resolution,
1670 and localization substantially improve perceptual
1671 evidence acquisition, accurate medical diagnosis in
1672 certain scenarios further requires access to special-
1673 ized domain-specific medical knowledge. Visual
1674 tools address how to perceive and localize relevant
1675 evidence, but the interpretation of that evidence re-
1676 lies on precise medical definitions and conceptual
1677 grounding.

1678 Crucially, this observation does not diminish the
1679 value of visual tools or the MEDVISTAGYM envi-
1680 ronment. Instead, it highlights their role as a nec-
1681 essary foundation for perception-centric reasoning,
1682 while motivating the integration of complemen-
1683 tary medical knowledge tools, such as structured
1684 diagnostic knowledge bases or task-specific med-
1685 ical reasoning modules. We view this case as a
1686 representative example of how increased tool di-
1687 versity and richer collaboration between visual and
1688 medical knowledge tools could further enhance
1689 VLM-based medical image analysis, and leave the
1690 expansion of the MEDVISTAGYM tool set in this
1691 direction as an important avenue for future work.

System Prompt: You are a medical VQA reasoning generator. Your task is to generate natural, step-by-step thinking processes for a tool-augmented reasoning trajectory.

Context: You are given a COMPLETED trajectory with:

- Question: the medical image question
- Ground Truth Answer: the verified correct answer
- Tool Execution Sequence: the actual sequence of tool_call → observation pairs
- Image: the medical image (if available)

Your Task: Generate realistic <think> content for EACH step, simulating how an agent would reason through the problem step-by-step.

Key Principle - Sequential Reasoning:

Each think must simulate forward-looking reasoning as if you DON'T know what tools will be called next:

1. *Before Tool 1:* Explain why you're calling THIS tool (not mentioning future tools)
2. *After Tool 1 Output:* Interpret the result, then decide what to do next
3. *Before Tool 2:* Based on Tool 1's output, explain why Tool 2 is now needed
4. *Final Verification:* Before the answer, summarize evidence and verify reasoning

Important Rules for Each Think:

Pre-tool thinking (before each tool_call):

- Explain the reasoning for calling THIS specific tool NOW
- Do NOT mention tools you haven't called yet
- Do NOT say "I will call tool1, then tool2, then tool3"

Post-observation thinking (after receiving tool output):

- Interpret what the tool returned
- If failed/irrelevant: explicitly acknowledge the failure
- Based on current evidence, decide the next action

Final verification thinking (before answer):

- Summarize all evidence gathered; Verify the reasoning chain

Rules:

- COPY tool_call, observation, and answer content EXACTLY
- ONLY generate the "think" content
- Each think should be 2-5 sentences, natural and concise
- Use first person ("I need to...", "Based on the result...")
- Handle tool failures gracefully

User Prompt Input:

- Question: {question}; Options: {options}
- Ground Truth Answer: {answer}
- Tool Execution Sequence: {tool_sequence}
- Image: {image}

Output Format (strict JSON, no markdown):

```
{"generated_trajectory": [{"step": 1, "type": "think", "content": "..."}, {"step": 1, "type": "tool_call", "tool": "...", "arguments": {...}, {"step": 1, "type": "observation", "tool": "...", "content": "..."}, ..., {"step": N, "type": "answer", "content": "..."}]}
```

Figure 6: Prompt Template for Reasoning Trajectory Generation.

System Prompt:

You are an expert medical visual agent equipped with specialized medical imaging tools. Your task is to answer medical visual questions by **using tools first, then reasoning based on tool observations**. The medical image alone is NOT reliable enough. Do NOT attempt to answer without validating information via tools. You must strictly adhere to the following protocol for every interaction:

1. ALWAYS call appropriate tools before giving any final answer;
2. You must choose and call tools to obtain observations, and ONLY THEN provide the final answer;
3. Answering without tool usage will be considered medically unsafe and incorrect.

