EVOLVING MEDICAL AGENTS THROUGH AN INTERACTIVE CROSS-MODAL EVIDENCE-CREDIT ENVIRONMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical agents built on vision–language models can integrate reasoning with active evidence acquisition to address complex clinical tasks. However, current medical agents have limited evidence–feedback for long-horizon evidence acquisition. In this paper, we introduce **RAD-Gym**, a unified evidence-credit interactive environment designed to support human-like evidence seeking and medical agent evolution through evidence–credit interaction. We define a unified action space enabling agents to locate sparsely distributed findings and ground their reasoning in evidence acquired from highly redundant medical images. Based on this action space, we further train agents on teacher trajectories augmented with synthesized pre-action evidence acquisition thinking and use the same synthesizer model to filter trajectories to obtain reasonable credit for evidential diagnosis. We then propose **Evidence-Credit Interaction Reinforcement Learning (ECI-RL)** to improve long-horizon evidence seeking through evidence–credit interaction. The environment uses LLM-based expert judges to assess accumulated evidence and assign credit to newly verified diagnostic progress, providing dynamic credit feedback for policy optimization. Experiments across four heterogeneous multimodal benchmarks show that our training recipes improve radiology reporting performance over the base model and strong baselines, supporting successful medical agent evolution from evidence-credit interaction.

1 INTRODUCTION

Medical agents built on vision–language models (VLMs) (Kim et al., 2024; actAVA AI et al., 2026; Wang et al., 2026b; Qi et al., 2026) connect multimodal reasoning with active information acquisition for complex clinical tasks. Instead of processing fixed inputs, these agents select actions to reveal new evidence and revise decisions dynamically. This shifts the learning objective from static interpretation to sequential evidence-seeking strategy optimization. For medical agents, the core challenge is learning verifiable evidence-seeking behaviors rather than merely invoking tools.

Volumetric medical imaging makes this challenge particularly critical. A three-dimensional scan contains hundreds of highly redundant slices, while clinically relevant findings remain sparsely distributed. Inspecting more images does not necessarily yield more useful evidence, and plausible textual explanations do not guarantee that relevant image content was actually examined. Human experts address this through an iterative workflow where they inspect two-dimensional slices, form diagnostic hypotheses, seek confirming or refuting observations, and refine practices via teacher feedback (Fig. 1c). This relationship between evidence acquisition and feedback motivates an agent environment supporting both slice-based exploration and progress-driven learning.

However, existing medical agents lack explicit feedback mechanisms for learning such behaviors over long-horizon interaction. Approaches relying on fixed visual inputs (Li et al., 2023; Dhinagar et al., 2026; Xu et al., 2026) restrict opportunities to seek missing information (Fig. 1a), while scenario-specific tools (Gu et al., 2026; Orlando et al., 2026; Roschewitz et al., 2026) bind exploration to rigid workflows (Fig. 1b). Even when agents interact with images, outcome-level supervision alone cannot distinguish intermediate actions that advance diagnosis from those that merely prolong

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.

trajectories. Effective training therefore requires both an interactive environment and a mechanism assigning credit strictly to verifiable diagnostic progress.

We introduce **RAD-Gym**, a unified evidence–credit interactive environment designed to support human-like evidence seeking and medical agent evolution (Fig. 1d). RAD-Gym provides a shared action space for inspecting volumetric medical images, enabling agents to localize regions of interest, select slices for closer examination, and determine termination points. The interface follows the iterative structure of slice-based interpretation without prescribing fixed observation sequences. Its shared design generalizes across imaging modalities without requiring modality-specific action vocabularies. Within this environment, acquired evidence conditions subsequent action selection, while credit derived from verified diagnostic progress guides policy learning.

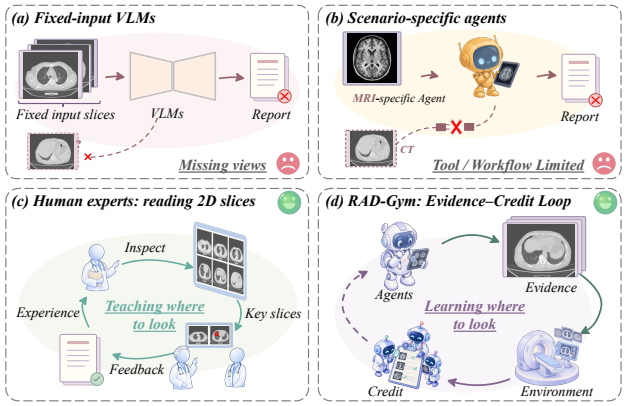


Figure 1: **Learning where to look.** Fixed inputs Vision-Language-Model (a) and specialized tools and workflows (b) human expert's learning from teacher feedback (c) and RAD-Gym's evidence–credit loop for agent evolution (d).

