

# ACTOR: Agentic CT Reasoning with Radiologist-Aligned Actions

Xingyu Qian<sup>1,5,6,\*</sup> Yu Sun<sup>2,\*</sup> Songyou Li<sup>1,5,6</sup> Zongzhao Li<sup>1,5,6</sup> Jiacheng Cen<sup>1,5,6</sup>  
Wenqian Wu<sup>3,7</sup> Tingyang Xu<sup>4,8</sup> Yu Rong<sup>4,8</sup> Deli Zhao<sup>4</sup> Wenbing Huang<sup>1,5,6,†</sup>

<sup>1</sup>Gaoling School of Artificial Intelligence, Renmin University of China

<sup>2</sup>The Chinese University of Hong Kong, Shenzhen

<sup>3</sup>Shanghai Jiao Tong University School of Medicine

<sup>4</sup>Alibaba DAMO Academy

<sup>5</sup>Beijing Key Laboratory of Research on Large Models and Intelligent Governance

<sup>6</sup>Engineering Research Center of Next-Generation Intelligent Search and Recommendation, MOE

<sup>7</sup>Shanghai General Hospital <sup>8</sup>Hupan Lab

## Abstract

Radiologists read CT scans as a sequential decision process, actively scrolling, zooming, adjusting windows, comparing with prior scans, and measuring structures, where each observation guides the next action. Current VLMs instead answer from passively sampled views, missing the multi-step reasoning that clinical reading demands. We address this by aligning the agent’s action space to radiologists’ operations within the Picture Archiving and Communication System (PACS), enabling active, multi-step CT reading with inherently transparent reasoning for clinical decision support. In particular, we build **ACTOR**, a unified framework comprising: (1) a PACS-like interactive CT environment with five radiologist-aligned tools; (2) **ActiveCTR-Bench**, spanning 7 tasks across a Perception–Interpretation–Reasoning hierarchy (7,112 items with visual ground truth and verified vision necessity); and (3) **ActiveCTRReasoner**, a 7B model trained on 108K synthesized trajectories via SFT and reinforcement learning. ActiveCTRReasoner achieves 73.2% close-ended accuracy, substantially outperforming GPT-5.4 (47.3%). In open-ended evaluations, it scores 44.2% compared to 23.8% for the best baseline. A radiologist collaboration study provides preliminary evidence for the decision-support utility of radiologist-aligned agentic reasoning.

## 1 Introduction

Clinical interpretation of CT scans is fundamentally an active, sequential exploration process. Radiologists do not passively observe static images. Instead, they engage in a dynamic workflow: scrolling through hundreds of slices to locate lesions, switching window presets to differentiate

tissues, zooming in to confirm morphological details, retrieving prior scans for temporal comparison, and calculating precise geometric changes of nodule growth. This reading workflow is inherently a *sequential decision process*, where initial observations shape subsequent findings, which in turn dictate the next steps. An agentic system should mirror this sequential decision-making process to achieve robust and context-aware CT reasoning. However, current VLM-based systems (Li et al., 2023; Bai et al., 2024; Chen et al., 2024a; Pan et al., 2025) rely on a limited set of passively sampled views to answer CT queries, a paradigm that fundamentally conflicts with the active, multi-step nature of clinical reading.

Recent agentic approaches (Mao et al., 2026; Gu et al., 2026; Orlando et al., 2026) take an important step by introducing tool-calling into CT question answering. Yet their tools are task-specific rather than aligned with standard clinical operations; moreover, the agentic behaviors are merely modulated via prompt engineering where initial observations guide subsequent exploration. On the evaluation front, existing benchmarks (Bai et al., 2024; Zhu et al., 2026) rely heavily on report text to derive ground truth, thereby restricting assessment to perceptual tasks and failing to assess cognitive reasoning.

We ground our approach in radiologist alignment, ensuring that every model action corresponds to standard Picture Archiving and Communication System (PACS) operations in routine clinical workflows. This alignment ensures that the agent reads CT volumes through the same active, multi-step workflow as clinical practice, and that its reasoning trajectory is inherently transparent—each intermediate state is a rendered image readable by clinicians.

To this end, we propose Agentic CT reasOn-

\*Equal contribution.

†Corresponding author.

Email: hwenbing@ruc.edu.cn.

ing with Radiologist-aligned actions (**ACTOR**)<sup>1</sup>, a unified framework containing environment, benchmark, and training. Specifically, we build a PACS-like interactive CT environment equipped with five radiologist-aligned tools. Inside this environment, we introduce **ActiveCTR-Bench**, a benchmark spanning 7 tasks across a Perception–Interpretation–Reasoning cognitive hierarchy, featuring 7,112 vision-necessary items paired with rich visual ground truth. Leveraging this infrastructure, we train **ActiveCTReasoner**, a 7B VLM that internalizes agentic reasoning via 108K trajectory Supervised Fine-Tuning (SFT) followed by GRPO reinforcement learning.

To sum up, our contributions are as follows:

- **Radiologist-aligned agentic paradigm.** A PACS-like interactive environment where every agent action mirrors a standard radiological operation, unifying benchmark construction, model training, and clinical decision support through inherent transparency.
- **Co-designed benchmark and training.** ActiveCTR-Bench, the first 3D CT benchmark with verified vision necessity, visual ground truth, and P-I-R cognitive coverage (7 tasks, 7,112 items); paired with 108K trajectory SFT and GRPO that internalize agentic reasoning.
- **Comprehensive empirical and clinical validation.** Extensive evaluation across 21 models, supported by rigorous ablation and RL analyses, demonstrates that ActiveCTReasoner achieves 73.2% close-ended accuracy (vs. 47.3% for GPT-5.4) and a 1.9× gain in open-ended LLM-Judge scores over the leading baseline. Furthermore, a radiologist collaboration study provides preliminary evidence that transparent agent outputs can support readers under the study protocol, improving answer quality while reducing reading time.

## 2 Related Work

**Medical Image VQA Benchmarks.** 3D medical VQA has advanced through large-scale datasets (Bai et al., 2024; Yu et al., 2025), hierarchical CT evaluation (Zhu et al., 2026), and

tool-calling benchmarks (Gu et al., 2026). Notably, RadImageNet-VQA (Butsanets et al., 2025) introduced blind-test vision necessity verification. However, existing benchmarks predominantly derive ground truth from report text and focus on perception-level tasks. ActiveCTR-Bench complements these with visual ground truth, P-I-R cognitive coverage, and longitudinal temporal support (detailed comparison in § 3.3).

**Agentic VLM.** Tool-augmented agents (Schick et al., 2023; Yao et al., 2022; Liu et al., 2025) have been adopted in medical imaging for chest X-ray interpretation (Chen et al., 2024b), multimodal medical reasoning (Li et al., 2024; Fallahpour et al., 2025), and 3D CT analysis. In the CT domain, CT-Agent (Mao et al., 2026) decomposes QA through anatomy-specific tools, CT-Flow (Gu et al., 2026) implements tool-chain orchestration via MCP, and MedScribe (Orlando et al., 2026) acquires evidence through hypothesis-driven reasoning. However, these works primarily leverage existing VLM capabilities through prompting or lightweight instruction tuning, without internalizing agentic reasoning via large-scale trajectory training and RL. Moreover, their tool designs are task-specific rather than aligned with the radiologists’ standard PACS operations.

**RL-based Multimodal Reasoning.** RL has progressed from RLHF alignment to reasoning-chain optimization (Guo et al., 2025) and is extending to medical VLMs: MedVLM-R1 (Pan et al., 2025) and Med-R1 (Lai et al., 2026) apply GRPO for medical reasoning—all single-turn without tool interaction. In tool-augmented RL, Ophiuchus (Jiang et al., 2025b) teaches models to call tools within reasoning chains, and Med-VistaGym (Lu et al., 2026) provides a scalable RL environment—but both target 2D images. To our knowledge, no prior work has conducted multi-turn tool-augmented RL on 3D medical volumes.

## 3 ACTOR

We present **ACTOR**, a unified framework spanning environment, data, and training, as illustrated in Fig. 1. Every model action is aligned to a radiologist’s standard PACS operation, and every intermediate state is rendered as a human-readable image, ensuring inherently transparent reasoning.

<sup>1</sup>Code will be released at <https://github.com/qianxingyucode/ACTOR>.

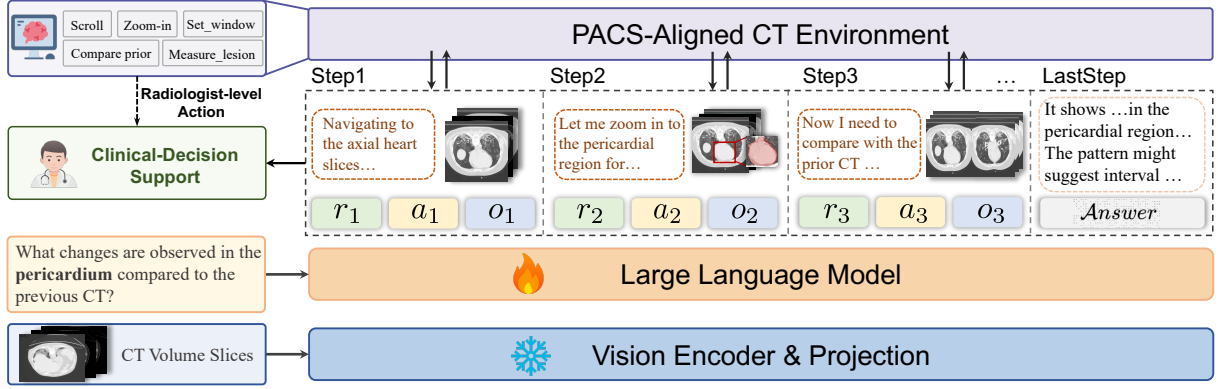


Figure 1: ACTOR framework. At each step  $t$ , the agent produces reasoning  $r_t$ , selects action  $a_t$  from five radiologist-aligned tools, and receives observation  $o_t$  (a rendered 2D image) from the PACS-like CT environment. The loop continues until the agent issues a final answer.

### 3.1 PACS-Aligned CT Environment

We build a PACS-aligned CT simulator as the interactive environment. It wraps 3D CT volumes, tracks the current slice, zoom region, and window preset, and renders 2D observations after each tool call. Each *window preset* is specified by a width-level pair, i.e., width ( $W$ ) and level ( $L$ ), mapping raw Hounsfield Units to display intensities. For example, the lung window ( $W = 1500, L = -600$ ) highlights airspace, whereas the mediastinum window ( $W = 350, L = 50$ ) emphasizes soft tissue. For longitudinal tasks, the simulator jointly loads current and prior scans.

**Problem Formalization.** Given a CT volume  $V \in \mathbb{R}^{D \times H \times W}$ , where  $D$  denotes the slice dimension and  $H \times W$  specifies the spatial resolution of each slice, the agent performs sequential decision-making conditioned on the task query  $q$ .