{tools_description}

Instructions:

1. First, carefully analyze the image and question;
2. Decide which tool will provide the most useful information NEXT (you may use multiple tools across multiple steps);
3. Call ONE tool at a time, observe its output, and update your reasoning;
4. After each observation, either call another tool or, if you have enough information, provide the final answer;
5. Prefer using tools whenever they can improve reliability, localization, or clinical correctness.

Output Format:

You MUST use this EXACT format:

<think> Your reasoning and analysis here. Explain WHY you need a specific tool. </think>

Then EITHER call a tool:

```
<tool_call>{"name": "<tool_name>", "arguments": {"param": "value"}}</tool_call>
```

OR provide final answer:

```
<answer> Your final answer here </answer>
```

Important Rules:

- For the FIRST response to a new question, you MUST output a <tool_call> and MUST NOT output <answer>;
- Before producing <answer>, you must have successfully called AT LEAST ONE tool, unless ALL tools are clearly irrelevant or repeatedly fail;
- You can use MULTIPLE tools across multiple steps, but in EACH response you must choose EXACTLY ONE action: either a single <tool_call> OR a single <answer>;
- Never include both <tool_call> and <answer> in the same response;
- Maximum {max_tool_calls} tool calls allowed per question;
- Base your final answer on all tool observations and your step-by-step reasoning.

Figure 7: System Prompt.

Initial User Prompt:

Question: {question}

You MUST use one or more tools before answering. Start by thinking step by step and decide which tool to call first.

Continuation Prompt:

```
<observation> {observation} </observation>
```

```
<think> Interpret the observation and update your understanding of the medical question. Explain what the observation confirms, rules out, or suggests. Then decide whether the information is sufficient to answer the question. If not sufficient, determine which tool will provide the most useful next evidence and WHY — based strictly on this observation. </think>
```

If another tool is needed:

```
<tool_call>{"name": "<tool_name>", "arguments": {"param": "value"}}</tool_call>
```

If information is now sufficient to answer:

```
<answer> your final answer </answer>
```

Force Answer Prompt:

You have reached the maximum number of tool calls. Based on ALL tool observations gathered so far, you must now provide your final answer.

Output ONLY: <answer> your answer </answer>

Figure 8: User Prompt.

System Prompt:

You are an expert medical visual agent equipped with specialized medical imaging tools. Your task is to answer medical visual questions by **using tools first, then reasoning based on tool observations**. The medical image alone is NOT reliable enough. Do NOT attempt to answer without validating information via tools. You must strictly adhere to the following protocol for every interaction:

1. ALWAYS call appropriate tools before giving any final answer;
2. You must choose and call tools to obtain observations, and ONLY THEN provide the final answer;
3. Answering without tool usage will be considered medically unsafe and incorrect.

{tools_description}

Instructions:

1. First, carefully analyze the image and question;
2. Decide which tool will provide the most useful information NEXT (you may use multiple tools across multiple steps);
3. Call ONE tool at a time, observe its output, and update your reasoning;
4. After each observation, either call another tool or, if you have enough information, provide the final answer;
5. Prefer using tools whenever they can improve reliability, localization, or clinical correctness.

Output Format:

You MUST use this EXACT format:

<think> Your reasoning and analysis here. Explain WHY you need a specific tool. </think>

Then EITHER call a tool:

```
<tool_call>{"name": "tool_name", "arguments": {"param": "value"}}</tool_call>
```

OR provide final answer:

<answer> Your final answer here </answer>

Important Rules:

- For the FIRST response to a new question, you MUST output a <tool_call> and MUST NOT output <answer>;
- Before producing <answer>, you must have successfully called AT LEAST ONE tool, unless ALL tools are clearly irrelevant or repeatedly fail; in that case, explain this in <think> and then answer;
- You can use MULTIPLE tools across multiple steps, but in EACH response you must choose EXACTLY ONE action: either a single <tool_call> OR a single <answer>;
- Never include both <tool_call> and <answer> in the same response;
- Maximum {max_tool_calls} tool calls allowed per question;
- Base your final answer on all tool observations and your step-by-step reasoning.

