X-RAGENT: TOOL-AUGMENTED REINFORCEMENT LEARNING FOR EVIDENCE-GROUNDED CHEST X-RAY REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Chest X-ray (CXR) interpretation often requires integrating visual findings with task-specific evidence rather than relying on direct visual recognition alone. While tool-augmented medical agents can supply such evidence via external modules, they often lack explicit mechanisms for learning when auxiliary evidence is needed and for preventing inapplicable quantitative evidence from entering diagnostic reasoning. To address this, we propose **X-Ragent**, a post-training framework for tool-augmented reinforcement learning that decouples learned auxiliary-evidence seeking from validity-gated quantitative measurement acquisition for evidence-grounded CXR reasoning. We construct **Rad-CoT**, a 7.7K-sample corpus combining causally ordered tool-augmented trajectories with report-style diagnostic reasoning, and use it to initialize Qwen3-VL-8B via agentic supervised fine-tuning (SFT). We then optimize the policy with Group Relative Policy Optimization (GRPO) using an accuracy-gated reward that applies process-level terms for measurement consistency and auxiliary-tool cost only to correct trajectories, preventing plausible process behavior from compensating for an incorrect diagnosis. Across three CXR benchmarks, X-Ragent achieves the strongest results, including an 8.6-point improvement on the fully held-out ChestAgentBench. By seeking evidence selectively and admitting only valid measurements, X-Ragent supports more faithful and evidence-grounded CXR reasoning.

1 INTRODUCTION

Chest X-ray (CXR) is the most frequently ordered imaging examination, supporting clinical tasks from thoracic disease detection to postoperative monitoring (Yang et al., 2025). Automated CXR analysis has progressed from closed-set disease classification (Ma et al., 2025) to report generation (Jin et al., 2024; Liu et al., 2025b) and visual question answering (Hu et al., 2024), and is now moving toward open-ended interpretation with explanations (Chen et al., 2024; Liu et al., 2025a). However, reliable CXR reasoning hinges on localized and quantitative evidence, which current vision-language models still handle poorly. They struggle to ground diagnostic criteria in the relevant anatomy (Lee et al., 2025) and frequently hallucinate measurements (Heiman et al., 2025).

Recent advances make such evidence-grounded reasoning increasingly feasible. General and medical multimodal large language models (MLLMs) (Bai et al., 2025; Park et al., 2025) provide strong vision-language backbones, and GRPO-based reinforcement learning effectively elicits multi-step reasoning, including in CXR (Shao et al., 2024; Fan et al., 2025). Meanwhile, specialized models supply fine-grained visual evidence through anatomical segmentation (Seibold et al., 2023) and text-prompted lesion localization (Zhao et al., 2025), which tool-augmented agents compose into diagnostic workflows (Fallahpour et al., 2025; Li et al., 2024).

Despite these advances, three gaps remain when current approaches are applied to complex CXR diagnostic tasks: ❶ **Visual-Diagnostic Gap**: MLLMs globally perceive radiographs but struggle to link broad visual cues with localized, fine-grained anatomical evidence or quantitative metrics. This stems from a lack of structured cognitive mapping connecting perception to diagnostic hypotheses, leading to fragmented reasoning. ❷ **Evidence–Diagnosis Inconsistency**: Existing medical agents may invoke external tools but frequently ignore or misinterpret their objective outputs. Driven by

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

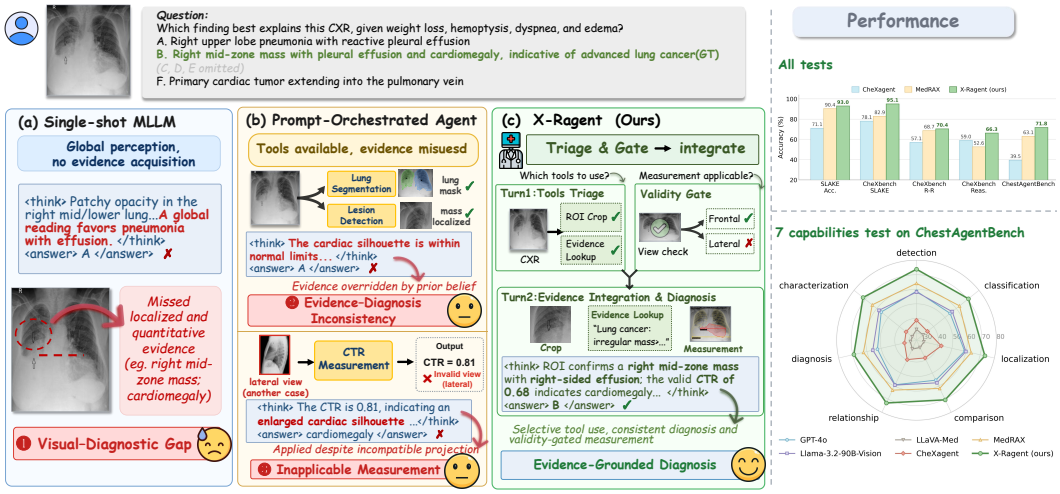

Figure 1: **Evidence-grounded CXR reasoning with X-Ragent.** (a) Single-shot MLLMs read the radiograph globally and miss localized evidence, leaving a visual–diagnostic gap. (b) Tool-augmented agents invoke tools but override the returned evidence with textual priors (top), or apply a measurement under incompatible acquisition conditions (bottom). (c) X-Ragent routes task-specific evidence on demand and integrates it with validity-gated quantitative measurements for evidence-grounded diagnosis.

textual priors rather than grounded evidence, they generate plausible-sounding but unsupported diagnoses, suffering from imprecise alignment between visual findings and semantic conclusions. ❸ **Inapplicable Measurement**: Deterministic measurements are meaningful only under their intended acquisition conditions. Without an explicit applicability check, inapplicable quantitative evidence can enter the reasoning process and mislead the final diagnosis.

To resolve these limitations, we argue that an effective CXR diagnostic agent requires three synergistic capabilities: ❶ **Evidence-Triage Planning**: The capability to assess whether auxiliary evidence is needed and selectively route task-specific evidence accordingly. Conditioned on the clinical query and initial radiographic assessment, the policy invokes visual refinement or condition-specific radiographic knowledge only when additional evidence is needed, alleviating the visual-diagnostic gap. ❷ **Faithful-Evidence Integration**: The capability to reason over heterogeneous tool observations and reconcile them with visual findings. This evidence-grounded process helps the agent resolve visual-semantic ambiguities rather than blindly trusting tool outputs, thereby mitigating cross-modal hallucination. ❸ **Applicability-Gated Measurement**: A validity gate that automatically provides basic CXR measurements only when acquisition conditions are satisfied, preventing incompatible quantitative evidence from entering diagnostic reasoning.

To realize these capabilities, we introduce **X-Ragent**, a post-training framework for evidence-grounded CXR reasoning. As shown in Fig. 1, X-Ragent selectively acquires task-relevant evidence, gates quantitative measurements by clinical applicability, and integrates the resulting evidence into diagnostic reasoning. First, we construct the **Rad-CoT** dataset for **supervised fine-tuning (SFT)**, comprising structured diagnostic trajectories that teach the policy when to invoke optional visual-refinement and knowledge-lookup tools and how to integrate their observations into diagnostic reasoning. Second, quantitative CXR evidence is supplied through a deterministic, validity-gated pathway: measurements such as the cardiothoracic ratio are exposed only when the applicability check succeeds. Finally, we optimize the policy with Group Relative Policy Optimization (GRPO) using a composite **gated reward** that prioritizes diagnostic correctness while encouraging selective auxiliary-tool use and faithful interpretation of valid quantitative evidence. This design decouples learned evidence seeking from validity-constrained measurement acquisition: X-Ragent decides *which* optional evidence to seek, while the validity gate determines *whether* quantitative measurements are admitted into reasoning. Our contributions are summarized as follows:

- **Evidence-Grounded CXR Reasoning.** We propose **X-Ragent**, a post-training framework that decouples learned auxiliary-evidence seeking from validity-gated quantitative measurement, enabling selective and evidence-grounded CXR reasoning.

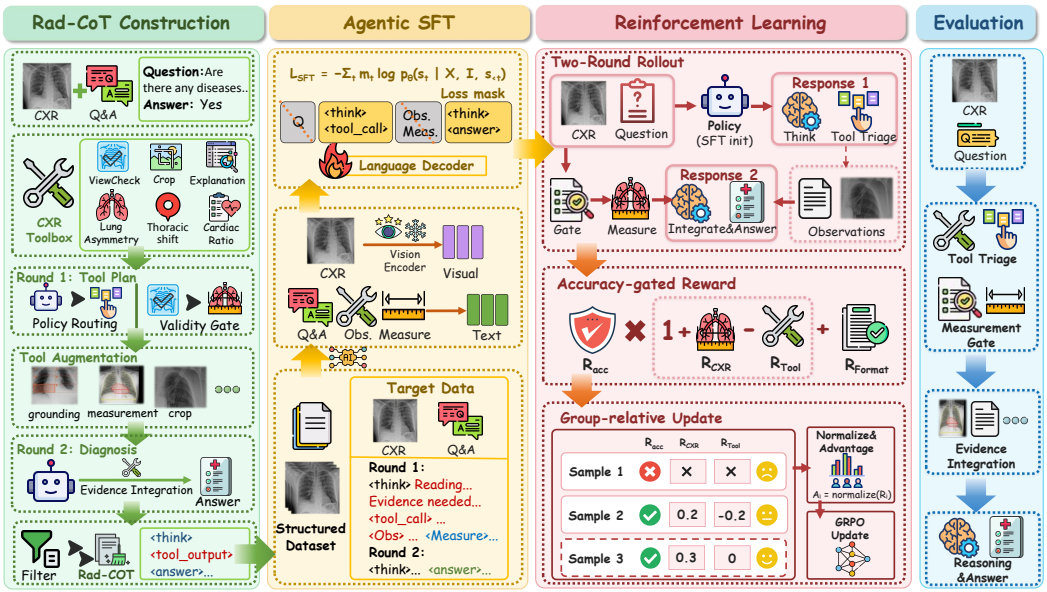

Figure 2: **Overview of X-Ragent.** (i) Rad-CoT constructs structured CXR reasoning data through evidence triage and gating, evidence integration, and filtering. (ii) Agentic SFT cold-starts Qwen3-VL-8B on Rad-CoT, masking injected observations from the language-modeling loss. (iii) GRPO optimizes the policy with accuracy and format supervision, while tool-use cost and radiographic evidence consistency are gated by answer correctness. (iv) At inference, the policy selectively routes auxiliary tools, while an independent validity gate decides whether quantitative measurements are exposed; both are integrated in the second turn.