▷ **State:** At time step  $t$ , the agent state is defined as  $s_t = (z_t, b_t, w_t)$ , where  $z_t \subseteq \{1, \dots, D\}$  denotes the set of slice indices selected at time step  $t$ , while  $b_t$  and  $w_t$  denote the bounding box and radiological window preset shared across these selected slices, respectively.

▷ **Action space:** The action space is defined as  $\mathcal{A} := \mathcal{A}_{\text{tool}} \cup \mathcal{A}_{\text{terminal}}$ , where  $\mathcal{A}_{\text{tool}}$  includes five radiological tools: *scroll*, *zoom in*, *set window*, *compare with prior*, and *measure lesion*. Under this design,  $\mathcal{A}_{\text{terminal}}$  contains only one action that ends the episode, and produces the final response.

▷ **Observation:** After executing action  $a_t$ , the simulator returns an observation  $o_t$ , either a 2D image for navigation or comparison, or structured quantitative output for measurement, from which the next state is updated as  $s_{t+1} = f_{\text{state}}(s_t, o_t)$ .

▷ **Policy:** The VLM-based policy  $\pi_\theta$  :

Table 1: Mapping of tools to PACS operations and their corresponding outputs. Output of each tool is either a rendered 2D image or structured data.

| Tool                            | PACS Operation      | Output            |
|---------------------------------|---------------------|-------------------|
| <code>scroll(z)</code>          | Scroll to slice     | Axial view at $z$ |
| <code>zoom_in(region)</code>    | Magnify ROI         | Cropped view      |
| <code>set_window(preset)</code> | Adjust window/level | Re-rendered image |
| <code>compare_prior(z)</code>   | Retrieve prior      | Side-by-side pair |
| <code>measure_lesion(t)</code>  | Measurement tool    | Structured data   |

$(q, s_t, h_t) \mapsto a_t$  maps the task query, current state, and interaction history to the next action, where  $h_t = \{a_i, o_i\}_{i=0}^{t-1}$  records previous actions and observations. This process induces a trajectory  $\tau = \{r_i, a_i, o_i\}_{i=1}^T$ , in which the model generates reasoning  $r_t$ , selects an action  $a_t \in \mathcal{A}$ , and receives simulator feedback  $o_t = \text{Sim}(V, s_t, a_t)$ . The episode ends when  $a_T \in \mathcal{A}_{\text{terminal}}$ , producing the final response  $y$  evaluated against  $y^*$  via a reward function  $R(y, y^*)$ .

**Tool Definition.** We design five tools that directly mirror the radiologists' standard PACS operations, as summarized in Tab. 1. Since all intermediate outputs are rendered 2D images, every reasoning step is human-readable and auditable, enabling radiologists to directly review agent reasoning as clinical decision support.

### 3.2 Data Pipeline

**Raw Data Collection.** We collect data from two complementary sources, as shown in Step 1 of Fig. 2. *CT-RATE* (Hamamci et al., 2024) is one of the largest chest CT datasets, comprising 21,304 patients, 25,692 volumes, and approximately 8.9M slices, with paired reports and abnormality labels. *TotalSegmentator* (Wasserthal et al., 2023) extracts anatomy masks for 104 structures

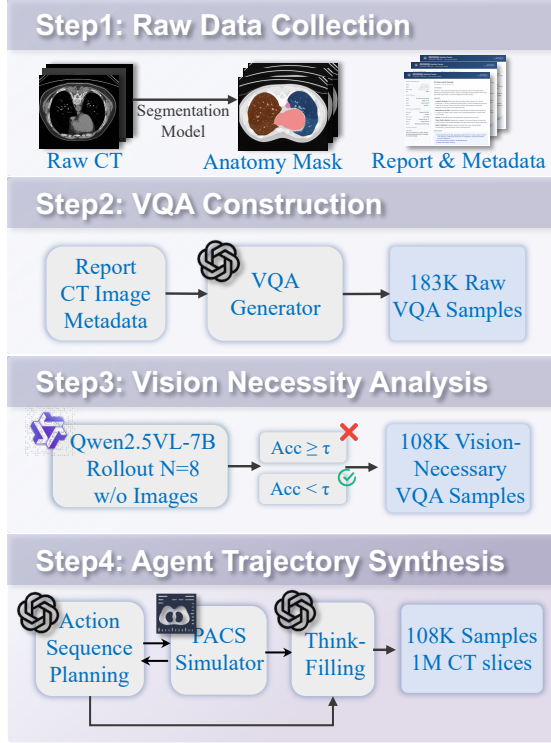


Figure 2: Data pipeline overview. Raw CTs are processed with TotalSegmentator to obtain anatomy masks, from which a VQA generator produces 183K samples. Vision necessity filtering retains 108K samples, which are then converted into multi-turn agentic trajectories (108K trajectories, 1M rendered slices) via action planning, PACS rendering, and think-filling.

from raw CTs, providing visual ground truth such as lesion volumes, lobe masks, and organ-level changes. We further construct longitudinal scan pairs from the same patient at different time points, yielding 4,125 training pairs and 261 validation pairs with a median interval of 74 days. These pairs support temporal tasks that require comparison between current and prior scans.

**VQA Construction.** Using the collected data, we build a VQA generator that takes anatomy masks, report text, and metadata as input, producing 183K raw VQA samples across all seven tasks, as shown in Step 2 of Fig. 2. For each sample, the LLM is provided with structured mask-derived features, including lesion volumes, lobe distributions, and longitudinal changes, together with report context, and then generates both the question and the corresponding answer. *Quality control* is performed by cross-checking all generated answers against rule-based ground truth deterministically computed from mask voxels, such as lesion volume from segmentation and lobe membership from spatial coordinates. Samples that fail these

consistency checks are discarded.

**Vision Necessity Filtering.** To ensure genuine visual reasoning, as shown in Step 3 of Fig. 2, we run Qwen2.5-VL-7B in text-only mode with 8 rollouts and no images, removing samples answerable without visual evidence. Questions are first rewritten with visual-anchoring prefixes to reduce text leakage; samples exceeding task-specific text-only accuracy thresholds are then discarded, with  $\tau = 0.6$  for 4-way MCQ and  $\tau = 0.7$  for 3-way MCQ and Yes/No questions. Open-ended tasks are retained because their ground truth contains mask-derived quantitative information. This yields 108K vision-necessary VQA samples.

**Trajectory Synthesis.** Vision-necessary VQA samples are converted into multi-turn agentic trajectories in three steps, as shown in Step 4 of Fig. 2. First, GPT-5.4 plans an interaction skeleton with the tool sequence and target observations. Second, the PACS simulator executes each tool call and renders the corresponding CT views, with TotalSegmentator providing volumetric measurements for `measure_lesion`. Third, an LLM fills in turn-level reasoning traces within `<think>...</think>`. This process produces 108,085 complete trajectories, averaging 14.5 turns and covering 1M rendered CT slices.

**RL Data Curation.** To construct RL training data, we roll out the SFT model on training samples and retain moderately difficult samples with `pass_rate`  $\in [0.25, 0.625]$ , where the model succeeds intermittently but not consistently. This selection yields 12,011 samples for policy optimization.

### 3.3 ActiveCTR-Bench

**P-I-R Cognitive Framework.** We propose a three-level framework that models the radiologists reading process, with task examples shown in Fig. 3 and statistics detailed in Tab. A.1: *Perception*, attending to relevant regions (lesion detection, anatomical localization); *Interpretation*, translating visual evidence into clinical judgments (measurement grading, change classification, region-level reporting); *Reasoning*, integrating multi-step clinical inference across temporal intervals and imaging changes.

**Design Criteria.** ActiveCTR-Bench follows five criteria: (1) *Vision Necessity*, where removing

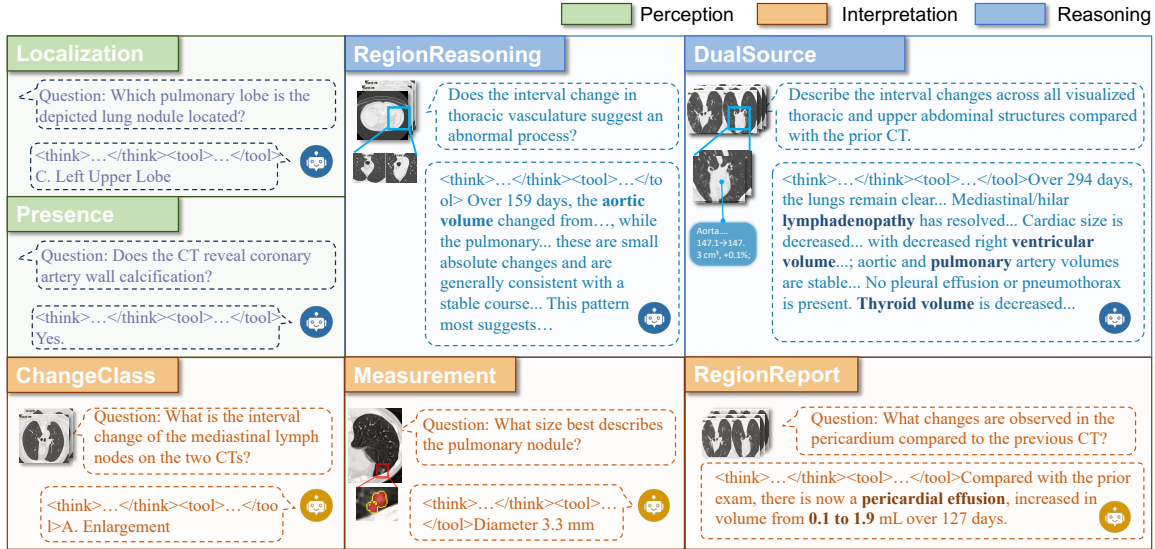


Figure 3: ActiveCTR-Bench task overview. Each card shows a representative question with the agent’s multi-turn agentic trajectory (think → tool call → answer).

Table 2: Benchmark comparison across five dimensions: blind-test verification (Blind), visual ground truth (Vis. GT), cognitive levels (Cog.), temporal reasoning (Temp.), and agentic evaluation (Agent.). Symbols ✓, ✗, and ✗ indicate full support, partial support, and absence, respectively.

|              | Blind | Vis. GT | Cog.         | Temp. | Agent. |
|--------------|-------|---------|--------------|-------|--------|
| M3D-VQA      | ✗     | ✗       | P            | ✗     | ✗      |
| 3D-RAD       | ✗     | ✗       | P            | ✗     | ✗      |
| CT-Bench     | ✗     | ✗       | P            | ✗     | ✗      |
| RadImNet-VQA | ✓     | ✓       | P            | ✗     | ✗      |
| MedFrameQA   | ✗     | ✗       | P            | ✗     | ✗      |
| CT-FlowBench | ✗     | ✗       | P            | ✗     | ✓      |
| <b>Ours</b>  | ✓     | ✓       | <b>P-I-R</b> | ✓     | ✓      |

images reduces accuracy to near chance; (2) *Visual Ground Truth*, where answers come from image tools rather than report text; (3) *Non-Trivial Reasoning*, requiring at least one intermediate step; (4) *Clinical Validity*, ensuring genuine radiological judgment; and (5) *Evaluability*, with unique MCQ answers and structured rubrics for open-ended questions. Tab. 2 compares ActiveCTR-Bench with existing benchmarks along these axes.