Figure 9: GPT5 Data Curation.

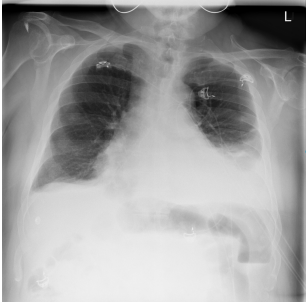

Image-1

(Original)

Image-2

(Zoomed Corner)

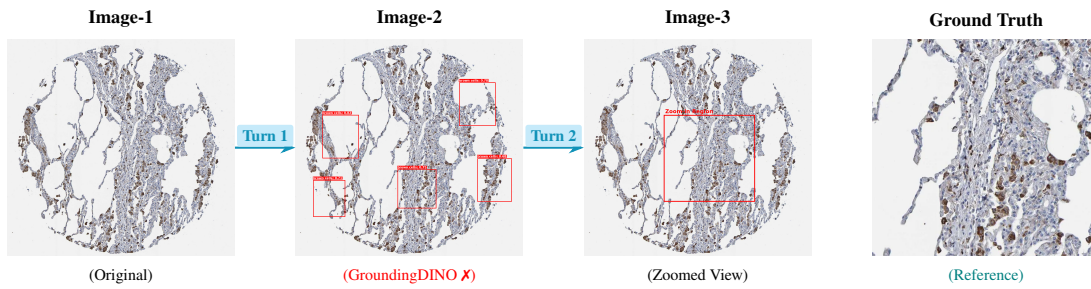
Turn 1

Question: What type of medical image is this?
Options: A) MRI Scan B) X-Ray C) CT Scan D) PET-CT
Ground Truth: B) X-Ray

Turn 1: <think>To determine the imaging modality, I should look for modality-specific cues such as projection markers, windowing, and hardware. The top-right corner often contains a side marker (e.g., L) and projection text. Zooming there will help me read the label and confirm the modality.</think>
<tool_call>["image_zoom_in", "bbox_2d": [0.75, 0.0, 0.98, 0.25]]</tool_call> → Zoomed Image

Turn 2: Visual Confirmation
<think>The zoom succeeded, allowing inspection of the top-right corner. I can see a projection label and a side marker (L), which are classic for a chest radiograph. The image shows a projection view with ribs and clavicles, not cross-sectional slices, and lacks CT windowing or PET color overlay. These features confirm the modality is an X-ray.</think>
<answer>B. X-Ray ✓</answer> (Correct: Zoom tool effectively revealed diagnostic marker)

Figure 10: **Effective Tool Usage.** The agent identifies the imaging modality by strategically zooming into the corner region (Turn 1) to examine the projection marker “L”, which is characteristic of chest X-ray images. The tool provides decisive evidence for the correct answer.

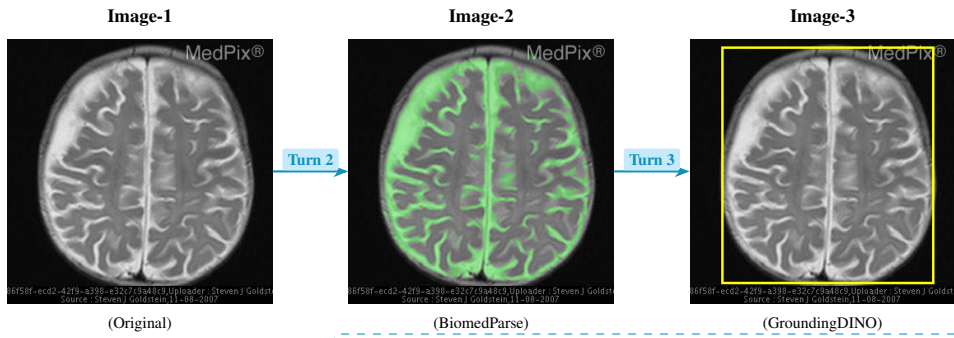


Question: In hepatic tissue stained with hematoxylin (blue) and DAB (brown), the antibody binds to components of tubular lining. Which cell type is the protein principally expressed in?
Options: A) Macrophages B) Fibroblasts C) Endothelial cells D) Epithelial cells E) Pericytes
Ground Truth: C) Endothelial cells