- **Structured Agentic Post-Training.** We construct **Rad-CoT**, a 7.7K-sample dataset of structured diagnostic trajectories, and develop a two-stage SFT–GRPO training scheme with an accuracy-gated reward that promotes selective tool use and faithful interpretation of verifiable radiographic measurements.
- **Empirical Effectiveness.** X-Ragent achieves state-of-the-art performance across three CXR benchmarks, raising accuracy on the fully held-out **ChestAgentBench** from 63.1% to 71.7% (+8.6 points) over the strongest baseline, with consistent gains across all seven CXR interpretation capabilities.

2 Methodology

Overview. As illustrated in Fig. 2, we first formulate tool-augmented CXR reasoning (Sec. 2.1), then construct Rad-CoT (Sec. 2.2), and post-train the policy via agentic SFT (Sec. 2.3) and GRPO with a gated reward (Sec. 2.4).

2.1 Problem Formulation & Workflow Design

Problem Formulation. We formulate chest X-ray (CXR) interpretation as a tool-augmented visual reasoning process. Given a CXR image $I \in \mathbb{R}^{H\times W\times 3}$ and a clinical query Q, the model is required to generate a structured output $O = \langle \mathcal{R}, A \rangle$, where $\mathcal{R}$ denotes the reasoning trajectory including tool invocations and their observations, and A denotes the final answer. The answer may take the form of a multiple-choice selection, a short-answer response, or an image–text reasoning judgment, depending on the downstream task. Instead of directly mapping the image to a prediction, we instantiate Qwen3-VL-8B as an agentic policy π_θ. The policy selectively invokes auxiliary tools from an action set $\mathcal{T}_\pi$, while deterministic quantitative measurements are supplied through an independent validity-gated path.

Agentic Inference. X-Ragent performs diagnosis through a two-turn autoregressive process that separates policy-routed auxiliary acquisition from validity-gated quantitative measurement.

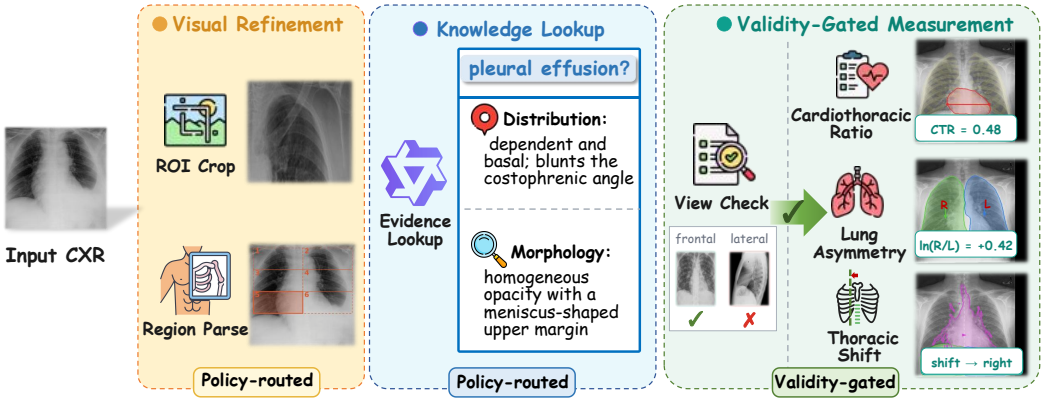

Figure 3: **X-Ragent tool suite.** The policy selectively invokes visual-refinement and evidence-lookup tools, while a separate validity-gated backend provides quantitative measurements.

Stage 1: Evidence Triage. Given an image I and query Q, the policy first performs a global radiographic assessment z_1 and selects an optional auxiliary-tool plan C, a possibly empty sequence of calls from $\mathcal{T}_\pi = \{\texttt{region_parse}, \texttt{roi_crop}, \texttt{evidence_lookup}\}$ with their arguments (e.g., target region or candidate conditions):

$$(z_1, C) \sim \pi_\theta(\cdot \mid I, Q; \mathcal{T}_\pi), \qquad C = (c_1, \ldots, c_K),\ c_k \in \mathcal{T}_\pi,\ 0 \le K \le K_{\max}, \tag{1}$$

where $K = 0$ when visual inspection alone suffices and $K_{\max} = 5$. Independently of this routing decision, a validity gate $g(I) \in \{0, 1\}$ exposes the quantitative measurement bundle only when the validity gate passes.

Stage 2: Evidence Integration and Diagnosis. The routed calls in C are executed and their observations are returned to the policy; when $g(I) = 1$, the backend additionally supplies the quantitative measurement bundle. Conditioned on its first-turn output, the policy integrates these observations with the original image and query to generate the reasoning trajectory $\mathcal{R}$ and final answer A:

$$O = \langle \mathcal{R}, A \rangle \sim \pi_\theta\left(\cdot \mid I \oplus F_C,\ Q \oplus z_1 \oplus C \oplus E_C \oplus M(I)\right), \tag{2}$$

where $\oplus$ denotes appending to the multimodal context, F_C and E_C denote visual and semantic feedback from the routed calls, respectively, and $M(I)$ denotes the quantitative measurement bundle, with $M(I) = \varnothing$ when $g(I) = 0$.

Tool Suite. As shown in Fig. 3, the policy can invoke three auxiliary tools: `roi_crop` magnifies relevant regions, `region_parse` supports localized anatomical inspection, and `evidence_lookup` retrieves condition-specific radiographic knowledge from a locally served Qwen3.5-27B. Separately, a deterministic backend computes the cardiothoracic ratio, lung-field asymmetry, and thoracic shift, exposing them only when `view_check` confirms applicability. Only these verifiable measurements receive explicit process supervision; further details are provided in Appendix C.

2.2 Training Data Construction

Existing MLLM training data mainly supervise final answers, without capturing how diagnostic evidence is acquired and integrated into reasoning. We therefore construct **Rad-CoT**, a 7,715-sample SFT corpus comprising 3,715 routed two-turn examples and 4,000 single-turn report-style CoT examples.

Data Collection. Rad-CoT combines PA radiographs from MIMIC-CXR-JPG (Johnson et al., 2024; Pollard et al., 2026) with structured findings derived from LLaVA-Rad (Zambrano Chaves et al., 2025; Pollard et al., 2026) annotations and publicly available medical VQA data. It contains 1,995 multiple-choice examples, 720 closed-ended VQA examples, 1,000 binary-choice VQA examples, and 4,000 report-style diagnostic CoT examples.

Generating Pipeline. We use locally deployed Qwen3.5-122B-A10B-FP8 to construct Rad-CoT reasoning data, with routed examples following the two-turn protocol in Sec. 2.1:

- **Evidence Triage and Acquisition.** For routed examples, the first turn performs an initial radiographic assessment and selects optional visual-refinement or knowledge-lookup actions, while quantitative measurements are exposed only when `view_check` verifies their applicability.
- **Evidence Integration.** The second turn reconciles the returned observations with the original radiograph to produce the final reasoning and answer. Observations are revealed only after the corresponding action is generated, preserving inference-time causal ordering.
- **Trajectory Filtering.** We filter malformed, leaky, or unverifiable generations; for the report-style subset, this reduces approximately 6,000 candidates to 4,000 valid examples. We also remove report-derived label leakage from multiple-choice trajectories.

Construction and quality-control details are provided in Appendix B.2.

2.3 Agentic Supervised Fine-Tuning

After constructing Rad-CoT, we adopt SFT as a cold start before GRPO. Prompt-elicited reasoning underperforms direct answering (Appendix B.3), while direct GRPO from the base policy also underperforms supervised initialization (Tab. 2). We therefore first use Rad-CoT to establish structured reasoning and tool-use behaviors before RL optimization.

Formally, each training instance is denoted as $\mathcal{T} = (\mathcal{X}, \mathcal{I}, \mathcal{S}, \mathcal{Y})$, where $\mathcal{X}$ denotes the visual input, $\mathcal{I}$ is the clinical task instruction, $\mathcal{S} = \{s_1, \ldots, s_T\}$ represents the policy-generated tokens spanning both the routing decision and the reasoning trajectory, and $\mathcal{Y}$ denotes the final answer. We apply a selective masking strategy: the loss is computed exclusively over policy-generated tokens, with injected observation spans masked out. The objective minimizes:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{\mathcal{T}\sim\mathcal{D}}\left[\sum_{t=1}^{T} \log p_\theta(s_t \mid \mathcal{X}, \mathcal{I}, s_{<t})\right], \tag{3}$$

where $p_\theta(\cdot)$ denotes the conditional probability distribution of the parameterized policy. This phase anchors auxiliary-tool routing and evidence-integrated diagnostic reasoning, providing a calibrated policy initialization for subsequent reinforcement learning.