**Evaluation Protocol.** For *closed-ended tasks*, we report accuracy based on exact answer-letter matching. For *open-ended tasks*, we adopt LLM-J, an LLM-as-Judge metric, as the primary evaluation. GPT-5.5 scores each prediction against the ground truth along five rubric dimensions: finding coverage, anatomical and clinical precision, quantitative accuracy, change and temporal reasoning, and faithfulness. Each dimension is scored from

0 to 2, yielding a total score from 0 to 10 that is linearly scaled to 0 to 100, as detailed in Appendix D.4. For completeness, we provide the definitions of four auxiliary N-gram-based metrics in the appendix: ROUGE-L, Token-F1, Numerical Recall, and Direction Match.

## 4 Training of ActiveCTReasoner

ActiveCTReasoner is built upon the Qwen2.5-VL-7B architecture and trained via a two-stage paradigm: while SFT instructs the model what tools to invoke and how to format interactions, RL optimizes how to strategically deploy them. This progressive training mirrors the professional trajectory of a radiologist advancing from a novice trainee mastering basic PACS operations to an attending physician strategically gathering evidence to formulate definitive clinical decisions.

### 4.1 Supervised Fine-Tuning

SFT teaches the base model fundamental tool-calling patterns for interacting with the PACS-aligned CT environment (§ 3.1): when to invoke each tool, how to format arguments, and how to interpret rendered observations. We fine-tune Qwen2.5-VL-7B on the 108K trajectories (§ 3.2), freezing the vision tower and projector. The training objective is causal LM with response-only loss: cross-entropy is computed exclusively over response tokens (reasoning, tool calls, and answers) while prompt tokens are masked. After SFT, the model reliably produces well-formed tool calls (99.8% normal termination), establishing the

behavioral foundation for RL to optimize tool-use strategies. Configuration in Tab. A.4.

## 4.2 Agentic Reinforcement Learning

Building on the SFT model, RL further optimizes *tool-use strategies* by encouraging the model to actively gather sufficient visual evidence and incorporate observations into clinical reasoning, similar to how radiologists decide which views to inspect and how many measurements to perform. We adopt GRPO (Shao et al., 2024) for this stage, which performs policy optimization through group-relative advantage estimation without requiring a critic model. This makes it well suited for multi-turn agentic trajectories where rewards are only available at trajectory termination. During RL, the model interacts with the environment (§ 3.1): each rollout yields a complete trajectory, and the final answer is scored against the ground truth. Training uses the 12,011 RL samples curated by pass-rate filtering (§ 3.2).

**Reward design.** We jointly train the model on all seven tasks with task-specific rewards. For *closed-ended* tasks, we define the multiple-choice reward as follows:

$$R_{MCQ} = \mathbb{I}[\text{correct}] - 0.1 \cdot \mathbb{I}[\text{invalid}], \quad (1)$$

where  $\mathbb{I}[\cdot]$  is the indicator function, returning 1 when the condition holds and 0 otherwise. Invalid responses refer to outputs that do not contain a valid answer option. For *open-ended* tasks, we use a composite reward to emphasize both textual quality and quantitative correctness:

$$R_{\text{open}} = \sum_{m \in \mathcal{M}} w_m R_m, \quad (2)$$

where  $\mathcal{M} = \{\text{RL}, \text{NR}, \text{F1}, \text{DM}\}$  denotes four reward metrics: ROUGE-L, Numerical Recall, Token-F1, and Direction Match. We use weights  $(w_{\text{RL}}, w_{\text{NR}}, w_{\text{F1}}, w_{\text{DM}}) = (0.3, 0.3, 0.2, 0.2)$  to measure structural completeness, quantitative accuracy, lexical coverage, and change-direction correctness. The quantitative terms  $R_{\text{NR}}$  and  $R_{\text{DM}}$  receive a total weight of 0.5, discouraging fluent but numerically inaccurate responses. We report the training configuration in Tab. A.4.

## 5 Experiments

In this section, we first describe the experimental setup in § 5.1, then evaluate ActiveCTR-Bench from four perspectives: overall perfor-

mance in § 5.2, tool-calling behavior in § 5.3, ablation settings in § 5.4, and radiologist collaboration in § 5.5.

### 5.1 Setup

**Models.** We evaluate **21 external baseline models**: 5 proprietary models, 5 general-purpose VLMs, and 11 medical VLMs, and compare them with two ACTOR variants in the main leaderboard. These include GPT-5.4 (OpenAI, 2026)/5.2 (OpenAI, 2025), Gemini-3.1-Pro (Google DeepMind, 2026)/Flash (Google DeepMind, 2025), Claude Opus 4.6 (Anthropic, 2026); InternVL3.5-38B/14B/8B (Zhu et al., 2025) and Qwen2.5-VL-32B/7B (Bai et al., 2025); and medical VLMs including Lingshu (Xu et al., 2025), MedGemma (Sellersgren et al., 2025), HuatuoGPT-Vision (Chen et al., 2024a), Hulu-Med (Jiang et al., 2025a), MedVLM-R1 (Pan et al., 2025), LLaVA-Med (Li et al., 2023), M3D-LaMed (Bai et al., 2024), and RadFM (Wu et al., 2025).

**Metrics.** For closed-ended tasks, we report accuracy across four tasks. For open-ended tasks, we use LLM-J as the primary metric, where GPT-5.5 scores each prediction on five rubric dimensions, each ranging from 0 to 2, and the total score is scaled to 0100.

**Protocol.** We evaluate on 7,112 items, including 4,214 closed-ended and 2,898 open-ended items. Baselines receive single-turn pre-rendered best-view images with per-task image counts matched to the agentic setting, while our model explores CT volumes through multi-turn tool calls. All final answers are scored with the same metrics.

### 5.2 Main Results

Full results are shown in Tab. 3. ActiveCTRReasoner achieves 73.2% closed-ended accuracy, a 55% relative improvement over GPT-5.4, and 44.2 open-ended LLM-J, which is  $1.9\times$  the best baseline. The advantage is most pronounced on tasks requiring tool-derived quantitative evidence: Measurement, with 98.6% vs. 59.4% for GPT-5.4, and Localization, with 61.4% vs. 41.0% for InternVL3.5-38B, where tool-based volumetric computation and spatial grounding replace imprecise visual estimation. On open-ended tasks, DualSource shows the largest gap, 29.3 vs. 15.2 LLM-J, as longitudinal quantification requires an explicit workflow combining `compare_with_prior`

Table 3: Performance evaluation on the ActiveCTR-Bench (7,112 items). We report accuracy (%) for closed-ended tasks and LLM-J (a 0–100 score judged by GPT-5.5 on 5 dimensions) for open-ended tasks. The best and second-best results are **bolded** and underlined, respectively. **Green subscripts** denote improvement over the base model (Qwen2.5-VL-7B). The evaluated task categories include Presence (Pres.), Localization (Local.), Measurement (Meas.), ChangeClass (Change.), RegionReport (RegRpt.), RegionReasoning (RegReas.), and DualSource (DualSrc.).