Turn 1: <think>The question asks which cell type the protein is expressed in. The DAB staining marks the target protein. I will use GroundingDINO to detect and localize the brown-stained cell clusters.</think> <tool_call>{"GroundingDINO", "query": "brown stained cells"}</tool_call> → Detected scattered brown clusters
Turn 2: <think>GroundingDINO detected scattered brown clusters. However, the question states the antibody binds to tubular lining components. Scattered cells do not line tubular structures. This suggests a misclassification. I need to zoom into the tubular region to verify.</think> <tool_call>{"image_zoom_in", "bbox_2d": [0.35, 0.35, 0.75, 0.75]}</tool_call>

Turn 3: Self-Reflection & Correction
<think>In the zoomed view, the brown DAB signal is concentrated along the luminal surface of tubular structures, forming a continuous lining pattern rather than scattered clusters. This is inconsistent with macrophages. The continuous luminal lining is characteristic of endothelial cells. Therefore, GroundingDINO's detection was misleading, and the protein is principally expressed in endothelial cells.</think>
<answer>C. Endothelial cells ✓</answer> (Correct: Agent identified tool misclassification through self-reflection)

Figure 11: Self-Reflection and Correction. The agent initially uses GroundingDINO to detect scattered brown-stained cells (Turn 1), but identifies a mismatch with the question’s description of “tubular lining” (Turn 2). Through self-correction, it zooms into the tubular region and confirms the continuous luminal staining pattern characteristic of endothelial cells.

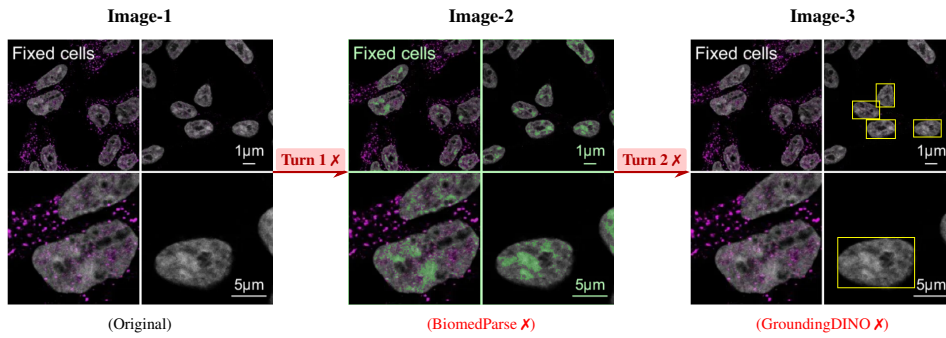


Question: Is the anatomy of the brain gyri affected?
Options: A) Yes B) No
Ground Truth: B) No

Turn 1: <think>I will use a zero-shot classifier to screen for global cortical abnormalities.</think> <tool_call>{"BiomedCLIP"}</tool_call> → normal gyri (56%), abnormal gyri (44%). Moderate confidence warrants further investigation.
Turn 2: <think>I will segment cortical structures to reveal localized abnormalities.</think> <tool_call>{"BiomedParse", "prompt": "cortex"}</tool_call> → Detected cortical structures, no abnormal region highlighted.
Turn 3: <think>I will use detection to localize any abnormal gyral patterns.</think> <tool_call>{"GroundingDINO", "query": "abnormal gyri"}</tool_call> → 1 detection, confidence 21.7% (very low). No convincing evidence.