2.4 Agentic Reinforcement Learning

Group Relative Policy Optimization (GRPO). We adopt the GRPO algorithm (Shao et al., 2024) for policy optimization. GRPO innovatively employs a groupwise comparison framework to evaluate candidate responses. Specifically, for each query q paired with its ground-truth solution a from dataset D, the algorithm generates a set of rollout trajectories $\{o_1, o_2, \ldots, o_G\}$ based on the previous policy $\pi_{\theta_{\text{old}}}$. The policy π_θ is then refined through the optimization of this objective function:

$$\begin{aligned}\mathcal{L}_{GRPO}(\theta) = -\mathbb{E}_{q\sim P(Q),\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{old}}(O|q)}\Bigg[\frac{1}{G}\sum_{i=1}^{G} \\ \left(\min\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}A_i, \text{clip}\left(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{old}}(o_i|q)}, 1-\epsilon, 1+\epsilon\right)A_i\right) - \beta\mathbb{D}_{KL}(\pi_\theta\|\pi_{ref})\right)\Bigg],\end{aligned} \tag{4}$$

$$\mathbb{D}_{KL}(\pi_\theta||\pi_{ref}) = \frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - \log\frac{\pi_{ref}(o_i|q)}{\pi_\theta(o_i|q)} - 1, \tag{5}$$

where β is adopted to balance the trade-off between exploration and stability during optimization. Then the advantage estimator A_i is calculated using normalized rewards from the trajectory group:

$$A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \ldots, r_G\})}{\text{std}(\{r_1, r_2, \ldots, r_G\})}. \tag{6}$$

Each trajectory o_i receives a composite reward through rule-based verification designed as follows.

Reward Design. Outcome-level rewards supervise final diagnostic correctness but provide limited guidance on whether acquired evidence is interpreted consistently. We therefore retain answer accuracy as the primary optimization target while introducing lightweight process-level signals for evidence consistency and auxiliary-tool efficiency. The reward comprises answer accuracy R_{acc},

format compliance R_{format}, radiographic evidence consistency R_{CXR}, and tool-use cost R_{tool}. $R_{\text{acc}} \in \{0, 1\}$ evaluates final-answer correctness, while $R_{\text{format}} \in \{0, -1\}$ penalizes malformed outputs.

R_{CXR} applies only to objective measurements whose applicability and interpretation can be deterministically verified. It rewards consistency with the query-relevant measurements exposed by the validity-gated backend, where query relevance affects reward eligibility only, not measurement execution. In parallel, R_{tool} counts policy-generated auxiliary calls in $\mathcal{T}_\pi$, excluding backend validity checks and deterministic measurement routines. Detailed definitions are provided in Appendix B.4. The total reward is

$$R(\tau) = R_{\text{format}}(\tau) + R_{\text{acc}}(\tau)\left[1 + \lambda_m R_{\text{CXR}}(\tau) - \lambda_t R_{\text{tool}}(\tau)\right]. \tag{7}$$

The multiplicative accuracy gate restricts the process-level terms to refining preferences among correct trajectories. Without this gate, an incorrect trajectory could still gain reward from superficially desirable process behavior, allowing process quality to compensate for diagnostic error. Visual-refinement and knowledge-lookup tools receive no explicit process supervision and are optimized indirectly through task success.

3 Experiment

Benchmarks and protocol. We evaluate X-Ragent on three CXR benchmarks, with further details provided in Appendix B.1. ChestAgentBench (Fallahpour et al., 2025) contains 2,500 six-choice questions spanning seven capabilities and is entirely excluded from both SFT and RL training, serving as an unseen benchmark. For CheXbench (Chen et al., 2024), following MedRAX (Fallahpour et al., 2025), we use 238 VQA questions from Rad-ReStruct (Pellegrini et al., 2023) and SLAKE and 380 fine-grained reasoning questions from OpenI (Demner-Fushman et al., 2016). SLAKE (Liu et al., 2021) provides 114 closed-ended CXR test questions, scored by exact match following M4CXR (Park et al., 2025). For CheXbench and SLAKE, training uses only the official training partitions of their source datasets, and all evaluation images are excluded from training (Appendix B.2).

Implementation Details. X-Ragent is initialized from Qwen3-VL-8B-Instruct and post-trained with SFT followed by GRPO. For SFT, we train on 7.7K samples for one epoch with a learning rate of 1×10^{-5} and an effective batch size of 16. For GRPO, we optimize the SFT checkpoint on 4.7K training examples for 300 steps, using 4 rollouts per example and at most 5 policy-routed auxiliary calls per trajectory. We use a learning rate of 2×10^{-6}, a rollout batch size of 16, and an actor update batch size of 8. Additional training details are provided in Appendix B.1.

3.1 Main Results

As shown in Table 1, X-Ragent achieves state-of-the-art performance across all three benchmarks, reaching 71.7% on ChestAgentBench, averaging 77.3% across the three CheXbench subsets, and 93.0% on SLAKE. We highlight three observations. ❶ **Strong performance across diverse CXR tasks.** X-Ragent improves over the best reported results by 8.6, 9.2, and 2.6 points on ChestAgentBench, CheXbench, and SLAKE, respectively. Relative to its Qwen3-VL-8B base model, post-training yields gains of 16.6, 11.8, and 20.2 points, enabling an 8B model to surpass substantially larger general-purpose VLMs and the GPT-4o-based MedRAX agent. On CheXbench, the gains come from the SLAKE (+19.5) and OpenI reasoning (+15.8) subsets, whereas on the 115-question Rad-Restruct subset X-Ragent matches its base model (70.4%). ❷ **Strong out-of-domain generalization.** ChestAgentBench is entirely excluded from post-training, yet X-Ragent outperforms the strongest baseline across all seven interpretation capabilities in Table 1 and improves over its base model by 14.6–18.3 points on every capability. This suggests that the learned reasoning behaviors transfer beyond the benchmarks represented during optimization. ❸ **Learned and selective auxiliary-evidence routing.** X-Ragent learns when to invoke optional visual-refinement and knowledge-lookup tools under GRPO, while quantitative measurements follow a separate validity-gated path. Tab. 4 shows that acquisition is conditioned on evidence need rather than applied by default. Sec. 3.2 further analyzes the contributions of tool-augmented training, evidence-consistency supervision, and applicability gating.

Table 1: **Main results on three CXR benchmarks.** Accuracy (%). CheXbench reports Rad-Restruct (R-R), SLAKE and OpenI reasoning (Reas.); ChestAgentBench is fully held out (OOD) and reported per capability. Following MedRAX, ChestAgentBench "All" is question-level accuracy over all 2,500 questions; since capability labels overlap, it is not the mean of the category columns. †Our base model, evaluated zero-shot with the direct-answer prompt (Appendix B.3). We denote **best** and second-best results, and highlight ours.

Model	CheXbench			SLAKE	ChestAgentBench (OOD)							
	R-R	SLAKE	Reas.	Acc.	Det.	Cls.	Loc.	Comp.	Rel.	Diag.	Char.	All
General-purpose VLMs												
GPT-4o	53.9	85.4	51.1	–	58.7	54.6	59.0	55.5	59.0	52.6	56.1	56.4
Llama-3.2-90B-Vision	62.6	74.0	49.2	–	58.1	56.5	59.9	57.5	59.3	55.9	58.0	57.9
Qwen3-VL-8B†	**70.4**	75.6	50.5	72.8	56.5	53.3	57.9	55.9	57.2	52.1	53.8	55.1
Medical VLMs												
LLaVA-Med	34.9	55.5	45.8	–	32.4	30.8	30.2	30.6	31.8	29.3	28.8	28.7
CheXagent	57.1	78.1	59.0	71.1	38.7	34.7	42.5	38.5	39.8	33.5	34.2	39.5
Agentic CXR systems												
MedRAX	68.7	82.9	52.6	90.4	64.1	62.9	63.6	61.8	63.1	62.5	64.0	63.1
X-Ragent (Ours)	**70.4**	**95.1**	**66.3**	**93.0**	**73.8**	**70.7**	**73.3**	**70.5**	**72.9**	**69.3**	**72.1**	**71.7**

Table 2: **Post-training pipeline ablation.** ChestA. is OOD; CheXb. and SLAKE are ID test splits. Avg. is the mean over the three benchmarks.

Method	ChestA.	CheXb.	SLAKE	Avg.
(a) Zero-shot	55.1	65.5	72.8	64.5
(b) RL only	59.8	68.4	78.9	69.0
(c) SFT only	64.4	72.0	89.5	75.3
(d) SFT+RL	67.4	73.2	92.1	77.6
(e) SFT+RL+Tools	**71.7**	**77.3**	**93.0**	**80.7**

Table 3: **Reward design ablation (Eq. 7).** R_{acc} is always used; rows (c)–(e) add process terms additively, while (f) gates them by R_{acc}.