| Model                                                 | closed-ended (Accuracy %)   |                             |                             |                             |                             | Open-ended (LLM-J)          |                             |                            |                             |
|-------------------------------------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|----------------------------|-----------------------------|
|                                                       | Pres.                       | Local.                      | Meas.                       | Change.                     | Avg.                        | RegRpt.                     | RegReas.                    | DualSrc.                   | Avg.                        |
| <i>Proprietary API Models</i>                         |                             |                             |                             |                             |                             |                             |                             |                            |                             |
| GPT-5.4 (OpenAI, 2026)                                | 68.4                        | 29.3                        | 59.4                        | 32.3                        | 47.3                        | 21.9                        | 26.9                        | 15.2                       | 22.7                        |
| GPT-5.2 (OpenAI, 2025)                                | 65.3                        | 34.1                        | 53.5                        | 30.4                        | 45.8                        | 24.6                        | 27.2                        | 14.5                       | 23.8                        |
| Gemini-3.1-Pro (Google DeepMind, 2026)                | 59.9                        | 34.6                        | 57.3                        | 30.9                        | 45.7                        | 19.8                        | 13.0                        | 12.4                       | 15.7                        |
| Claude Opus 4.6 (Anthropic, 2026)                     | 58.7                        | 32.4                        | 47.7                        | 31.8                        | 42.7                        | 14.0                        | 6.9                         | 8.7                        | 10.1                        |
| Gemini-3-Flash (Google DeepMind, 2025)                | 59.6                        | 7.0                         | 30.5                        | 4.6                         | 25.4                        | 20.0                        | 19.5                        | 19.8                       | 19.7                        |
| <i>General-Purpose VLMs</i>                           |                             |                             |                             |                             |                             |                             |                             |                            |                             |
| InternVL3.5-38B (Zhu et al., 2025)                    | 59.5                        | 41.0                        | 32.4                        | 29.7                        | 40.6                        | 19.5                        | 12.7                        | 7.6                        | 14.6                        |
| InternVL3.5-14B (Zhu et al., 2025)                    | 57.4                        | 34.8                        | 41.0                        | 29.9                        | 40.8                        | 16.5                        | 12.8                        | 6.6                        | 13.2                        |
| InternVL3.5-8B (Zhu et al., 2025)                     | 57.9                        | 33.5                        | 31.2                        | 22.7                        | 36.3                        | 17.5                        | 15.5                        | 6.7                        | 14.7                        |
| Qwen2.5-VL-32B (Bai et al., 2025)                     | 51.5                        | 35.1                        | 29.8                        | 27.1                        | 35.9                        | 17.7                        | 12.6                        | 12.0                       | 14.6                        |
| Qwen2.5-VL-7B (Bai et al., 2025)                      | 49.3                        | 31.8                        | 42.0                        | 20.3                        | 35.8                        | 22.7                        | 20.4                        | 20.9                       | 21.4                        |
| <i>Medical VLMs</i>                                   |                             |                             |                             |                             |                             |                             |                             |                            |                             |
| Lingshu-32B (Xu et al., 2025)                         | 57.8                        | 34.2                        | 44.9                        | 26.4                        | 40.8                        | 20.0                        | 15.5                        | 14.1                       | 17.1                        |
| MedGemma-27B (Sellingren et al., 2025)                | 57.7                        | 31.1                        | 37.4                        | 23.5                        | 37.4                        | 21.4                        | 14.0                        | 5.5                        | 15.5                        |
| Lingshu-7B (Xu et al., 2025)                          | 54.8                        | 32.0                        | 35.4                        | 24.6                        | 36.7                        | 16.4                        | 14.1                        | 6.9                        | 13.8                        |
| Hulu-Med-7B (Jiang et al., 2025a)                     | 56.4                        | 30.0                        | 36.4                        | 22.2                        | 36.2                        | 22.4                        | 20.9                        | 20.0                       | 21.4                        |
| HuatuoGPT-V-34B (Chen et al., 2024a)                  | 51.7                        | 27.3                        | 32.0                        | 30.1                        | 35.3                        | 16.0                        | 15.7                        | 11.7                       | 15.1                        |
| MedVLM-R1-2B (Pan et al., 2025)                       | 48.2                        | 30.5                        | 39.5                        | 21.3                        | 34.9                        | 16.1                        | 9.1                         | 4.9                        | 11.2                        |
| MedGemma-4B (Sellingren et al., 2025)                 | 56.4                        | 27.5                        | 29.3                        | 22.4                        | 33.9                        | 17.3                        | 10.7                        | 9.6                        | 13.2                        |
| HuatuoGPT-V-7B (Chen et al., 2024a)                   | 60.1                        | 22.6                        | 20.0                        | 22.4                        | 31.3                        | 14.4                        | 9.4                         | 6.0                        | 10.8                        |
| LLaVA-Med-7B (Li et al., 2023)                        | 49.6                        | 0.1                         | 4.6                         | 16.3                        | 17.7                        | 17.1                        | 19.2                        | 22.6                       | 18.9                        |
| M3D-LaMed-4B (Bai et al., 2024)                       | 26.5                        | 6.0                         | 5.1                         | 4.6                         | 10.6                        | 14.0                        | 13.7                        | 16.5                       | 14.3                        |
| RadFM (Wu et al., 2025)                               | 5.9                         | 4.0                         | 0.3                         | 0.0                         | 2.6                         | 10.9                        | 17.0                        | 14.3                       | 14.0                        |
| <i>Ours (Agentic Multi-turn, base: Qwen2.5-VL-7B)</i> |                             |                             |                             |                             |                             |                             |                             |                            |                             |
| ActiveCTReasoner-SFT                                  | 67.8 <sub>+18.5</sub>       | 58.8 <sub>+27.0</sub>       | 97.3 <sub>+55.3</sub>       | 50.6 <sub>+30.3</sub>       | 68.6 <sub>+32.8</sub>       | 45.6 <sub>+22.9</sub>       | 44.4 <sub>+24.0</sub>       | 22.6 <sub>+1.7</sub>       | 41.0 <sub>+19.6</sub>       |
| ActiveCTReasoner                                      | <b>75.4<sub>+26.1</sub></b> | <b>61.4<sub>+29.6</sub></b> | <b>98.6<sub>+56.6</sub></b> | <b>57.4<sub>+37.1</sub></b> | <b>73.2<sub>+37.4</sub></b> | <b>48.5<sub>+25.8</sub></b> | <b>46.4<sub>+26.0</sub></b> | <b>29.3<sub>+8.4</sub></b> | <b>44.2<sub>+22.8</sub></b> |

and repeated measure\_lesion calls that single-turn models cannot replicate. GRPO further improves over SFT by 4.6pp closed-ended and 3.2pp LLM-J. The gain concentrates on measurement-intensive tasks, including DualSource with 6.7pp and Measurement with 1.3pp, indicating that RL optimizes evidence gathering, as analyzed in § 5.3.

### 5.3 Tool Analysis

We analyze the differences in tool-calling behavior between the SFT and RL models on the ActiveCTR-Bench validation set. Fig. 4 visualizes the distribution of tool calls across different task types for both models.

#### SFT establishes task-driven specialization.

The solid bars in Fig. 4 show that SFT learns structured tool recipes mirroring radiologist workflows: (1) scroll is the universal first step (1.00 in every task); (2) perception tasks (Presence/Localization) use only scroll+zoom\_in (1.78–2.00 steps); (3) Measurement always invokes measure\_lesion (100%); (4) all longitudinal

tasks always call compare\_with\_prior; (5) complex tasks chain multiple tools: DualSource averages 7.50 steps with 4.0 measure\_lesion calls. The model achieves 99.8% normal termination (7,098 / 7,112 items).

**RL selectively intensifies measurement.** Comparing solid and hatched bars, RL’s effect is concentrated on a single tool: measure\_lesion increases by +0.51 calls/trajectory overall, while all other tools remain unchanged ( $\Delta < 0.05$ ). The increase is task-adaptive (DualSource +2.3, RegionReasoning +1.3, RegionReport +0.6) while closed-ended tasks stay flat.

This mirrors how radiologists strategically repeat measurements to ensure quantitative confidence—RL aligns the model toward this evidence-gathering behavior without explicit supervision.

**Two mechanisms of RL improvement.** Fig. 5 shows average tool-calling rounds rising steadily during RL training, confirming that RL learns to explore more. Yet closed-ended tasks maintain

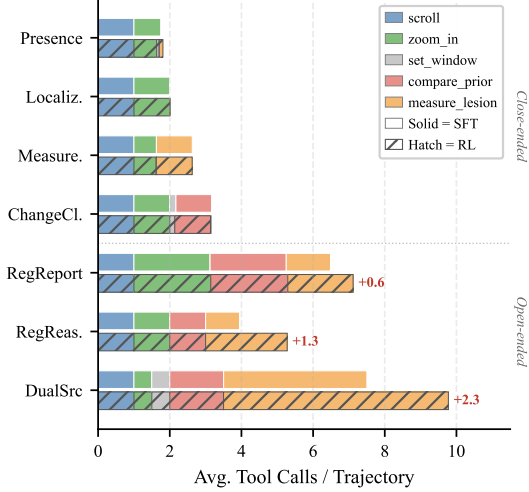


Figure 4: Per-task average tool calls per trajectory. Solid bars and hatched bars represent the SFT model (avg. 3.69 steps) and RL model (avg. 4.19 steps), respectively. Notably, the RL model selectively increases the usage of `measure_lesion` in open-ended tasks.

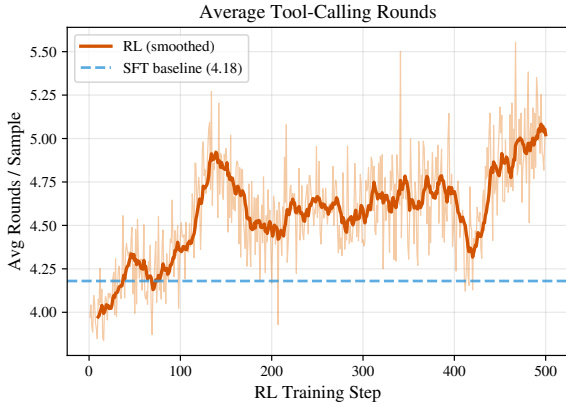


Figure 5: Average number of tool-calling rounds during RL training. The count increases steadily from 3.97 to 4.69 (+18%), surpassing the SFT baseline (4.18) at an early stage.

stable tool patterns while improving significantly (Presence +7.6pp, ChangeClass +6.8pp), indicating these gains come from better *decision quality* over the same observations rather than additional exploration. In summary: SFT teaches *what* tools to use; RL teaches *how much* to use them.

#### 5.4 Ablation Study

As shown in Tab. 4, DirectQA and TextCoT evaluate the same SFT model in non-agentic, single-turn settings with oracle images. In contrast, the agentic multi-turn setting achieves an LLM-J score of 41.0, more than doubling DirectQA (17.7) and TextCoT (18.2). This indicates that interactive tool access, rather than the reasoning format alone, is the primary driver of open-ended answer

Table 4: Ablation study. DirectQA/TextCoT: non-agentic single-turn with pre-rendered oracle images.

|                                    | Acc. (%) | LLM-J (0-100) |
|------------------------------------|----------|---------------|
| <i>Non-Agentic (oracle images)</i> |          |               |
| DirectQA                           | 44.7     | 17.7          |
| TextCoT                            | 41.5     | 18.2          |

Table 5: Results of the radiologist collaboration study involving  $N = 3$  radiologists and 100 cases. Performance is measured using multiple-choice question (MCQ) accuracy for closed-ended tasks and LLM-J quality scores for open-ended tasks.

|                          | Acc. (%)                          | LLM-J (0-100)                    | Time (min)                      |
|--------------------------|-----------------------------------|----------------------------------|---------------------------------|
| <b>100-case results</b>  |                                   |                                  |                                 |
| RAD.                     | $68.8 \pm 15.0$                   | $33.0 \pm 4.0$                   | $4.2 \pm 1.3$                   |
| Agent                    | $71.4 \pm 8.3$                    | $40.0 \pm 11.0$                  | $0.4 \pm 0.1$                   |
| RAD. + Agent             | <b><math>76.0 \pm 16.7</math></b> | <b><math>44.0 \pm 7.0</math></b> | <b><math>2.8 \pm 0.9</math></b> |
| <i>p</i> -value vs. RAD. | 0.0039                            | $\leq 0.016$                     | $< 0.01$                        |

quality. The agentic setting also substantially improves closed-ended accuracy from 44.7 to 68.6 (+23.9pp), largely driven by the pronounced gain on Measurement (97.3 vs. 35.1).

Additional analyses are reported in Appendix C. Tool-prompted VLMs with access to the same five PACS tools remain far below ActiveCTReasoner, and structured-tool removal mainly degrades the tasks that require the removed computation, indicating that the gains are not explained by tool access alone.

#### 5.5 Radiologist Collaboration Study

To obtain preliminary evidence for the decision-support utility of the radiologist-aligned paradigm, we conduct a radiologist collaboration study comparing three conditions on 100 cases sampled from the ActiveCTR-Bench validation set, and additionally report variability estimates and paired statistical tests.

**Setup.** Professional radiologists independently evaluated each case under three conditions: (A) *Radiologist alone*: standard PACS-based reading; (B) *Radiologist + Agent*: ActiveCTReasoner provides a candidate report with step-by-step reasoning, and the radiologist reviews, corrects, and finalizes; (C) *Agent alone*: no human involvement. All conditions are evaluated on closed-ended accuracy, open-ended LLM-J, and time per case.