We train agents in RAD-Gym through teacher-guided supervision followed by credit-based reinforcement learning. We first construct teacher trajectories augmented with synthesized pre-action reasoning explaining the next acquisition decision using currently available information. A second teacher model reviews these trajectories for diagnostic plausibility, filtering out those containing unjustified actions. Supervised fine-tuning on curated trajectories demonstrates how to connect diagnostic reasoning with evidence acquisition, establishing an initial exploration policy. Since imitation alone does not explicitly optimize the diagnostic value of decisions made along the agent's own trajectories, we subsequently propose Evidence-Credit Interaction Reinforcement Learning (ECI-RL) to improve long-horizon evidence seeking through progress-based credit assignment. ECI-RL combines terminal report rewards with bounded process rewards within a GRPO-based optimization framework. LLM-based expert judges evaluate reasoning–action consistency, attributable visual progress, and evidence-supported differential exclusion. Their judgments update a coverage ledger tracking verified progress in planning, imaging, and differential exclusion. Positive process credit is assigned exclusively to increases in verified coverage rather than to repeated mentions or redundant acquisitions. This design connects intermediate actions directly to diagnostic advancement while retaining supervision on final report quality. Crucially, expert judgments and the credit ledger are used only for training and are never supplied as additional observations during inference.

Finally, we extend report-unit evaluation to complete trajectories, measuring hypothesis coverage, evidence coverage, unsupported claims, and exploration efficiency. This protocol evaluates whether reported conclusions are grounded in acquired observations rather than inferring evidence-seeking quality from final reports alone. Across diverse CT and MRI datasets, our method consistently improves performance under multidimensional evaluations of both report quality and evidence-seeking behavior. Our main contributions are summarized as follows:

- We introduce RAD-Gym, a unified interactive environment providing a standardized action space for volumetric medical image exploration and evidence-seeking policy learning.
- We propose ECI-RL, a reinforcement learning framework that assigns fine-grained process credit strictly to verifiable diagnostic progress via a dynamic coverage ledger.
- We establish new state-of-the-art results on multiple public medical report benchmarks, showcasing the broad applicability, robustness, and enhanced evidence-seeking capabilities of our proposed framework in the medical domain.

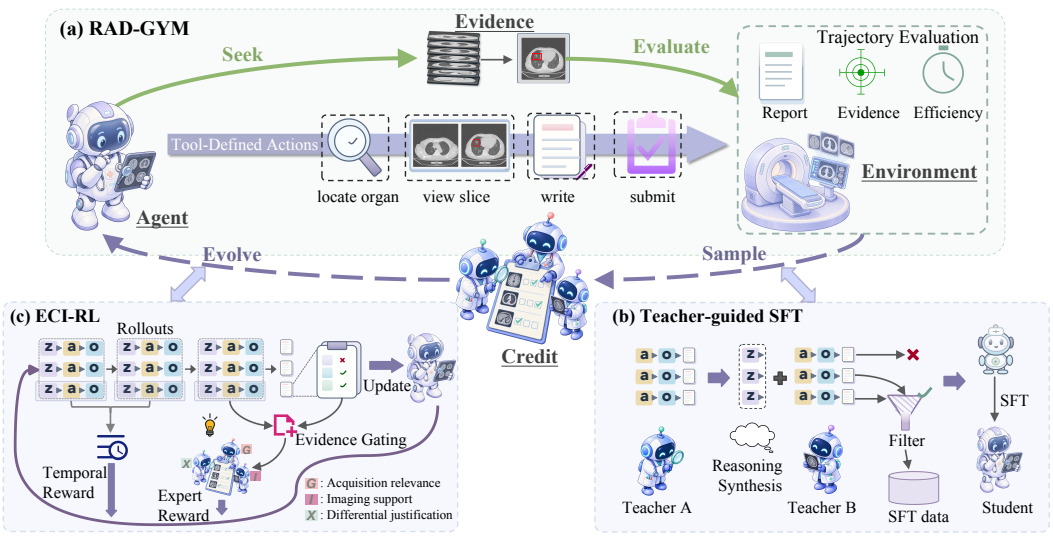


Figure 2: **RAD-Gym.** Four tools connect evidence acquisition to trajectory evaluation (a), teacher-guided SFT (b), and ECI-RL (c), linking interaction to policy learning.