	R_{fmt}	R_{tool}	R_{CXR}	Gate	ChestA.	CheXb.	SLAKE
(a)					67.2	73.8	90.4
(b)	✓				69.4	75.3	91.2
(c)	✓	✓			70.1	75.9	92.1
(d)	✓		✓		70.6	76.4	92.1
(e)	✓	✓	✓		69.8	75.6	91.2
(f)	✓	✓	✓	✓	**71.7**	**77.3**	**93.0**

3.2 Ablation Study

Training pipeline. In rows (a)–(e) of Tab. 2, we compare zero-shot, RL-only, SFT-only, and the two-stage model evaluated with and without tool access. SFT in row (c) yields the largest single gain (+10.8), improving both in-domain SLAKE (+16.7) and held-out ChestAgentBench (+9.3), whereas RL in row (d) adds more on held-out ChestAgentBench (+3.0) than on in-domain CheXbench (+1.2). This suggests that SFT establishes the core reasoning behavior, whereas RL mainly improves its transfer. RL-only training in row (b) falls 6.3 points below SFT-only, which we attribute to the difficulty of eliciting structured reasoning from reward alone, motivating our cold-start design. Enabling tools in row (e) adds 4.3 points on ChestAgentBench and 4.1 on CheXbench, but only 0.9 on SLAKE, where accuracy is already near saturation at 92.1% and the 114-question test set makes a single item worth 0.88%. The gains are therefore concentrated on the two benchmarks that involve multi-step and fine-grained reasoning, where external evidence is most likely to be informative.

Reward design. Tab. 3 isolates each reward component. R_{acc} alone performs worst on all three benchmarks. The tool-use cost (c) and the evidence reward (d) each improve over (b), but combining them without the gate (e) underperforms either alone: ungated, incorrect trajectories can still earn R_{CXR}, while the tool cost penalizes evidence acquisition regardless of outcome. Gating both terms by answer correctness (f) gives the best results on all three benchmarks, supporting the multiplicative form of Eq. 7.

Selective evidence acquisition. Tab. 4 stratifies questions by the correctness of direct-answer Qwen3-VL-8B, which we use as a proxy for evidence need. On ChestAgentBench, X-Ragent invokes auxiliary tools on 64.9% of questions missed by the base model, 21.3 percentage points higher

Table 4: **Selective auxiliary-tool routing.** Questions are stratified by base-model correctness; TIR (tool invocation rate) and Acc. are those of X-Ragent.

Benchmark	Base	n	TIR (%)	Acc. (%)
ChestA.	Wrong	1123	**64.9**	54.1
	Correct	1377	43.6	**86.0**
CheXb.	Wrong	252	**61.5**	54.4
	Correct	366	45.9	**85.5**

Table 5: **Ablation on the composition of Rad-CoT.** (b) strips tool calls and observations; (c)–(e) each remove one task type.

Corpus	ChestA.	CheXb.	SLAKE
(a) Full	**71.7**	**77.3**	**93.0**
(b) w/o tool traj.	67.2	71.2	86.0
(c) w/o report CoT	66.2	72.6	90.4
(d) w/o MCQ	69.8	74.9	92.1
(e) w/o VQA	70.4	74.2	82.5

Table 6: **Controlled ablation of acquisition-validity gating on ChestAgentBench.** The gated and no-gate variants use the same anatomy-measurement backend and differ only in whether view validity suppresses anatomy evidence. G-only / NG-only denote paired cases correct only with / without gating.

View group	N	No gate	Gated	Δ	G-only / NG-only
All	2500	71.44	**71.68**	+0.24	17 / 11
Frontal	1887	72.18	72.18	+0.00	8 / 8
Non-frontal / atyp.	613	69.17	**70.15**	+0.98	9 / 3

than the 43.6% invocation rate on questions it already answers correctly (1.49×). A similar pattern holds on CheXbench, where the invocation rate increases from 45.9% on base-correct questions to 61.5% on base-error questions (1.34×). Meanwhile, corrections substantially outnumber regressions, yielding rescue-to-damage ratios of 3.15× on ChestAgentBench and 2.58× on CheXbench. These results indicate that learned routing concentrates auxiliary acquisition on cases where the base policy is more likely to fail, rather than invoking tools uniformly.

Data composition. Tab. 5 removes one supervision component at a time while resampling the remainder to keep the corpus size fixed. Removing tool-trajectory supervision in row (b) lowers accuracy by 4.5, 6.1, and 7.0 points on ChestAgentBench, CheXbench, and SLAKE, respectively. Removing report-style CoT in row (c) causes the largest drop on the held-out ChestAgentBench (5.5 points), while removing VQA supervision in row (e) mainly affects SLAKE (-10.5 points), reflecting its closed-ended VQA format.

Controlled validity-gate ablation. On ChestAgentBench, `view_check` prevents measurement execution for 24.5% of inputs, with 31% ultimately withheld after the mask-quality check. In a controlled comparison with the same measurement backend (Tab. 6), validity gating raises accuracy on 613 non-frontal or atypical CXRs from 69.17% to 70.15% (+0.98 points), leaves 1,887 frontal CXRs unchanged (72.18%), and improves overall accuracy from 71.44% to 71.68% (+0.24 points). Paired prediction changes show 8 corrections and 8 regressions on frontal inputs, versus 9 corrections and 3 regressions on non-frontal or atypical inputs, so all six net corrections arise from the validity-sensitive subset. Fig. 4 provides a representative example: exposing an inapplicable lung-field asymmetry measurement flips the answer from the correct option C to the incorrect option A, whereas withholding it preserves the visually supported answer. Additional qualitative examples of auxiliary-evidence acquisition and integration are provided in Appendix B.6.

4 Related Work

Multimodal Vision–Language Reasoning. Multimodal large language models increasingly employ explicit intermediate reasoning. Multimodal-CoT separates rationale generation from answer related inference (Zhang et al., 2024), while LLaVA-CoT structures inference into summarization, visual interpretation, logical reasoning, and conclusion generation (Xu et al., 2025). VLM-R1 further shows that rule-based reinforcement learning improves visual reasoning and generalization over supervised fine-tuning (Shen et al., 2025). However, these methods reason over fixed multimodal inputs rather than acquiring task-specific evidence when needed.

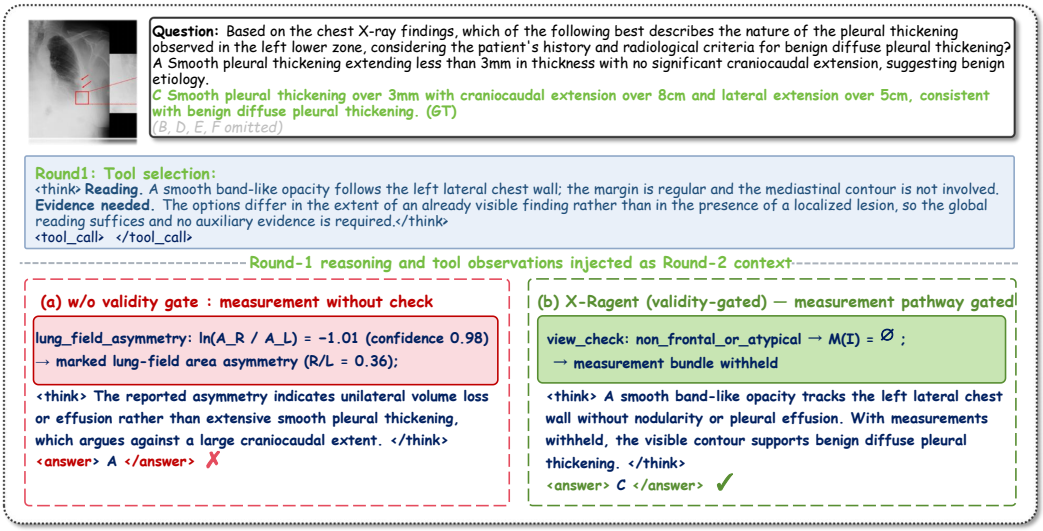

Figure 4: **Counterfactual case study of validity-gated quantitative evidence.** Without the validity gate, an inapplicable lung-field asymmetry measurement ($\ln(A_R/A_L) = -1.01$) is exposed to the policy and shifts the prediction from the correct option C to the incorrect option A. X-Ragent instead withholds the invalid measurement bundle and returns the correct answer based on the radiographic evidence.

Tool-Augmented Agentic Systems. Tool-augmented agents extend static reasoning into an interaction loop involving reasoning, action, and returned observations. ReAct interleaves reasoning with task-specific actions (Yao et al., 2023), Toolformer learns when and how to invoke external APIs from self-supervised examples (Schick et al., 2023), and HuggingGPT uses an LLM controller to plan tasks, select specialized models, and integrate their outputs across modalities (Shen et al., 2023). However, these general systems do not address clinical validity, acquisition-dependent applicability, or evidence integration for medical imaging tools.

Medical Vision–Language Reasoning and Agents. Medical VLMs have extended image interpretation toward instruction-driven reasoning, as exemplified by CheXagent for chest X-ray understanding (Chen et al., 2024). Tool-augmented systems further combine general reasoning with specialized medical models: MMedAgent learns to select tools across imaging modalities (Li et al., 2024), whereas MedRAX integrates classification, segmentation, and grounding modules within a ReAct-style CXR reasoning loop (Fallahpour et al., 2025). More recent methods such as ViTAR and Ophiuchus use supervised fine-tuning and reinforcement learning to train iterative visual or tool-mediated reasoning (Chen et al., 2025; Jiang et al., 2025). In contrast, our method learns a compact two-turn CXR interaction: the first turn performs preliminary interpretation and tool planning, while the second integrates returned observations into the final decision. Its reward jointly considers diagnostic correctness, tool cost, and consistency with verifiable radiographic measurements.