Preliminary results are reported in Tab. 5. Radiologist + Agent improves LLM-J to 44.0, 11.0 points over Radiologist alone, and closed-ended accuracy to 76.0%, 7.2pp higher, while reducing average reading time by about one third, from 4.2

to 2.8 min per case. These gains suggest that the agent’s pre-computed measurements and structured reasoning provide a strong draft for radiologists to verify and refine. Radiologist + Agent also outperforms Agent alone on open-ended tasks, with 44.0 vs. 40.0 LLM-J, validating the transparency design: each agent step is rendered as a human-readable image, enabling radiologists to inspect and improve the reasoning with clinical expertise. On closed-ended tasks, Agent alone exceeds Radiologist alone, 71.4% vs. 68.8%, mainly driven by Measurement, where tool-based quantification replaces imprecise visual estimation.

## 6 Conclusion

We present ACTOR, a CT reasoning framework that aligns agent actions with standard PACS operations, making each intermediate state a human-readable image for transparent clinical decision-making. We build ActiveCTR-Bench with 7 tasks and 7,112 vision-verified items, synthesize 108K multi-turn trajectories, and train ActiveCTReasoner, a 7B model optimized with SFT and GRPO. ActiveCTReasoner achieves 73.2% closed-ended accuracy and 44.2 LLM-J, substantially outperforming strong baselines. Further analyses show that tool images provide task-critical visual evidence, while RL improves quantitative reasoning by encouraging more intensive measurement. A radiologist collaboration study provides preliminary evidence that transparent step-level outputs can improve answer quality (+11.0 LLM-J) and efficiency (−33% reading time) under the study protocol, suggesting a path from passive single-turn CT QA toward active, multi-step decision-support research.

## Limitations

Our study is currently limited to a 7B-scale backbone due to computational constraints. Although the proposed framework already demonstrates strong performance, exploring larger multimodal models and longer-horizon training may further improve reasoning and tool-use capabilities. In addition, progress in agentic medical reasoning is constrained by the limited availability of large-scale open-source clinical datasets with high-quality annotations and interaction supervision. Future work may benefit from richer open medical datasets and more diverse real-world clinical trajectories.

## Acknowledgements

The research is financially supported by the National Natural Science Foundation of China (No. 62376276), the Beijing Nova Program (No. 20230484278), and Big Data and Responsible Artificial Intelligence for National Governance, Renmin University of China.

## References

- Anthropic. 2026. [Introducing claude opus 4.6](#).
- Fan Bai, Yuxin Du, Tiejun Huang, Max Q-H Meng, and Bo Zhao. 2024. M3d: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2.5-vl technical report. *arXiv e-prints*, pages arXiv–2502.
- Léo Butsanets, Charles Corbière, Julien Khaut, Pierre Manceron, and Corentin Dancette. 2025. Radimagenet-vqa: A large-scale ct and mri dataset for radiologic visual question answering. *arXiv preprint arXiv:2512.17396*.
- Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, et al. 2024a. Towards injecting medical visual knowledge into multimodal llms at scale. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 7346–7370.
- Zhihong Chen, Maya Varma, Justin Xu, Magdalini Paschali, Dave Van Veen, Andrew Johnston, Alaa Youssef, Louis Blankemeier, Christian Bluethgen, Stephan Altmayer, et al. 2024b. A vision-language foundation model to enhance efficiency of chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*.
- Adibvafa Fallahpour, Jun Ma, Alif Munim, Hongwei Lyu, and Bo Wang. 2025. Medrax: Medical reasoning agent for chest x-ray. *arXiv preprint arXiv:2502.02673*.
- Google DeepMind. 2025. [Gemini 3.1 flash](#).
- Google DeepMind. 2026. [Gemini 3.1 pro](#).
- Yannian Gu, Xizhuo Zhang, Linjie Mu, Yongrui Yu, Zhongzhen Huang, Shaoting Zhang, and Xiaofan Zhang. 2026. Ct-flow: Orchestrating ct interpretation workflow with model context protocol servers. *arXiv preprint arXiv:2603.00123*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in

- llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Seval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, et al. 2024. Developing generalist foundation models from a multimodal dataset for 3d computed tomography. *arXiv preprint arXiv:2403.17834*.
- Xiaoke Huang, Juncheng Wu, Hui Liu, Xianfeng Tang, and Yuyin Zhou. 2025. Medvlthinker: Simple baselines for multimodal medical reasoning. *arXiv preprint arXiv:2508.02669*.
- Songtao Jiang, Yuan Wang, Sibao Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, Jin Hao, Zijian Chen, Ruijia Wu, Tao Tang, Junhui Lv, Hongxia Xu, Hongwei Wang, Jun Xiao, Bin Feng, and 6 others. 2025a. [Hulu-med: A transparent generalist model towards holistic medical vision-language understanding](#). *Preprint*, arXiv:2510.08668.
- Yankai Jiang, Yujie Zhang, Peng Zhang, Yichen Li, Jintai Chen, Xiaoming Shi, and Shihui Zhen. 2025b. Incentivizing tool-augmented thinking with images for medical image analysis. *arXiv preprint arXiv:2512.14157*.
- Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, Yuheng Li, Konstantinos Psounis, and Xiaofeng Yang. 2026. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *IEEE transactions on medical imaging*.
- Binxu Li, Tiankai Yan, Yuanting Pan, Jie Luo, Ruiyang Ji, Jiayuan Ding, Zhe Xu, Shilong Liu, Haoyu Dong, Zihao Lin, et al. 2024. Mmedagent: Learning to use medical tools with multi-modal agent. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8745–8760.
- 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. *Advances in Neural Information Processing Systems*, 36:28541–28564.
- Xiao Liu, Tianjie Zhang, Yu Gu, Iat Long Iong, Song XiXuan, Yifan Xu, Shudan Zhang, Hanyu Lai, Jiadai Sun, Xinyue Yang, et al. 2025. Visualagent-bench: Towards large multimodal models as visual foundation agents. In *International Conference on Learning Representations*, volume 2025, pages 95650–95707.
- Meng Lu, Yuxing Lu, Yuchen Zhuang, Megan Mullins, Yang Xie, Guanghua Xiao, Charles Fleming, Wenqi Shi, and Xuan Wang. 2026. Medvistagym: A scalable training environment for thinking with medical images via tool-integrated reinforcement learning. *arXiv preprint arXiv:2601.07107*.
- Yuren Mao, Wenyi Xu, Yuyang Qin, and Yunjun Gao. 2026. Ct-agent: a multimodal-llm agent for 3d ct radiology question answering. *Science China Information Sciences*, 69(5):150107.
- OpenAI. 2025. [Introducing gpt5.2](#).
- OpenAI. 2026. [Introducing gpt5.4](#).
- Giuseppe A Orlando, Paolo Papotti, Maria A Zuluaga, Olivier Humbert, and Marco Lorenzi. 2026. Medscribe: Clinically grounded ct reporting through agentic workflows. *arXiv preprint arXiv:2605.01779*.
- Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. 2025. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 337–347. Springer.
- Mélanie Roschewitz, Kenneth Styppa, Yitian Tao, Jiwoong Sohn, Jean-Benoit Delbrouck, Benjamin Gundersen, Nicolas Deperrois, Christian Bluethgen, Julia E Vogt, Bjoern Menze, et al. 2026. Rada-gent: A tool-using ai agent for stepwise interpretation of chest computed tomography. *arXiv preprint arXiv:2604.15231*.
- Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36:68539–68551.
- Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. 2024. Deepseek-math: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Ziyue Wang, Linghan Cai, Chang Han Low, Haofeng Liu, Junde Wu, Jingyu Wang, Rui Wang, Lei Song, Jiang Bian, Jingjing Fu, et al. 2026. 3dmedagent: Unified perception-to-understanding for 3d medical analysis. *arXiv preprint arXiv:2602.18064*.
- Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. 2023. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5(5):e230024.

- Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. 2025. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 16(1):7866.
- Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. 2025. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.
- Suhao Yu, Haojin Wang, Juncheng Wu, Luyang Luo, Jingshen Wang, Cihang Xie, Pranav Rajpurkar, Carl Yang, Yang Yang, Kang Wang, et al. 2025. Medframeqa: A multi-image medical vqa benchmark for clinical reasoning. *arXiv preprint arXiv:2505.16964*.
- Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.
- Qingqing Zhu, Qiao Jin, Tejas S Mathai, Yin Fang, Zhizheng Wang, Yifan Yang, Maame Sarfo-Gyamfi, Benjamin Hou, Ran Gu, Praveen TS Balamuralikrishna, et al. 2026. Ct-bench: A benchmark for multimodal lesion understanding in computed tomography. *arXiv preprint arXiv:2602.14879*.

## A Implementation Details

In this section, we provide comprehensive details regarding the experimental setup to ensure reproducibility and clarity. We begin by characterizing the composition of the ActiveCTR-Bench validation split and its underlying three-level reading framework in Appendix A.1. Subsequently, we outline the specific configurations employed during the SFT and GRPO stages in Appendix A.3. Finally, we analyze the task distribution strategies in Appendix A.4 to balance performance across different cognitive levels.

### A.1 Benchmark Setup.

To characterize the composition of the validation split, we summarize the statistics of ActiveCTR-Bench in Tab. A.1. The benchmark is organized according to the proposed three-level reading framework, covering perception, interpretation, and reasoning tasks. For compactness, we denote these three levels as P, I, and R, respectively. Each task is associated with visual ground truth derived from automated visual tools, ensuring that evaluation is grounded in verifiable image-based evidence rather than purely textual annotations.

Table A.1: Statistics of ActiveCTR-Bench validation set.

| Task         | Lvl | Type    | Val                  | GT Source            |
|--------------|-----|---------|----------------------|----------------------|
| Localization | P   | 4-MCQ   | 1,295                | Lobe masks           |
| Presence     | P   | Yes/No  | 1,000                | AI clf. labels       |
| Measurement  | I   | 3/4-MCQ | 958                  | Mask volume          |
| ChangeClass  | I   | 4-MCQ   | 961                  | Report + mask verif. |
| RegionReport | I   | Open    | 1,188                | Multi-source         |
| RegionReas.  | R   | Open    | 1,188                | Multi-source         |
| DualSource   | R   | Open    | 522                  | Multi-source         |
| <b>Total</b> |     |         | <b>7,112 samples</b> |                      |

### A.2 Benchmark Provenance and Radiologist Audit

We make the provenance of each task explicit in Tab. A.2. LLMs are used for clinically natural wording or synthesis conditioned on structured evidence; they are not the sole source of benchmark ground truth. Reference answers are anchored in CT-RATE reports and labels, TotalSegmentator-derived masks, lobe masks, metadata, longitudinal scan pairing, and deterministic mask-derived measurements.

To further assess benchmark validity, three board-certified radiologists audit 210 sampled cases, stratified as 30 cases per task. As shown in

Tab. A.3, each item is rated on a 0–2 scale, where 0 denotes unsupported or invalid, 1 denotes acceptable with minor caveats, and 2 denotes fully acceptable. An item is counted as majority-passed when at least two of the three radiologists assign a score of 1 or higher.

All audit criteria achieve majority-pass rates above 91%. The only error rate above 5% is visual necessity, where 8.6% of audited cases were judged partially inferable from clinical context but still requiring image confirmation for full correctness.