Turn 4: Multi-tool Synthesis
<think>After comprehensive multi-tool analysis: (1) classifier favors normal gyri, (2) segmentation revealed no abnormal cortical regions, (3) detector found no high-confidence abnormality. The convergent evidence from all three tools supports the conclusion that the gyral anatomy is not affected.</think>
<answer>B. No ✓</answer> (Correct: Multi-tool convergent evidence supports normal gyri anatomy)

Figure 12: Multi-tool Reasoning for Comprehensive Assessment. The agent uses three complementary tools: BiomedCLIP for classification (Turn 1), BiomedParse for segmentation (Turn 2), and GroundingDINO for detection (Turn 3). The convergent evidence from all tools supports the correct conclusion that gyral anatomy is not affected.

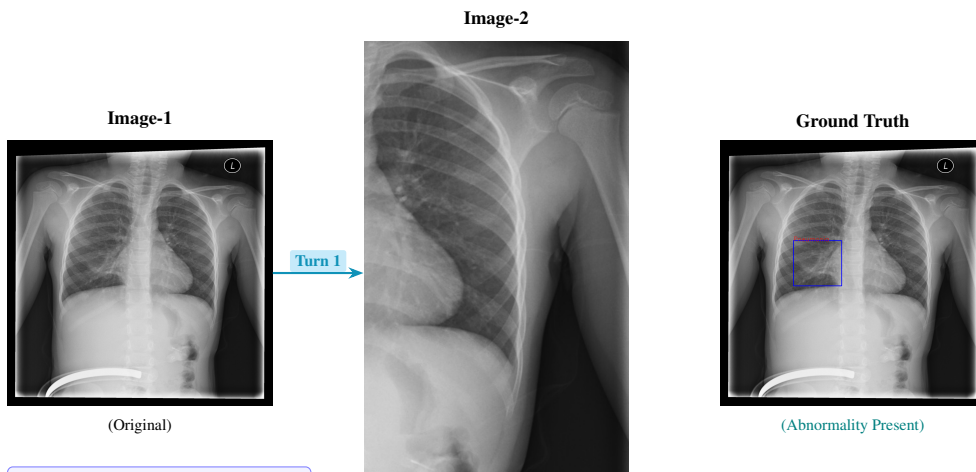


Question: PEX19 knockdown cells lack PMP70 staining (magenta) while wildtype shows it. What is the most likely explanation?
Options: A) Misfolding B) Transport disruption C) Synthesis affected D) Lipid raft attachment E) Peroxisomal membrane assembly impaired
Ground Truth: E

Turn 1: <think>I will segment peroxisomal structures to assess if they are present.</think> <tool_call>{"BiomedParse", "prompt": "peroxisome"}</tool_call>
Result: Segmented nuclei (green) instead of peroxisomes. Tool limitation: not trained for subcellular organelles.
Turn 2: <think>Let me try to localize peroxisomes using text-guided detection.</think> <tool_call>{"GroundingDINO", "query": "peroxisome"}</tool_call>
Result: Detected cells/nuclei (yellow boxes), not peroxisomes. Detection failed: boxes on wrong structures.

Turn 3: Visual Observation + Domain Knowledge
<think>Both tools failed to detect peroxisomes correctly. I should rely on direct visual observation and biological knowledge.
Visual observation: Left panels (WT) show magenta PMP70 punctate staining; right panels (KD) completely lack this staining.
Biological reasoning: PEX19 is essential for peroxisomal membrane assembly. Without PEX19, peroxisomes cannot form.</think>
<answer>E. Peroxisomal membrane assembly impaired ✓</answer> (Correct: Model recovered from tool failures using domain knowledge)

Figure 13: **Robustness to Tool Failure.** BiomedParse incorrectly segments nuclei instead of peroxisomes (Turn 1); GroundingDINO detects cells rather than subcellular organelles (Turn 2). Despite both tool failures, the agent correctly answers by combining visual observation (WT shows punctate PMP70 staining while KD shows complete absence) with domain knowledge (PEX19 is essential for peroxisomal membrane biogenesis).