5 Conclusion

In this work, we propose X-Ragent, a post-training framework that combines learned evidence seeking with validity-gated quantitative augmentation for evidence-grounded CXR reasoning. Trained on structured diagnostic trajectories with agentic SFT and GRPO, X-Ragent learns when to acquire auxiliary evidence and how to integrate it, while admitting quantitative measurements only when clinically valid. X-Ragent achieves state-of-the-art performance across three CXR benchmarks, including an 8.6-point gain on the fully held-out ChestAgentBench and consistent improvements across all seven interpretation capabilities. These results highlight the value of selective evidence acquisition and validity-aware grounding for reliable and generalizable CXR reasoning.

AI USE STATEMENT

We clarify the role of Large Language Models (LLMs) in this work. During manuscript preparation, LLMs were used for translation, language polishing and providing pre-submission feedback on the manuscript. In the research pipeline, a locally deployed Qwen3.5-122B-A10B-FP8 was used to generate the synthetic Rad-CoT reasoning trajectories, which were filtered and verified by the authors; a locally deployed Qwen3.5-27B serves as the knowledge source of `evidence_lookup`. LLMs were not used to independently formulate research hypotheses, design methodological innovations, or determine scientific conclusions.

ETHICS STATEMENT

This work uses only de-identified, retrospective data; no new data were collected. MIMIC-CXR-JPG (Johnson et al., 2024) and LLaVA-Rad annotations (Zambrano Chaves et al., 2025) were accessed under the PhysioNet Credentialed Health Data Use Agreement, and other datasets are used under their respective licenses. Following PhysioNet's guidance on LLM use[1], all LLMs in our pipeline were deployed locally, so no MIMIC data were shared with third-party services; any release of Rad-CoT will be made only through PhysioNet. X-Ragent is a research prototype, not a medical device, and should not be used for clinical decision-making.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Isarun Chamveha, Treethep Promwiset, Trongtum Tongdee, Pairash Saiviroonporn, and Warasinee Chaisangmongkon. Automated cardiothoracic ratio calculation and cardiomegaly detection using deep learning approach, 2020. URL https://arxiv.org/abs/2002.07468.

Kaitao Chen, Shaohao Rui, Yankai Jiang, Jiamin Wu, Qihao Zheng, Chunfeng Song, Xiaosong Wang, Mu Zhou, and Mianxin Liu. Think twice to see more: Iterative visual reasoning in medical VLMs. *arXiv preprint arXiv:2510.10052*, 2025. doi: 10.48550/arXiv.2510.10052. URL https://arxiv.org/abs/2510.10052.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, Emily B. Tsai, Andrew Johnston, Cameron Olsen, Tanishq Mathew Abraham, Sergios Gatidis, Akshay S Chaudhari, and Curtis Langlotz. Chexagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*, 2024. URL https://arxiv.org/abs/2401.12208.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.

Adibvafa Fallahpour, Jun Ma, Alif Munim, Hongwei Lyu, and Bo Wang. MedRAX: Medical reasoning agent for chest x-ray. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 15661–15676. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/fallahpour25a.html.

Ziqing Fan, Cheng Liang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Chestx-reasoner: Advancing radiology foundation models with reasoning through step-by-step verification. *arXiv preprint arXiv:2504.20930*, 2025.

[1] https://physionet.org/news/post/llm-responsible-use/

Alice Heiman, Xiaoman Zhang, Emma Chen, Sung Eun Kim, and Pranav Rajpurkar. Factchexcker: Mitigating measurement hallucinations in chest x-ray report generation models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30787–30796. IEEE, 2025.

Xinyue Hu, Lin Gu, Kazuma Kobayashi, Liangchen Liu, Mengliang Zhang, Tatsuya Harada, Ronald M. Summers, and Yingying Zhu. Interpretable medical image visual question answering via multi-modal relationship graph learning. *Medical Image Analysis*, 97:103279, 2024. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2024.103279. URL https://www.sciencedirect.com/science/article/pii/S1361841524002044.

Yankai Jiang, Yujie Zhang, Peng Zhang, Wenjie Li, Yichen Li, Jintai Chen, Xiaoming Shi, and Shihui Zhen. Ophiuchus: Incentivizing tool-augmented "think with images" for joint medical segmentation, understanding and reasoning. *arXiv preprint arXiv:2512.14157*, 2025. doi: 10.48550/arXiv.2512.14157. URL https://arxiv.org/abs/2512.14157.

Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. Promptmrg: Diagnosis-driven prompts for medical report generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 2607–2615, 2024.

Alistair Johnson, Matthew Lungren, Yifan Peng, Zhiyong Lu, Roger Mark, Seth Berkowitz, and Steven Horng. MIMIC-CXR-JPG - chest radiographs with structured labels. *PhysioNet*, March 2024. doi: 10.13026/jsn5-t979. URL https://doi.org/10.13026/jsn5-t979. Version 2.1.0.

Hyungyung Lee, Geon Choi, Jung Oh Lee, Hangyul Yoon, Hyuk Gi Hong, and Edward Choi. Cxreasonbench: A benchmark for evaluating structured diagnostic reasoning in chest x-rays. Technical report, PhysioNet, 2025.

Binxu Li, Tiankai Yan, Yuanting Pan, Jie Luo, Ruiyang Ji, Jiayuan Ding, Zhe Xu, Shilong Liu, Haoyu Dong, Zihao Lin, and Yixin Wang. MMedAgent: Learning to use medical tools with multi-modal agent. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8745–8760, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.510. URL https://aclanthology.org/2024.findings-emnlp.510/.

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

Bo Liu, Ke Zou, Li-Ming Zhan, Zexin Lu, Xiaoyu Dong, Yidi Chen, Chengqiang Xie, Jiannong Cao, Xiao-Ming Wu, and Huazhu Fu. Gemex: A large-scale, groundable, and explainable medical vqa benchmark for chest x-ray diagnosis. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21310–21320. IEEE, 2025a.

Kang Liu, Zhuoqi Ma, Xiaolu Kang, Yunan Li, Kun Xie, Zhicheng Jiao, and Qiguang Miao. Enhanced contrastive learning with multi-view longitudinal data for chest x-ray report generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10348–10359. IEEE, 2025b.

DongAo Ma, Jiaxuan Pang, Michael B Gotway, and Jianming Liang. A fully open ai foundation model applied to chest radiography. *Nature*, 643(8071):488–498, 2025.

Jonggwon Park, Soobum Kim, Byungmu Yoon, Jihun Hyun, and Kyoyun Choi. M4cxr: exploring multitask potentials of multimodal large language models for chest x-ray interpretation. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

Chantal Pellegrini, Matthias Keicher, Ege Özsoy, and Nassir Navab. Rad-restruct: A novel vqa benchmark and method for structured radiology reporting. In *International conference on medical image computing and computer-assisted intervention*, pp. 409–419. Springer, 2023.

Tom Pollard, Benjamin E. Moody, Li-wei H. Lehman, Brian J. Gow, Chrystinne Fernandes, Chen Xie, Alistair Johnson, Roger G. Mark, and Thomas Heldt. PhysioNet as a global platform for biomedical research. *Nature Health*, 1(8):792–795, August 2026. ISSN 3005-0693. doi: 10.1038/s44360-026-00096-z. URL https://doi.org/10.1038/s44360-026-00096-z.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36, pp. 68539–68551. Curran Associates, Inc., 2023. doi: 10.52202/075280-2997. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/d842425e4bf79ba039352da0f658a906-Abstract-Conference.html.

Constantin Seibold, Alexander Jaus, Matthias A Fink, Moon Kim, Simon Reiß, Ken Herrmann, Jens Kleesiek, and Rainer Stiefelhagen. Accurate fine-grained segmentation of human anatomy in radiographs via volumetric pseudo-labeling. *arXiv preprint arXiv:2306.03934*, 2023.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. VLM-R1: A stable and generalizable R1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025. doi: 10.48550/arXiv.2504.07615. URL https://arxiv.org/abs/2504.07615.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. HuggingGPT: Solving AI tasks with ChatGPT and its friends in Hugging Face. *arXiv preprint arXiv:2303.17580*, 2023. doi: 10.48550/arXiv.2303.17580. URL https://arxiv.org/abs/2303.17580.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. LLaVA-CoT: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2087–2098, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Xu_LLaVA-CoT_Let_Vision_Language_Models_Reason_Step-by-Step_ICCV_2025_paper.html.

Zefan Yang, Xuanang Xu, Jiajin Zhang, Ge Wang, Mannudeep K Kalra, and Pingkun Yan. Chest x-ray foundation model with global and local representations integration. *IEEE transactions on medical imaging*, 44(12):4787–4799, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, Hany Awadalla, Julia Gong, Houdong Hu, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Yu Gu, Cliff Wong, Mu-Hsin Wei, Tristan Naumann, Muhao Chen, Matthew Lungren, Akshay Chaudhari, Serena Yeung, Curtis Langlotz, Sheng Wang, and Hoifung Poon. LLaVA-Rad MIMIC-CXR Annotations. *PhysioNet*, January 2025. doi: 10.13026/4ma4-k740. URL https://doi.org/10.13026/4ma4-k740. Version 1.0.0.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=y1pPWFVfvR.