### A.3 Training configuration.

Tab. A.4 summarizes the hyperparameters for both SFT and GRPO stages. SFT trains on 108K multi-turn trajectories (avg. 14.5 turns) for 2 epochs with frozen vision components. GRPO starts from the SFT model and runs for 6 epochs with 8 rollouts per prompt. We freeze the vision tower and MM projector throughout both stages to preserve visual grounding while focusing capacity on reasoning and tool-calling.

### A.4 Task distribution.

The rightmost panel of Tab. A.5 shows per-task counts. SFT is roughly uniform across tasks (13%–18% each) except DualSource (3.8%, limited by longitudinal pairs). RL upsamples Localization and ChangeClass (25% each) to focus on tasks where the model has moderate difficulty; Presence is resampled to 50:50 Yes/No to correct label imbalance.

## B Tool Specifications

This section details the suite of external tools integrated into our framework to enhance visual grounding and quantitative reasoning. We first specify the API interfaces for each tool, describing their input parameters and multimodal return formats in Appendix B.1. Subsequently, we analyze the empirical usage patterns of these tools during inference in Appendix B.2.

### B.1 API specification.

Tab. B.6 provides the complete tool API. Each tool returns a text–image tuple, where the text contains structured data or execution confirmation and the images are rendered 2D views from the PACS simulator. The default window preset is lung ( $W = 1500, L = -600$ ). The scroll tool renders six uniformly spaced key frames within the requested

Table A.2: Per-task ground-truth provenance. LLMs are used for wording or trajectory synthesis; reference answers are anchored in traceable CT/report/mask/metadata evidence.

| Task            | Question source                     | Answer / GT source                                | LLM role            |
|-----------------|-------------------------------------|---------------------------------------------------|---------------------|
| Presence        | Report finding + visual evidence    | Abnormality label / report-mask consistency       | Question phrasing   |
| Localization    | Lesion/anatomy metadata             | Lobe mask / spatial location                      | Question phrasing   |
| Measurement     | Lesion/structure target             | Segmentation-derived measurement                  | Question phrasing   |
| ChangeClass     | Prior-current reports + masks       | Longitudinal report/mask consistency              | Question phrasing   |
| RegionReport    | Region-specific evidence            | Report + mask-derived visual evidence             | Wording / synthesis |
| RegionReasoning | Region evidence + metadata          | Visual/quantitative evidence + clinical rationale | Wording / synthesis |
| DualSource      | Prior-current multi-source evidence | Longitudinal visual/mask/report evidence          | Wording / synthesis |

Table A.3: 210-case benchmark validity audit by three radiologists.

| Audit criterion                                      | Mean score (0–2) | Majority-pass | Error rate |
|------------------------------------------------------|------------------|---------------|------------|
| Clinically meaningful question                       | 1.88             | 100.0%        | 0.0%       |
| Clear and answerable question                        | 1.19             | 98.1%         | 1.9%       |
| Correct/defensible answer                            | 1.34             | 99.5%         | 0.5%       |
| GT source appropriate and traceable                  | 1.49             | 100.0%        | 0.0%       |
| Answer supported by CT/mask/report/metadata evidence | 1.15             | 99.0%         | 1.0%       |
| Visual evidence necessary                            | 1.76             | 91.4%         | 8.6%       |

Table A.4: Training configurations for SFT and GRPO.

| SFT Configuration       |                           |
|-------------------------|---------------------------|
| Base model              | Qwen2.5-VL-7B             |
| Training data           | 108,085 samples           |
| Avg. turns              | 14.5                      |
| Hardware                | 16×A800                   |
| Framework               | LLaMA-Factory             |
| Parallelism             | DeepSpeed ZeRO-3          |
| Optimizer               | linear warmup             |
| Learning rate           | 1e-5                      |
| Vision tower / MM proj. | Frozen                    |
| GRPO Configuration      |                           |
| Training data           | 12,011 samples            |
| Data filtering          | pass_rate ∈ [0.25, 0.625] |
| Presence resample       | Yes:No = 50:50            |
| Hardware                | 16×A800                   |
| KL coefficient          | 0.01                      |
| Learning rate           | 1e-6                      |
| Clip ratio              | [0.2, 0.3]                |
| Max rounds              | 12                        |

range, with each frame shown in dual windows, i.e., lung and mediastinum, to provide a comprehensive overview. The `measure_lesion` tool leverages `TotalSegmentator` masks for volumetric computation and returns calibrated measurements, including volume in  $\text{cm}^3$ , equivalent diameter, or longitudinal change percentage.

## B.2 Calling statistics.

As shown in Tab. B.7, navigation tools (`scroll+zoom_in`) account for 56.1% of all calls, reflecting their role as universal first steps. Quantitative tools (`compare_with_prior+measure_lesion`) comprise 42.2%, activated selectively for interpre-

Table A.5: Task distribution statistics of ActiveCTR-Bench. We filter RL samples based on a pass rate range of [0.25, 0.625] and perform resampling to ensure a 50:50 balance ratio between positive and negative instances.

| Task            | SFT     |          | RL     |          |
|-----------------|---------|----------|--------|----------|
|                 | Count   | Pct. (%) | Count  | Pct. (%) |
| Localization    | 19,513  | 18.1     | 3,000  | 25.0     |
| RegionReasoning | 18,600  | 17.2     | 947    | 7.9      |
| RegionReport    | 18,600  | 17.2     | 1,521  | 12.7     |
| ChangeClass     | 17,563  | 16.2     | 3,000  | 25.0     |
| Presence        | 14,992  | 13.9     | 2,576  | 21.4     |
| Measurement     | 14,692  | 13.6     | 730    | 6.1      |
| DualSource      | 4,125   | 3.8      | 237    | 2.0      |
| Total           | 108,085 | 100      | 12,011 | 100      |

tation and reasoning tasks. `set_window` is rare (1.7%) as most tasks are adequately served by the default lung window; it is primarily invoked for mediastinal structures (pericardial effusion, lymph nodes).

Table B.6: Summary of Tool APIs. Each tool yields a structured multimodal output consisting of textual descriptions and corresponding 2D images. The default viewing window is set for the lung region.

| Tool                        | Parameters                                             | Return                            |
|-----------------------------|--------------------------------------------------------|-----------------------------------|
| <code>scroll</code>         | <code>z_start, z_end</code> : int                      | Text + 6 key frames (dual window) |
| <code>zoom_in</code>        | <code>bbox</code> : List[int]; <code>mode</code> : str | Magnified frames                  |
| <code>set_window</code>     | <code>preset</code> : str                              | Re-rendered confirmation          |
| <code>compare_prior</code>  | <code>z_indices</code> : List[int]                     | Side-by-side (curr. vs. prior)    |
| <code>measure_lesion</code> | <code>target</code> : str; <code>bbox</code> : List    | Volume, diameter, or change       |

Table B.7: Frequency statistics of tool usage of the SFT model on 7,112 trajectories.

| Tool               | Calls of Tool | Percentage (%) |
|--------------------|---------------|----------------|
| scroll             | 7,112         | 27.2           |
| zoom_in            | 7,579         | 28.9           |
| set_window         | 457           | 1.7            |
| compare_with_prior | 5,468         | 20.9           |
| measure_lesion     | 5,567         | 21.3           |
| <b>Total</b>       | <b>26,183</b> | <b>100</b>     |

## C Additional Tool and Evidence Analyses

We add analyses to separate the contribution of tool availability, learned action selection, and structured tool computation. First, Tab. C.8 evaluates open-source medical agents and prompted-tool VLMs under the same ActiveCTR-Bench protocol. Despite access to the same five PACS tools, the best prompted-tool baseline reaches only 33.3% closed-ended accuracy and 22.2 LLM-J, far below ActiveCTReasoner’s 73.2% and 44.2.

Second, Tab. C.9 disables structured tools at inference. Removing `measure_lesion` primarily degrades Measurement accuracy, while removing `compare_with_prior` mainly affects Change-Class. This separates tool-backed computation from the learned policy that selects and integrates evidence.

## D Prompt Templates

This section presents the detailed prompt templates employed throughout our experimental pipeline. We provide the core system prompt that defines the agentic behavior and tool interaction protocol in Appendix D.1. Additionally, we specify the distinct prompts used for non-agentic VQA baselines in Appendix D.2, the specialized instructions for generating Chain-of-Thought reasoning traces during data synthesis in Appendix D.3, and the rubric-based evaluation prompts utilized by the LLM-as-Judge framework in Appendix D.4.

### D.1 Agentic System Prompt

The following system prompt defines the agent’s role and tool interface, used for both trajectory synthesis and agentic evaluation.

#### Agentic System Prompt

You are an expert radiologist and an AI medical agent capable of dynamically interacting with a

PACS simulator to explore 3D CT scans. Your environment allows you to look at 2D axial slices. You can scroll through slices, adjust the CT window, zoom in on regions of interest, take physical measurements of lesions, and compare the current scan with a prior study side-by-side.

For EVERY step, first explain your reasoning using `<think>...</think>` tags, then choose exactly one action:

– A tool call:  
`<tool\_call>\{"name": "tool\_name", "arguments": \{...\}\}</tool\_call>`  
– A final answer:  
`<answer>...</answer>`

#### Available Tools:

- `scroll` --- Navigate through axial slices  
`\quad <tool\_call>\{"name": "scroll", "arguments": \{"z\_start": 0, "z\_end": 100\}\}</tool\_call>`
- `zoom_in` --- Magnify a region of interest  
`\quad <tool\_call>\{"name": "zoom\_in", "arguments": \{"bbox": [x0,y0,x1,y1], "mode": "highlight"\}\}</tool\_call>`
- `set_window` --- Change CT window preset  
`\quad <tool\_call>\{"name": "set\_window", "arguments": \{"preset": "mediastinum"\}\}</tool\_call>`
- `compare_with_prior` --- Side-by-side comparison  
`\quad <tool\_call>\{"name": "compare\_with\_prior", "arguments": \{"z\_indices": [50, 60, 70]\}\}</tool\_call>`
- `measure_lesion` --- Quantitative measurement  
`\quad <tool\_call>\{"name": "measure\_lesion", "arguments": \{"target": "liver"\}\}</tool\_call>`

#### Important Rules:

- You MUST use the exact `<tool\_call>JSON</tool\_call>` format\
- You MUST wrap your final answer in `<answer>...</answer>` tags
- For MCQ questions, output ONLY the letter: `<answer>A</answer>`
- Always think step by step before acting

Table C.8: Comparison of open-source medical agents, tool-prompted VLMs, and ActiveCTReasoner on ActiveCTR-Bench. CE Avg = closed-ended MCQ accuracy (%); OE columns = LLM-Judge scores (0–100).