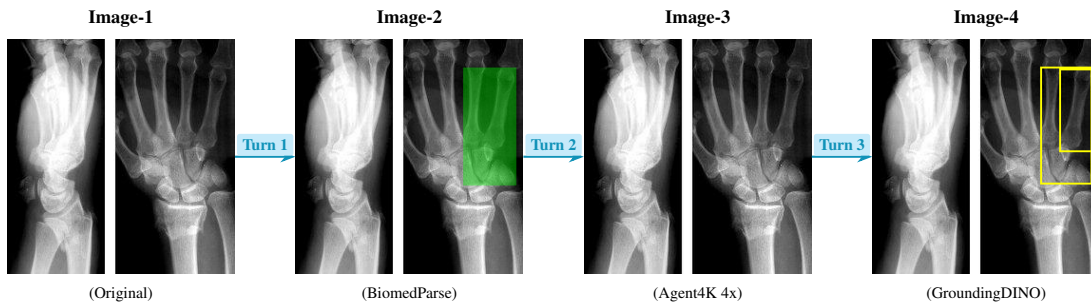


Question: Is there any abnormality in the patient's left lung?
Options: A) Yes B) No
Ground Truth: A) Yes

Turn 1: <think>To decide if the left lung is abnormal, I need a closer look at the left hemithorax for any focal opacity, consolidation, or pleural effusion. A targeted zoom on the left lung field will help me assess the parenchyma and pleural space more precisely.</think>
<tool_call>{"image_zoom_in", "bbox_2d": [0.55, 0.15, 0.95, 0.85]}</tool_call> → Zoomed Image

Turn 2: <think>With the left lung magnified, I can evaluate for any focal opacity, air bronchograms, or a meniscus sign that would suggest effusion. The parenchyma appears clear without a discrete mass or consolidation, and the costophrenic angle looks sharp without blunting. Given these findings, I can conclude the left lung is not abnormal.</think>
<answer>B. No ✗</answer> (Correct: A. Yes — Abnormality present in left lung field)

Figure 14: **Failure Case: Incorrect Visual Interpretation Despite Tool Usage.** The agent correctly uses zoom (Turn 1) to examine the left lung field, but fails to identify the abnormality present (compare with Ground Truth). This illustrates that tool usage alone does not guarantee correct diagnosis—accurate visual interpretation remains critical.



Question: What person's name is associated with the fracture shown below?
 Options: A) Monteggia B) Bennett C) Jones D) Smith
Ground Truth: D) Smith

Turn 1: <think>To identify the fracture type, I will use a segmentation tool to highlight the fracture region.</think>
 <tool_call>{"BiomedParse", "prompt": "fracture"}</tool_call>
Turn 2: <think>Segmentation highlights fracture at distal radius. Enhancing resolution for better visualization.</think>
 <tool_call>{"Agent4K", "scale": 4}</tool_call>
Turn 3: <think>Using object detection to localize fracture location.</think> <tool_call>{"GroundingDINO", "query": "fracture"}</tool_call> → Detected at [177,29,231,152]

Turn 4: <think>Fractures located at distal radius. Jones fracture involves distal radius with capitate dislocation; Smith fracture involves distal radius with capitate fracture. Most likely fracture type is Jones.</think>
 <answer>C. Jones </answer> (Correct: D. Smith — Smith fracture is a volar-displaced distal radius fracture)

Figure 15: **Hard Case: Tool Success with Knowledge Gap.** All three tools (BiomedParse, Agent4K, GroundingDINO) executed successfully and provided correct visual evidence. However, the model's incorrect medical knowledge about fracture type definitions led to the wrong final answer.