Theodore Zhao, Yu Gu, Jianwei Yang, Naoto Usuyama, Ho Hin Lee, Sid Kiblawi, Tristan Naumann, Jianfeng Gao, Angela Crabtree, Jacob Abel, et al. A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. *Nature methods*, 22(1):166–176, 2025.

A Appendix

In this appendix, we provide more experimental details, related work, tool suite, and discussions for a comprehensive evaluation and understanding of our method. Detailed contents are as follows:

B Experiment **13**

B.1 More Experimental Details 13

B.2 Rad-CoT Construction and Quality Control 14

B.3 Effect of Prompt-Elicited Reasoning 16

B.4 Reward Details 16

B.5 More Ablation Studies 17

B.6 More Case Studies 17

C Tool Suite Details **18**

C.1 Visual Refinement Tools 19

C.2 Auxiliary Interpretation via Evidence Lookup 19

C.3 Validity-Gated Objective Measurements 19

C.4 Tool Execution and Observation Injection 20

D Prompts **20**

D.1 Prompts for Base-Model Reasoning Analysis 20

B Experiment

B.1 More Experimental Details

Datasets. ChestAgentBench contains 2,500 six-choice questions. Each question is associated with four category labels, including a reasoning label shared by all questions. Following the benchmark protocol, we report seven non-reasoning capability accuracies: detection (972), classification (1,517), localization (983), comparison (1,026), relationship (997), diagnosis (1,024), and characterization (981). These capability subsets overlap at the question level and are therefore not mutually exclusive.

CheXbench evaluation follows the MedRAX-style axis-2 subset and contains 618 two-choice questions from three sources: 380 OpenI fine-grained reasoning questions, 115 Rad-ReStruct VQA questions, and 123 SLAKE VQA questions. The standalone SLAKE evaluation contains 114 closed-ended chest-X-ray questions and is distinct from the SLAKE subset used within CheXbench.

Evaluation metrics. For ChestAgentBench, we report accuracy for each of the seven overlapping capability subsets, and report the overall score as question-level accuracy over all 2,500 questions. For CheXbench, we report source-specific accuracy and compute the overall score as the unweighted macro average across OpenI, Rad-ReStruct, and SLAKE. For standalone SLAKE, we report exact-match accuracy under the benchmark's closed-form answer protocol on the 114-question test set.

SFT configuration. We initialize from Qwen3-VL-8B-Instruct and perform one epoch of supervised fine-tuning with a learning rate of 1×10^{-5} and an effective batch size of 16. The maximum sequence length is 8,192 tokens. The vision encoder is kept frozen during SFT, while the language policy is optimized to learn both evidence routing and evidence-integrated reasoning.

Hyperparameters. Tab. 7 lists the key hyperparameters. Since $R_{\text{CXR}} \geq 0$ and $\lambda_t K_{\max} < 1$, any well-formed correct trajectory receives a higher reward than any incorrect one under Eq. 7, so tool cost never overrides diagnostic correctness.

Table 7: **Key hyperparameters of X-Ragent.**

Stage	Hyperparameter	Value
SFT	Base model	Qwen3-VL-8B-Instruct
	Trainable modules	LLM (vision encoder frozen)
	Epochs / learning rate	1 / 1×10^{-5}
	Effective batch size	16
	Max sequence length	8,192
	LR schedule / warmup ratio	cosine / 0.1
GRPO	Training prompts / steps	4,715 / 300
	Learning rate	2×10^{-6}
	Rollouts per prompt G	4
	Rollout / actor update batch size	16 / 8
	Max policy-routed calls $K_{\max}$	5
	Clip ratio ϵ / KL coefficient β	0.2 / 0.01
	Reward weights λ_m / λ_t	0.5 / 0.05
	Rollout temperature	1.0
	Max response length (per turn)	2,048
Inference	Decoding	greedy
	Max new tokens (per turn)	2,048

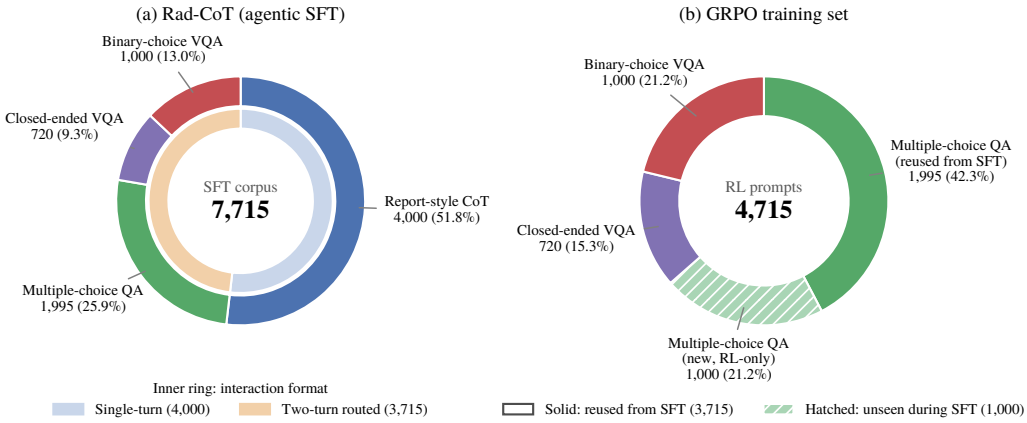

Figure 5: **Composition of the training data.** (a) The SFT corpus Rad-CoT. The outer ring shows the task type; the inner ring distinguishes single-turn report-style CoT from two-turn routed examples, in which the first-turn tool plan may be empty. (b) The GRPO corpus. Solid segments are reused from the routed SFT examples, while the hatched segment denotes multiple-choice questions unseen during SFT.

B.2 Rad-CoT Construction and Quality Control

Corpus composition. Rad-CoT contains 7,715 SFT instances, of which 3,715 follow the two-turn routed protocol and 4,000 are single-turn report-style CoT examples. The GRPO corpus contains 4,715 examples: all 3,715 routed SFT examples are reused, and 1,000 additional multiple-choice questions unseen during SFT are added. Report-style CoT is used only for SFT, since it lacks a rule-verifiable answer. Fig. 5 visualizes the composition of both corpora, and Tab. 8 lists the exact counts together with the interaction format and reward type of each subset. We use "7.7K" and "4.7K" in the main text as rounded descriptions of the two corpora.

Report-style reasoning distillation. The report-style subset is constructed from PA radiographs in MIMIC-CXR-JPG paired with reference findings derived from the LLaVA-Rad annotations. We first stratify 6,000 candidate studies and use locally deployed Qwen3.5-122B-A10B-FP8 to synthesize the intermediate diagnostic rationale conditioned on both the radiograph and its reference findings. Importantly, the teacher is used to generate the reasoning process rather than to freely

Table 8: **Subset statistics of the SFT and GRPO corpora.** "Routed" denotes the two-turn protocol in Sec. 2.1, where the first-turn tool plan may be empty. GRPO uses only subsets with rule-verifiable answers; all GRPO rewards are binary.

Subset	Format	Answer / reward	SFT	GRPO
Report-style diagnostic CoT	Single-turn	Free-text findings	4,000	–
Multiple-choice QA	Routed	Option-letter match	1,995	1,995
Multiple-choice QA (RL-only)	Routed	Option-letter match	–	1,000
Closed-ended VQA	Routed	Normalized exact match	720	720
Binary-choice VQA	Routed	Option-letter match	1,000	1,000
Total			**7,715**	**4,715**

regenerate the diagnostic target: after generation, the reference findings are retained as the target answer. This answer-conditioned construction reduces target drift while providing explicit supervision for mapping radiographic evidence to diagnostic conclusions.

Reasoning-quality filtering. We apply deterministic quality control to the generated report-style trajectories. Before generation, we retain PA-view studies with an available image and non-trivial reference findings. After generation, we remove prompt or metadata leakage, reject malformed or abnormal-length reasoning, and discard rationales that excessively reproduce the reference findings. In particular, candidate rationales are required to contain 40–900 words and reasoning–finding similarity above 0.82 is rejected. When multiple reasoning segments are produced, we retain the longest valid segment. Failed generations are retried once and discarded if they remain invalid. This filtering reduces the initial 6,000 candidates to 4,000 report-style reasoning trajectories.

Leakage-controlled MCQ construction and source disjointness. All 2,995 training MCQs (1,995 for SFT and 1,000 RL-only) are generated from MIMIC-CXR-JPG PA radiographs (Johnson et al., 2024). We remove report-derived cues unavailable at inference, such as sentences referring to hidden findings (e.g., "according to the report"), and strip findings metadata from the final records, so that answers must rely on the radiograph and query. ChestAgentBench (Fallahpour et al., 2025) is instead built from Eurorad cases, sharing no images, cases, or reports with our training data, and none of its questions or images is used during MCQ generation.

Closed-ended VQA construction. We include 720 chest-X-ray closed-ended VQA trajectories to preserve concise answering behavior. Images overlapping with the SLAKE evaluation split are removed during construction, preventing direct image overlap between this training subset and the corresponding held-out evaluation data.

Causally ordered tool trajectories. The 3,715 routed SFT examples, consisting of 1,995 MCQ, 720 closed-ended VQA, and 1,000 binary-choice VQA examples, follow the two-turn interaction protocol. In the first turn, the policy observes the radiograph and clinical query and produces an evidence-acquisition decision. Only after this decision is generated are the selected tool observations inserted into the conversation. The second assistant turn then integrates the returned evidence with the original image and query to produce the final reasoning and answer. Thus, tool observations never precede the actions that request them, matching the causal interaction structure used at inference time.