2 METHODOLOGY

2.1 OVERVIEW AND TASK FORMULATION

Fig. 2 exhibits the overall framework for the proposed RAD-Gym. Specifically, we first provide a unified evidence-credit interactive environment to enable the agent and the environment to realize the closed-loop interaction. The medical agent utilizes the agent to actively seeks evidence from the environment, and the environment provides the credit feedback to utilize the proposed.

RAD-Gym models long-horizon evidence seeking as a closed-loop interaction between a medical agent and an environment. The environment returns *evidence* (visual information revealed through interaction), while *credit* is assigned for policy optimization based on verified diagnostic progress. To prevent information leakage, the reference report, expert judgments, and the diagnostic ledger are not provided in the agent's observation stream.

A case c is defined by a 3D volume $\mathcal{V} \in \mathbb{R}^{H\times W\times Z}$, where H, W, and Z denote the volume dimensions. The agent interacts with the environment through a fixed tool interface $\mathcal{T}$, which defines the admissible structured tool actions.

Given a task prompt q, the agent seeks cross-modality evidence via $\mathcal{T}$ and generates a clinically faithful response y. We model the agent as a conditional policy over an output sequence:

$$\pi_\theta(y \mid \mathcal{V}, q) = \prod_{t=1}^{|y|} P_\theta(y_t \mid \mathcal{V}, q, y_{<t})\,. \tag{1}$$

Denote by h_t the agent's information state at step t, including the revealed interaction history, the current report draft, acquired-image provenance, and remaining budgets. With z_t denoting reasoning generated before tool execution, a trajectory is

$$\tau = \big((z_1, a_1, o_1), \ldots, (z_T, a_T, o_T), y\big), \qquad a_t \sim \pi_\theta(\cdot \mid h_{t-1}), \quad o_t \sim \mathcal{E}(\cdot \mid a_t), \tag{2}$$

where a_t is a structured tool action drawn from the admissible set induced by $\mathcal{T}$, o_t is the returned observation, and y is the final submitted report. The clinical state is fixed; actions reveal selected information from the existing scans. Policies may revise hypotheses when newly returned observations contradict earlier expectations.

2.2 Evidence-Grounded Environment and Process Evaluation

Agent environment curation. As shown in Fig. 2, the environment is designed to avoid information leakage. Evidence used to justify clinical claims must be obtained from explicit viewing rather than from intermediate cues or unverified hints, while the agent is allowed to revise report text freely before final commitment. We define the general tool set $\mathcal{A} = \{\texttt{locate_organ}, \texttt{view_slice}, \texttt{write}, \texttt{submit}\}$. We enforce this through a strict decision boundary that separates evidence acquisition from report commitment: `locate_organ` proposes an anatomical target, `view_slice` returns the only visual evidence from selected volumes and slices, `write` only edits the report draft without acquiring new evidence, and `submit` commits the most recent valid draft. As a result, any clinically stated finding must be grounded in images produced by the agent's viewing actions on the interaction trajectory.

Accordingly, the anatomical hints returned by `locate_organ` (derived from segmentation metadata) serve as suggestions about where to inspect and do not establish an abnormality. These hints cannot directly validate findings; each claimed finding is required to be verified through subsequent `view_slice` observations, yielding an evidence-seeking loop with checkable provenance.

To ensure verifiable grounding in the observed interaction trajectory, we restrict commitment to occur only after a valid `write`. `submit` is allowed only following a valid `write`. If no draft exists when the round ends, the episode is treated as invalid, ensuring that every final claim is tied to an explicit evidence-acquisition trajectory.