| Model                              | Pres.       | Local.      | Meas.       | Change.     | CE Avg.     | RegRpt.     | RegReas.    | DualSrc.    | OE Avg.     |
|------------------------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| MedVLThinker (Huang et al., 2025)  | 50.1        | 29.9        | 40.5        | 23.4        | 35.6        | 34.6        | 33.9        | 23.0        | 32.2        |
| RadAgent (Roschewitz et al., 2026) | 53.7        | 11.6        | 29.1        | 19.3        | 27.3        | 35.1        | 33.3        | 22.8        | 32.1        |
| 3DMedAgent (Wang et al., 2026)     | 51.6        | 16.9        | 44.9        | 43.0        | 37.5        | 39.8        | 30.5        | 22.9        | 32.9        |
| Hulu-Med-7B prompted               | 57.1        | 28.9        | 24.4        | 20.8        | 32.7        | 32.7        | 20.0        | 21.6        | 25.5        |
| Lingshu-7B prompted                | 53.1        | 34.8        | 22.1        | 22.0        | 33.3        | 20.7        | 19.6        | 19.9        | 20.1        |
| Lingshu-32B prompted               | 44.0        | 34.4        | 26.0        | 25.9        | 32.8        | 19.8        | 20.4        | 20.4        | 20.2        |
| Qwen2.5-VL-32B prompted            | 49.5        | 26.8        | 24.6        | 23.0        | 30.8        | 22.7        | 23.5        | 18.1        | 22.2        |
| InternVL3.5-8B prompted            | 44.0        | 29.2        | 19.9        | 23.7        | 29.3        | 19.1        | 21.0        | 19.4        | 19.9        |
| InternVL3.5-38B prompted           | 46.1        | 25.6        | 13.7        | 10.0        | 24.2        | 17.3        | 17.8        | 13.7        | 16.9        |
| <b>ActiveCTReasoner</b>            | <b>75.4</b> | <b>61.4</b> | <b>98.6</b> | <b>57.4</b> | <b>73.2</b> | <b>48.5</b> | <b>46.4</b> | <b>29.3</b> | <b>44.2</b> |

Table C.9: Tool-removal ablation on closed-ended tasks. Parentheses show change from full ActiveCTReasoner.

| Setting                | Pres. | Local. | Meas. | Change. | CE Avg. |
|------------------------|-------|--------|-------|---------|---------|
| Full ActiveCTReasoner  | 75.4  | 61.4   | 98.6  | 57.4    | 73.2    |
| w/o measure_lesion     | 74.6  | 60.7   | 54.9  | 56.8    | 61.8    |
| w/o compare_with_prior | 75.2  | 60.8   | 98.9  | 49.8    | 70.4    |
| w/o both               | 74.3  | 60.2   | 55.6  | 49.2    | 60.0    |

## D.2 VQA Evaluation Prompts

Non-agentic baselines receive oracle pre-rendered images with the following system prompts.

### Single-Scan System Prompt

You are an expert radiologist analyzing a chest CT scan. You are given several axial CT slices from the relevant anatomical region. Each slice is shown in two windows (lung and mediastinum) for comprehensive evaluation. Analyze the images carefully and answer the question.

### Longitudinal System Prompt

You are an expert radiologist comparing two CT scans of the same patient. You are given paired axial CT slices from the current scan and prior scan. For each z-position, you see: current scan (lung window), current scan (mediastinum window), prior scan (lung window), prior scan (mediastinum window). Analyze the changes between the two scans and answer the question.

*Question suffixes:* MCQ tasks append “Answer with the letter only (A/B/C/D).” Presence tasks append “Answer with Yes or No only.” Open-ended tasks use no suffix.

## D.3 Think-Filling Prompt

GPT-5.4 generates <think> reasoning for each trajectory step. Each task type has a specialized system prompt; below is the Measurement task as a representative example.

### Think-Filling Prompt (Measurement Task)

You are an expert radiologist AI agent performing chest CT interpretation. You will be given a quantitative measurement question, the correct answer, planned tool calls, and their results. Your job: write the internal reasoning (think text) that appears BEFORE each action.

\medskip  
TEMPORAL RULE: Each think is generated BEFORE the tool is called. You do NOT yet know the tool's result.

\medskip  
RULES FOR MEASUREMENT (MCQ --- quantitative elimination):\\  
Navigation steps:\\  
- measure\\_lesion: State your INTENT to measure. Do NOT quote the measurement result.

\medskip  
FINAL answer step ---  
QUANTITATIVE ELIMINATION:\\  
1. State the measured value from tool\\_response.\\  
2. For each option, compare measurement against that option's threshold/range.\\  
3. For each WRONG option: ``Option [X] ([range]) is excluded: [measured value] does not fall in this range.''\n\\  
4. Confirm matching option.

Other task types follow analogous patterns: Presence requires bidirectional evidence (FOR/A-

GAINST); RegionReasoning requires a clinical reasoning chain (Observation → Quantification → Rate Assessment → Clinical Significance).

#### D.4 LLM-as-Judge Prompt

GPT-5.5 scores open-ended predictions on 5 rubric dimensions per task.

##### LLM-as-Judge System Prompt

You are an expert radiologist evaluating AI-generated CT scan reports. You will be given a REFERENCE report (ground truth) and a PREDICTION (model output). Score the prediction on multiple dimensions using the rubrics below. Be strict but fair. The prediction does NOT need to match the reference word-for-word, but must convey the same clinical information accurately.

Each dimension is scored 0–2. Raw total (0 to 10) is linearly scaled to 0–100. Output format: {"dimension": score, ..., "total": sum}.

##### Rubric Dimensions

###### RegionReport:

1. **Finding Coverage** (0--2): main findings covered?
2. **Precision** (0--2): anatomical locations accurate?
3. **Quantitative Accuracy** (0--2): numbers correct (<5% error)?
4. **Change Reasoning** (0--2): direction + characterization?
5. **Faithfulness** (0--2): no fabricated findings?

###### RegionReasoning:

1. **Finding Coverage** (0--2): key findings identified?
2. **Precision** (0--2): clinical reasoning chain logical?
3. **Quantitative Accuracy** (0--2): numbers correct (<5% error)?
4. **Change Reasoning** (0--2): temporal change assessment?
5. **Faithfulness** (0--2): no fabricated findings?

###### DualSource:

1. **Finding Coverage** (0--2): all changes detected?
2. **Precision** (0--2): change direction (increase/decrease) correct?
3. **Quantitative Accuracy** (0--2): numerical changes correct?

4. **Change Reasoning** (0--2): clinical significance assessed?
5. **Faithfulness** (0--2): no fabricated findings?

#### D.5 LLM-Judge Human Agreement

To validate open-ended evaluation, three board-certified radiologists score 90 open-ended predictions sampled from RegionReport, RegionReasoning, and DualSource using the same five-dimension rubric. As shown in Tab. D.10, the GPT-5.5 judge shows strong agreement with radiologist ratings, with per-dimension Pearson correlations from 0.764 to 0.942. Faithfulness/hallucination has the lowest correlation, reflecting the finer-grained judgment needed to distinguish unsupported statements from acceptable clinical inference.

Table D.10: LLM-Judge validation against radiologist ratings on 90 open-ended predictions.

| Rubric dimension                | Pearson $r$ with radiologist scores |
|---------------------------------|-------------------------------------|
| Finding coverage                | 0.940                               |
| Anatomical / clinical precision | 0.803                               |
| Quantitative accuracy           | 0.820                               |
| Temporal / change reasoning     | 0.942                               |
| Faithfulness / hallucination    | 0.764                               |

## E Radiologist Collaboration Study

This section is organized as follows: Appendix E.1 outlines the within-subject experimental protocol designed to mitigate learning effects and ensure fair comparison. Appendix E.2 presents the comprehensive metrics used for evaluation, encompassing both quantitative diagnostic outcomes and qualitative user experience measures.

### E.1 Protocol

Professional radiologists participated in a within-subject design:

- **Condition A (Radiologist Alone):** Days 1–3. PDF materials with oracle key-frame images and clinical question. No AI assistance.
- **Condition B (Radiologist + Agent):** Days 4–6 ( $\geq 1$  day gap). Same materials plus the agent’s full reasoning trajectory (tool call images, measurements, candidate answer). Radiologist reviews, corrects, and finalizes.

- **Condition C (Agent Alone):** ActiveCTReasoner predictions (final RL model). No human involvement.

Answers were collected via structured Excel sheets.

## E.2 Metrics

- *Diagnostic performance:* MCQ/Yes-No → Accuracy; Open-ended → LLM-J (GPT-5.5) and Token-F1.
- *Time efficiency:* Minutes per case.
- *Subjective evaluation (Condition B only):* 5 dimensions on 1–5 Likert: Confidence, Reasoning Quality, Diagnostic Accuracy, Completeness, Helpfulness.

In the expanded 100-case analysis, we report mean and standard deviation for each condition in Tab. 5 and perform paired tests comparing Radiologist alone with Radiologist + Agent. For closed-ended tasks, we use McNemar’s test for paired correctness; for open-ended scores, we use Wilcoxon signed-rank tests; for reading time, we report paired-test significance.

Radiologist + Agent improves MCQ accuracy from 68.8% to 76.0% ( $\Delta = +7.2\text{pp}$ ), improves LLM-J from 33.0 to 44.0 ( $\Delta = +11.0$ ), and reduces reading time from 4.2 to 2.8 minutes. Interrater agreement under the Radiologist-alone condition is substantial for closed-ended decisions (Fleiss’  $\kappa = 0.635$ ) and good for open-ended quality scores (ICC(3,k)=0.81).

## F N-gram Metrics

Tab. F.11 reports four fine-grained metrics for the 21 external baseline models and two ACTOR variants across three open-ended tasks: Token-F1, ROUGE-L, Numerical Recall (NR), and Direction Accuracy (DA). While the LLM-as-a-Judge (LLM-J) metric evaluates high-level semantic alignment and reasoning, these metrics complement it by enforcing strict token-level correctness and clinical factual precision. Specifically, **Token-F1** and **ROUGE-L** quantify the lexical overlap and structural coherence of the generated reports compared to the ground truth, capturing the model’s fidelity to professional medical terminology. To penalize hallucinations in critical clinical details, we introduce **Numerical Recall**, which measures the model’s ability to accurately retrieve

essential quantitative measurements (e.g., lesion dimensions and attenuation values). Crucially, **Direction Accuracy** evaluates the model’s prognostic reasoning capability by verifying whether it correctly identifies the trajectory of disease progression (e.g., worsening, resolving, or stable), an essential requirement for dynamic and longitudinal tasks.