During SFT, loss is applied only to policy-generated assistant tokens, whereas externally injected tool observations are treated as context and masked from the language-modeling objective. This prevents the policy from being trained to imitate tool outputs and instead teaches it to decide when to acquire evidence and how to use the returned evidence.

Distribution and integrity checks. Finally, we use CheXbert to inspect the diagnostic-label distribution of the report-style subset and verify coverage across the 14 standard CXR categories. Together with leakage removal, evaluation-image deduplication, and generation-time filtering, these checks are intended to ensure that gains from Rad-CoT reflect learned radiographic reasoning and evidence integration rather than target leakage or duplicated evaluation content.

Evaluation-split integrity. For the CheXbench axis-2 evaluation, we construct the exclusion set from the images actually used by the evaluated subsets rather than excluding the entire CheXbench

Table 9: **Effect of prompt-elicited reasoning on the Qwen3-VL-8B base model.** We compare a direct-answer prompt with four variants that explicitly require intermediate reasoning. Accuracy (%) is reported using the same benchmark-level aggregation protocol as in the main experiments.

Prompting Strategy	ChestAgent	CheXbench	SLAKE	Avg.
Direct answering (v4)	**55.1**	**65.5**	**72.8**	**64.5**
Reasoning v0	47.6	56.7	59.6	54.6
Reasoning v1	47.3	56.7	45.6	49.9
Reasoning v2	51.1	57.3	62.3	56.9
Reasoning v3	50.3	60.1	59.6	56.7

image index. The OpenI fine-grained reasoning subset contains 363 unique image references, while the Rad-ReStruct evaluation contributes 94 OpenI-backed image references; their union contains 452 unique evaluation image references. We exclude this evaluation-specific image set from training and additionally check for OpenI report-level leakage through sibling images. For the final training set, both image-level overlap with this 452-image exclusion set and OpenI report-level sibling-image leakage are zero.

B.3 Effect of Prompt-Elicited Reasoning

We examine whether explicit diagnostic reasoning can be reliably elicited from the Qwen3-VL-8B base model through prompting alone, without any post-training. We compare four reasoning-oriented prompts (v0–v3), all of which explicitly require a `<think>...</think>` reasoning trace, against a direct-answer prompt (v4) that does not require an explicit reasoning block. All settings use the same base checkpoint, image–question inputs, decoding configuration, and evaluation protocol; only the system prompt is changed.

For consistency with the main experiments, ChestAgentBench accuracy is reported as question-level accuracy over all 2,500 questions, CheXbench accuracy as the unweighted mean across its three data sources, and SLAKE using closed-ended exact-match accuracy.

As shown in Tab. 9, explicitly requiring a reasoning trace does not improve the base model. Direct answering achieves 55.1%, 65.5%, and 72.8% on ChestAgentBench, CheXbench, and SLAKE, respectively. Even the strongest explicit-reasoning variant on each benchmark remains lower by 4.0, 5.4, and 10.5 percentage points, respectively.

These results indicate that simply eliciting longer or more explicit reasoning from the base model is insufficient to establish reliable diagnostic reasoning behavior. Together with the RL-only result in Tab. 2, where direct GRPO from the base policy remains substantially weaker than supervised initialization, this motivates our two-stage post-training strategy. We therefore first use Rad-CoT to initialize structured diagnostic reasoning and tool-use behaviors through SFT, and subsequently refine the policy with GRPO.

B.4 Reward Details

Radiographic evidence consistency. For measurement k, let $r_k \in \{0, 1\}$ indicate query relevance under a fixed measurement–task correspondence and $b_k \in \{0, 1\}$ indicate whether the measurement is exposed by the validity-gated backend. We define its reward eligibility as $a_k = r_k b_k$, where query relevance affects reward eligibility only and does not control measurement execution. The evidence-consistency reward is $R_{\mathrm{CXR}}(\tau) = \frac{\sum_k a_k s_k}{\max(1, \sum_k a_k)}$, where s_k is a rule-based score indicating whether the trajectory interprets measurement k consistently with its verified value. Each measurement contributes at most once.

Auxiliary tool-use cost. R_{tool} counts policy-generated auxiliary calls from $\mathcal{T}_\pi$, with repeated calls counted separately. Backend validity checks and deterministic measurement routines are excluded from this cost.

Table 10: **Ablation on evidence-acquisition modules at inference.** A single trained checkpoint is evaluated with individual modules selectively disabled. The validity-gated measurement backend is retained unless objective measurement is explicitly removed; row (f) statically invokes all policy-callable auxiliary tools on every question.

	Evidence acquisition	ChestAgent
(a)	**Full (learned routing)**	**71.7**
(b)	w/o visual enh.	70.1
(c)	w/o aux. interp.	69.5
(d)	w/o obj. meas.	70.8
(e)	w/o all evidence modules	67.4
(f)	All auxiliary tools (static)	71.0

Table 11: **Ablation on quantitative measurement routing.** We compare the default X-Ragent design, where quantitative measurements are supplied by an independent validity-gated backend, with a variant that additionally requires the first-stage policy to select the measurement tool before validity checking. Accuracy (%) is reported on ChestAgentBench and CheXbench.

Measurement pathway	ChestAgent	CheXbench
Independent validity gate (X-Ragent)	**71.7**	**77.3**
Policy routing + validity gate	71.0	76.1

B.5 More Ablation Studies

Module contribution. Tab. 10 disables individual evidence-acquisition modules at inference using a single trained checkpoint; row (e) is equivalent to row (d) of Tab. 2. With all modules removed, accuracy drops by 4.3 points to 67.4%, indicating that part of the gain is internalized during post-training. Auxiliary interpretation contributes most (2.2 points), followed by visual enhancement (1.6) and objective measurement (0.9), whose narrower scope and 31% withholding rate limit its effect. Forcing all auxiliary tools on every question in row (f) remains 0.7 points below learned routing, which invokes tools on only 53.2% of questions (Tab. 4), showing that selective routing loses no accuracy while avoiding unnecessary calls.

Quantitative measurement routing. We further examine whether quantitative measurements should be routed by the first-stage policy. In the default X-Ragent design, quantitative measurement is decoupled from policy-routed auxiliary evidence: the validity-gated backend is applied independently to each input, and measurements are exposed whenever their acquisition conditions are satisfied. We compare this design with a variant in which the measurement tool is included in the first-stage action space and is executed only when explicitly selected by the policy.

As shown in Tab. 11, the default design improves accuracy from 71.0% to 71.7% on ChestAgentBench and from 76.1% to 77.3% on CheXbench. The policy-routed variant is close to removing measurement entirely (70.8%, Tab. 10(d)), suggesting that routing tends to omit these cues even when informative. Since cardiac size, lung-field asymmetry, and thoracic shift are cheap, verifiable, and complementary to the MLLM's global visual representation, we supply them directly whenever acquisition conditions are satisfied, while the policy selectively acquires task-specific auxiliary evidence.

B.6 More Case Studies

Selective acquisition and evidence integration. Fig. 6 provides an additional example of the learned auxiliary-evidence pathway. Because the answer options differ in nodule distribution and cavitation, the first-turn policy determines that localized inspection is needed and requests a focused ROI via `region_parse` and `roi_crop`, together with condition-specific criteria via `evidence_lookup`. In parallel, the validity gate, which runs independently of the policy's routing decision, rejects the input and withholds the quantitative measurement bundle; had it been exposed, the ungated CTR would have been 1.79, a physically implausible value. The second turn then integrates the localized visual findings with the retrieved radiographic criteria. The absent measure-

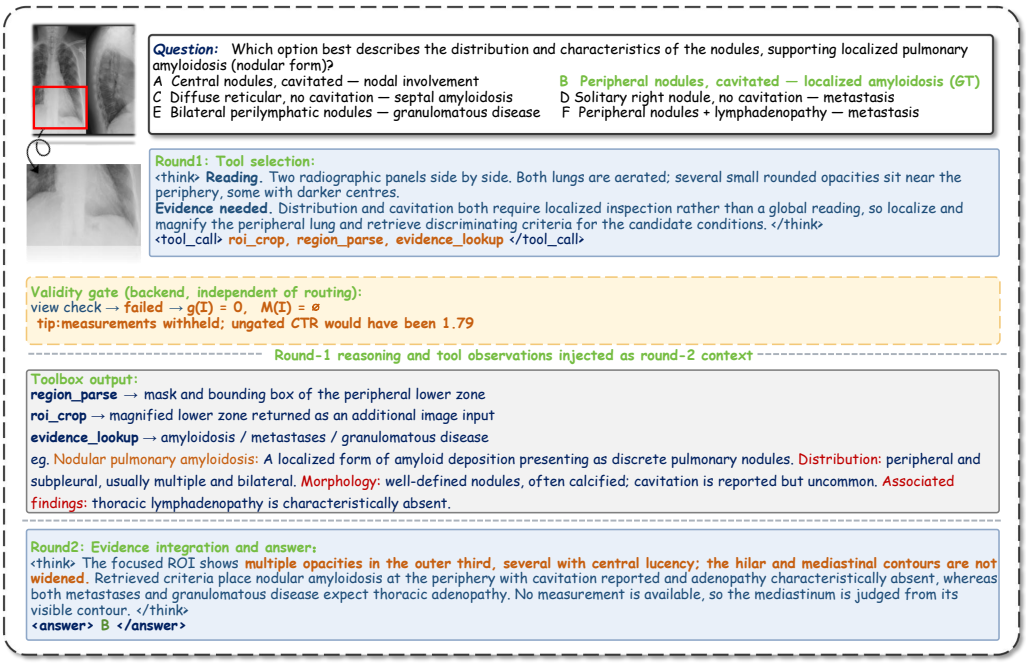

Figure 6: **Selective auxiliary-evidence acquisition and integration.** In the first turn, X-Ragent judges global inspection insufficient and routes localized visual refinement (`region_parse` → `roi_crop`) together with condition-specific `evidence_lookup`. Independently of this routing decision, the validity gate rejects the atypical input and withholds the measurement bundle ($M(I) = \varnothing$), preventing a physically implausible CTR (1.79 if ungated) from entering the reasoning. In the second turn, the policy integrates the magnified ROI with the retrieved criteria, identifying peripheral cavitated nodules without thoracic adenopathy, and selects the correct answer, localized nodular pulmonary amyloidosis.

Table 12: **Summary of the X-Ragent evidence-acquisition modules.** Policy-routed tools are selected by the learned policy, whereas the validity check and quantitative measurements are executed automatically by the backend.

Component	Execution	Backbone	Returned evidence
`region_parse`	Policy-routed	CXAS / BiomedParse	Region mask and bounding box
`roi_crop`	Policy-routed	Geometric crop	Local high-resolution view
`evidence_lookup`	Policy-routed	Qwen3.5-27B (local API)	Condition-specific radiographic knowledge
`view_check`	Backend	OpenCXR	Acquisition applicability
Measurement bundle	Backend	CXAS + deterministic rules	CTR, lung asymmetry, thoracic shift

ment is treated as unavailable rather than normal, and the mediastinum is assessed from its visible contour. The policy correctly identifies localized nodular pulmonary amyloidosis.

C Tool Suite Details

X-Ragent separates evidence acquisition into two operationally distinct pathways. First, the policy selectively invokes auxiliary tools from $\mathcal{T}_\pi = \{\texttt{roi_crop}, \texttt{region_parse}, \texttt{evidence_lookup}\}$ to acquire additional visual or interpretive evidence. These actions are generated by the policy and are therefore counted in R_{tool}. Second, quantitative radiographic measurements are computed by an independent backend whose execution is controlled by an acquisition-validity check. Backend validity checks and deterministic measurement routines are not policy actions and are excluded from R_{tool}.

C.1 Visual Refinement Tools

Region Parsing. `region_parse` provides localized anatomical or pathological evidence for region-specific CXR inspection. Following the specialized parsing tools used in prior medical agents (Jiang et al., 2025), chest anatomical structures are parsed using CXAS (Seibold et al., 2023), while text-specified pathological regions can be localized using BiomedParse (Zhao et al., 2025). Given the input radiograph and a target structure or finding, the parser returns a segmentation mask together with its tight axis-aligned bounding box. These outputs provide explicit region-level grounding and can subsequently be used for fine-grained visual inspection.

ROI Crop. `roi_crop` is a deterministic visual-refinement operator that magnifies a region identified by `region_parse`. Specifically, given the returned bounding box $b = (x_1, y_1, x_2, y_2)$, the corresponding region is cropped from the original radiograph,

$$I_{\text{crop}} = I[y_1 : y_2,\ x_1 : x_2], \tag{8}$$

and returned as an additional visual observation. The original CXR is retained in the multimodal context, allowing the policy to jointly reason over the global image and the localized high-resolution view. This parse–then–crop design follows the general segment/localize–then–zoom strategy used in tool-augmented visual reasoning (Jiang et al., 2025).

C.2 Auxiliary Interpretation via Evidence Lookup

To mitigate cross-modal hallucination, `evidence_lookup` queries a locally deployed Qwen3.5-27B through an OpenAI-compatible API, invoked online (i.e., on demand during inference). Given the candidate conditions specified by the policy, it returns a concise description of their characteristic CXR manifestations, including typical appearance, distribution, and discriminating imaging features. This tool supplies general condition-level knowledge rather than a patient-specific diagnosis; the returned description is appended to the query as E_C and reconciled with the visual findings by the policy in the second turn.

C.3 Validity-Gated Objective Measurements

Validity Gate. Unlike the auxiliary tools above, `view_check` is executed automatically and independently of the policy routing decision. We use the Image Sorter from OpenCXR[2], which distinguishes frontal PA, frontal AP, lateral, and non-CXR images. Because the quantitative measurements used in X-Ragent are defined for standard PA radiographs, the measurement pathway is enabled only when the image satisfies the required PA acquisition condition. Otherwise, the quantitative measurement bundle is withheld.

We additionally suppress the bundle when the anatomical masks required for measurement are unavailable or fail the backend quality checks. Thus, the validity gate can be written as

$$g(I) = g_{\text{view}}(I)\, g_{\text{quality}}(I), \tag{9}$$

and quantitative evidence is exposed only when $g(I) = 1$. Neither `view_check` nor the subsequent measurement routines contributes to R_{tool}.

Measurement Bundle. For an applicable input, the backend uses CXAS-derived anatomical masks to compute three deterministic radiographic measurements: cardiothoracic ratio (CTR), lung-field asymmetry, and thoracic shift.

Cardiothoracic ratio. Following mask-based CTR estimation (Chamveha et al., 2020), we identify the image row where the heart mask attains its maximum horizontal width. Let w_c denote the cardiac width on this row and w_t the horizontal span of the bilateral lung fields at the same level. The cardiothoracic ratio is

$$\text{CTR} = \frac{w_c}{w_t}. \tag{10}$$

On a valid PA radiograph, $\text{CTR} > 0.5$ is interpreted as an enlarged cardiac silhouette.

[2] https://github.com/DIAGNijmegen/opencxr

Lung-field asymmetry. Let A_L and A_R denote the projected areas of the left and right lung masks, respectively. We compute a log-ratio bilateral asymmetry index as

$$\mathrm{LFA} = \ln \frac{A_R}{A_L}, \tag{11}$$

where the magnitude reflects the degree of lung-field asymmetry and the sign indicates its direction.

Thoracic shift. Thoracic shift is estimated geometrically from CXAS-derived central thoracic structures relative to the bilateral thoracic midline. The result records both the magnitude and direction of displacement and is used as objective evidence for queries involving mediastinal or tracheal displacement.

The resulting measurements are converted to fixed categorical interpretations by deterministic rules and returned to the policy as a structured observation. These verified interpretations are also used to compute the evidence-consistency score s_k in R_{CXR} (Appendix B.4).

If the applicability check fails, we define

$$M(I) = \varnothing, \tag{12}$$

so that no potentially unreliable quantitative value enters the reasoning context. Importantly, an absent measurement therefore denotes *unavailable or inapplicable evidence*, rather than a normal radiographic finding.

C.4 Tool Execution and Observation Injection

For each example, policy-routed tools are executed only after their corresponding actions have been generated. Returned crops, parsing results, and knowledge-lookup observations are then injected into the second-turn context together with any valid quantitative measurements. During SFT, externally generated observation spans are masked from the language-modeling loss, such that the policy is optimized to learn tool selection and evidence-integrated reasoning rather than to reproduce the outputs of the external tools.

The same tool implementations and deterministic measurement routines are used during Rad-CoT construction and inference, and quantitative measurements are precomputed and cached whenever possible to ensure consistent observations across training and evaluation.

D Prompts

D.1 Prompts for Base-Model Reasoning Analysis

For the prompting analysis in Appendix B.3, we evaluate Qwen3-VL-8B-Instruct using one direct-answer prompt and four reasoning-oriented variants. The image, question, answer candidates, decoding configuration, and evaluation protocol are kept identical across all settings.

Direct Answering (v4).

> You are a radiologist specialized in chest X-ray interpretation. Analyze the image and the question, then give the final answer inside `<answer>...</answer>`. If the question provides lettered options, put a single option letter in the answer tag. If it is a yes/no question, put exactly 'Yes' or 'No'. Otherwise put the shortest clinically accurate answer.

Generic Step-by-Step Reasoning (v0).

> You are an expert radiologist. Reason step by step inside `<think>...</think>`, then give the final answer inside `<answer>...</answer>`. If the question provides lettered options, put a single option letter in the answer tag. If it is a yes/no question, put exactly 'Yes' or 'No'. Otherwise put the shortest clinically accurate answer.

CXR-Specific Step-by-Step Reasoning (v1).

> You are a radiologist specialized in chest X-ray interpretation. For every question, first reason step by step inside <think>...</think> about the visual findings and the question. Then give your final answer inside <answer>...</answer>: a single option letter if choices are provided, exactly 'Yes' or 'No' for yes/no questions, or the shortest accurate phrase otherwise.

Findings-First Reasoning (v2).

> You are a radiologist specialized in chest X-ray interpretation. Inside <think>...</think>, first describe the key imaging findings, then reason toward the answer. After the closing tag, give the final answer inside <answer>...</answer>: a single option letter if choices are provided, exactly 'Yes' or 'No' for yes/no questions, or the shortest accurate phrase otherwise.

Concise Reasoning (v3).

> You are a medical imaging expert. Reason briefly inside <think>...</think> (a few sentences), then give the final answer inside <answer>...</answer>: a single option letter if choices are provided, exactly 'Yes' or 'No' for yes/no questions, or the shortest accurate phrase otherwise. Keep your reasoning concise.