**Key observations.** (1) Numerical Recall (NR) is the most discriminative metric: baselines cluster near random ( $\sim 16.5$  for single-scan,  $\sim 0.6$  for DualSource) because they lack measurement tools, while our RL model reaches 70.8/66.1/54.2. (2) Direction Accuracy (DA) shows that even strong baselines (GPT-5.4: 34.9/28.9/42.5) struggle to assess temporal change direction without explicit longitudinal comparison. (3) Token-F1 and ROUGE-L correlate with LLM-J but are less sensitive to clinical correctness—e.g., LLaVA-Med achieves reasonable TF1 (15.6) despite low LLM-J (18.9) due to generic medical text without specific findings.

## G Extended Case Studies

We present four representative agentic trajectories illustrating how ACTOR chains radiologist-aligned tools to reach clinical conclusions. Each case demonstrates a distinct reasoning pattern.

### Case A: Tool-Backed Nodule Measurement

**Task:** Measurement (nodule size grading)

**Steps:** 3

1. `scroll(142)` → Axial slice at lung nodule region
2. `zoom_in(right_lower_lobe)` → Magnified view centering the nodule
3. `measure_lesion(lung_nodule)` → Volume 1.83 cm<sup>3</sup>, equiv. diameter 15.2 mm

**Answer:** “Large (>10 mm)” ✓ The canonical scroll→zoom→measure pattern achieves 98.6% on this task type.

Table F.11: Evaluation of N-gram metrics across all open-ended tasks (all values  $\times 100$ ). We report Token-F1 (TF1), ROUGE-L (RL), Numerical Recall (NR), and Direction Accuracy (DA). The best and second-best results are **bolded** and underlined, respectively. **Green** deltas denote improvement over the base model (Qwen2.5-VL-7B).

| Model                                                 | RegionReport     |                  |                  |                  | RegionReasoning  |                  |                  |                  | DualSource       |                  |                  |                  |
|-------------------------------------------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|------------------|
|                                                       | TF1              | RL               | NR               | DA               | TF1              | RL               | NR               | DA               | TF1              | RL               | NR               | DA               |
| <i>Proprietary API Models</i>                         |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |
| GPT-5.4                                               | 21.1             | 15.7             | 16.5             | 34.9             | 26.9             | 15.9             | 16.6             | 28.9             | 21.7             | 12.1             | 0.6              | 42.5             |
| GPT-5.2                                               | 21.6             | 15.6             | 16.5             | 33.2             | 27.1             | 16.1             | 16.6             | 20.8             | 20.5             | 11.6             | 0.6              | 36.7             |
| Gemini-3.1-Pro                                        | 15.3             | 12.1             | 16.5             | 31.1             | 10.2             | 7.2              | 16.6             | 21.8             | 8.1              | 5.3              | 0.6              | 45.5             |
| Claude Opus 4.6                                       | 11.6             | 7.7              | 16.5             | 40.7             | 19.4             | 11.1             | 16.7             | 42.0             | 19.5             | 9.9              | 0.6              | 37.3             |
| Gemini-3-Flash                                        | 5.1              | 3.8              | 16.5             | 34.3             | 1.2              | 0.7              | 16.6             | 35.1             | 0.2              | 0.1              | 0.6              | 0.0              |
| <i>General-Purpose VLMs</i>                           |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |
| InternVL3.5-38B                                       | 16.7             | 11.7             | 16.5             | 37.7             | 20.9             | 13.3             | 16.7             | 37.3             | 16.7             | 9.4              | 1.0              | 32.1             |
| InternVL3.5-14B                                       | 15.9             | 11.2             | 16.6             | 33.8             | 21.0             | 13.2             | 16.6             | 33.4             | 17.2             | 9.7              | 0.7              | 34.1             |
| InternVL3.5-8B                                        | 17.3             | 12.3             | 16.5             | 35.2             | 21.2             | 13.8             | 16.6             | 36.4             | 15.4             | 8.9              | 0.8              | 41.2             |
| Qwen2.5-VL-32B                                        | 9.5              | 6.9              | 16.5             | 17.6             | 14.4             | 8.7              | 16.7             | 18.1             | 13.4             | 7.6              | 0.6              | 12.2             |
| Qwen2.5-VL-7B                                         | 9.3              | 7.4              | 16.5             | 30.2             | 15.5             | 11.0             | 16.6             | 41.1             | 7.8              | 5.3              | 0.6              | 12.3             |
| <i>Medical VLMs</i>                                   |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |
| Lingshu-32B                                           | 13.6             | 10.6             | 16.5             | 25.8             | 16.0             | 10.0             | 16.6             | 31.5             | 9.5              | 6.0              | 0.7              | 31.8             |
| Lingshu-7B                                            | 15.9             | 11.1             | 16.5             | 38.3             | 20.7             | 13.1             | 16.7             | 36.7             | 14.8             | 8.4              | 0.6              | 38.5             |
| MedGemma-27B                                          | 15.9             | 11.4             | 16.5             | 22.3             | 18.1             | 10.4             | 16.9             | 30.9             | 15.5             | 8.3              | 0.8              | 38.0             |
| MedGemma-4B                                           | 11.7             | 8.7              | 16.5             | 27.8             | 11.9             | 7.5              | 16.6             | 24.3             | 12.0             | 6.9              | 0.8              | 20.1             |
| HuatuoGPT-V-34B                                       | 18.0             | 12.4             | 16.5             | 48.6             | 19.7             | 12.9             | 16.6             | 34.3             | 16.2             | 8.9              | 0.7              | 35.3             |
| HuatuoGPT-V-7B                                        | 16.9             | 11.5             | 16.6             | 50.3             | 19.5             | 12.0             | 16.8             | 39.2             | 14.3             | 8.3              | 0.7              | 38.3             |
| Hulu-Med-7B                                           | 11.8             | 9.7              | 16.5             | 30.7             | 8.6              | 6.8              | 16.6             | 28.6             | 3.7              | 2.6              | 0.6              | 5.8              |
| MedVLM-R1-2B                                          | 14.3             | 10.5             | 16.5             | 30.3             | 15.0             | 9.8              | 16.6             | 29.6             | 11.2             | 6.8              | 0.6              | 24.1             |
| LLaVA-Med-7B                                          | 15.6             | 11.9             | 16.5             | 50.7             | 16.1             | 10.9             | 16.6             | 33.3             | 8.2              | 5.5              | 0.6              | 7.4              |
| M3D-LaMed-4B                                          | 3.0              | 2.6              | 16.6             | 35.6             | 2.6              | 2.0              | 16.6             | 35.6             | 1.5              | 1.2              | 0.6              | 58.8             |
| RadFM                                                 | 3.7              | 3.3              | 16.5             | 25.8             | 2.0              | 1.8              | 16.6             | 30.6             | 3.2              | 2.6              | 0.6              | 45.2             |
| <i>Ours (Agentic Multi-turn, base: Qwen2.5-VL-7B)</i> |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |                  |
| ActiveCTReasoner-SFT                                  | 44.3+35.0        | 34.9+27.5        | 64.9+48.4        | <b>67.8+37.6</b> | <b>46.0+30.5</b> | 33.4+22.4        | 60.2+43.6        | <b>65.0+23.9</b> | 35.3+27.5        | 21.5+16.2        | 38.0+37.4        | 57.2+44.9        |
| ActiveCTReasoner                                      | <b>45.4+36.1</b> | <b>37.2+29.8</b> | <b>70.8+54.3</b> | <u>66.4+36.2</u> | <u>45.4+29.9</u> | <b>34.2+23.2</b> | <b>66.1+49.5</b> | <u>64.6+23.5</u> | <b>37.8+30.0</b> | <b>23.8+18.5</b> | <b>54.2+53.6</b> | <b>62.8+50.5</b> |

#### Case B: Window-Aware Temporal Comparison

**Task:** ChangeClass (pericardial effusion, 5-day interval)

**Steps:** 4

1. scroll(95)  $\rightarrow$  Navigate to cardiac level
2. set\_window(mediastinum)  $\rightarrow$  Switch for soft-tissue contrast
3. zoom\_in(pericardial\_region)  $\rightarrow$  Magnified pericardial view
4. compare\_with\_prior(95)  $\rightarrow$  Side-by-side: current shows fluid absent in prior

**Answer:** “New Appearance” ✓ Demonstrates clinically motivated window selection before comparison.

#### Case C: Multi-Organ Quantitative Reasoning

**Task:** RegionReasoning (cardiovascular, 180-day interval)

**Steps:** 6

1. scroll(110)  $\rightarrow$  Navigate to cardiac level
2. zoom\_in(heart)  $\rightarrow$  Magnified cardiac view
3. compare\_with\_prior(110)  $\rightarrow$  Side-by-side temporal comparison
4. measure\_lesion(heart)  $\rightarrow$  Total cardiac volume: 394.2 $\rightarrow$ 585.3 cm<sup>3</sup> (+48.5%)
5. measure\_lesion(right\_atrium)  $\rightarrow$  8.4 $\rightarrow$ 5.0 cm<sup>3</sup> (−40.2%)
6. measure\_lesion(right\_ventricle)  $\rightarrow$  3.0 $\rightarrow$ 0.5 cm<sup>3</sup> (−82.2%)

**Reasoning:** The model identifies a contradictory pattern (overall cardiac enlargement but individual chamber shrinkage)

and correctly infers segmentation artifact rather than true structural remodeling. All 12 numerical values match ground truth.

#### Case D: Comprehensive Longitudinal Assessment

**Task:** DualSource (full report, 452-day interval)

**Steps:** 11

1. `scroll(80)` → Navigate to upper lung level
2. `set_window(lung)` → Lung window for parenchymal assessment
3. `compare_with_prior(80)` → Bilateral lung comparison
4. `measure_lesion(left_upper_lobe)` → LUL: −30.8%
5. `measure_lesion(left_lower_lobe)` → LLL: −38.5%
6. `measure_lesion(right_upper_lobe)` → RUL: −26.5%
7. `measure_lesion(right_middle_lobe)` → RML: −26.7%
8. `measure_lesion(right_lower_lobe)` → RLL: −34.3%
9. `set_window(mediastinum)` → Switch for cardiac structures
10. `measure_lesion(heart)` → Cardiac volume change
11. `measure_lesion(pericardial_effusion)` → Effusion quantification

**Answer:** Comprehensive longitudinal report covering 5 lung lobes, cardiac structures, and pericardial effusion, matching 20 numerical values with ground truth.

## I Data Privacy and Legality

All data employed in this work are obtained exclusively from publicly available open-source datasets. We do not collect, access, or incorporate any private user data. The datasets are released for research use, and our use of them follows their applicable terms, licenses, and intended scope.

For data annotation, all annotators involved in this work are internal members of the research team rather than externally recruited crowdworkers, students, or paid third-party participants. Therefore, the study does not involve participant recruitment through crowdsourcing platforms, demographic-dependent compensation, or any payment-related disputes. All annotation activities were conducted as part of the internal research workflow.

## J Ethical Consideration

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

## H LLMs Usage

We only use Large Language Models to polish our writing.