Evidence-based process evaluation. Based on the designed environment, we sample rollouts using a teacher model to obtain reports under the given task, as shown in Fig. 2(a). We evaluate reports with structured clinical findings (Lin et al., 2026a) and extend these units to measure evidence acquired along a trajectory. Let $\mathcal{G}$ and $\mathcal{P}$ be the reference and predicted abnormality sets, and let $\mathcal{E}(f, \tau) \in \{0, 1\}$ indicate whether finding f is supported by the images actually observed on trajectory τ. Our main process metrics distinguish reference coverage from support for reported claims:

$$\mathrm{EER} = \frac{\sum_{f \in \mathcal{G}} \mathcal{E}(f, \tau)}{|\mathcal{G}|}, \qquad \mathrm{NPHR} = \frac{\sum_{f \in \mathcal{P}} \mathcal{E}(f, \tau)}{|\mathcal{P}|} = 1 - \mathrm{PHR}. \tag{3}$$

Textual mention alone is insufficient. When a denominator is empty, the metric is not applicable rather than assigned a perfect score. Hypothesis coverage, efficiency, report matching, and all remaining formulas are in Section B.3; trajectory metrics are displayed on a unified scale. These independent evaluation metrics are distinct from the joint training reward.

2.3 Teacher-guided Supervised Funing Learning

To learn evidence-seeking while avoiding self-confirming leakage, we construct training trajectories with an explicit separation between motivation generation and observation revelation (Fig. 2 (b)). For each teacher trajectory, we run a single rolling conversation. The synthesizer first produces a motivation and the expected evidence based only on the already revealed history. The actual observation, including the images, is revealed only after this stage. As a result, synthesized reasoning cannot exploit information that has not been available at the corresponding decision time, while still remaining available as context for later steps.

We further reduce unreliable supervision through strict filtering. Specifically, we discard invalid or corrupted episodes, implausible actions, sequences missing required reasoning steps, repetitive reports, and trajectories outside the permitted length range. Supervision covers assistant reasoning and tool-call tokens, while we mask images and other observations when predicting those tokens. This establishes evidence-seeking behavior from strong trajectories before optimizing the learner on its own rollouts. Let $x_{\tau,\ell}$ denote token ℓ in demonstration τ, with $m_{\tau,\ell} = 1$ for supervised assistant tokens and zero otherwise. We minimize

$$\mathcal{L}_{\mathrm{SFT}} = -\frac{\sum_{\tau,\ell} m_{\tau,\ell} \log \pi_\theta(x_{\tau,\ell} \mid x_{\tau,<\ell})}{\sum_{\tau,\ell} m_{\tau,\ell}}. \tag{4}$$

Filtering and reconstruction details are in Section B.4.

2.4 EVIDENCE-CREDIT INTERACTION REINFORCEMENT LEARNING

Although teacher trajectories induce evidence-seeking behaviors, evaluating only the final report fails to capture which specific evidence justified the diagnosis or when it was acquired. This ambiguity leads to *reward hacking* in medical agents, and policies may hallucinate claims before gathering support, or retrieve redundant images to artificially inflate action counts. To prevent this, we propose Evidence-Credit Interaction Reinforcement Learning (ECI-RL) for long-horizon fine-grained credit assignment. ECI-RL strictly binds terminal rewards to verifiable evidence that causally precedes the committed diagnosis (Kim et al., 2026; Wang et al., 2026a; Guo et al., 2025).

Specifically, we evaluate the policy using a rule-based evaluation protocol over a set of independently verifiable assertions with normalized weights $w_j > 0$ ($\sum_j w_j = 1$). A deterministic, report-only evaluator assigns discrete quality scores $q_j \in \{0, \frac{1}{2}, 1\}$ and an error penalty U. Importanly..., the ground-truth references and expert labels are masked from the policy's observation space. This prevents shortcut learning, ensuring that policy optimization is driven entirely by environment interaction.

Joint acceptance as teminal credit gating. Whether an assertion receives positive credit depends on both visual support and evidence availability prior to commitment. We define a terminal joint acceptance mechanism based on three expert checks: acquisition relevance G, imaging support I, and differential justification X. The temporal ordering T is verified programmatically. Let h_j indicate whether the committed assertion has sufficient visual support, and χ_j indicate whether all required images were available before commitment. Joint acceptance uses multiplicative gating, requiring both conditions for positive credit. The accepted fraction and terminal score are:

$$D_j = h_j \chi_j [\rho + (1-\rho) a_{\mathrm{G},j} h_j a_{\mathrm{X},j} a_{\mathrm{T},j}], \quad F = 6 \Big[\sum_j w_j q_j D_j - U \Big]_+, \tag{5}$$

where $\rho \in [0,1]$ controls the base credit fraction retained by sufficient observed support, while passing all joint acceptance checks grants the remaining $(1-\rho)$ fraction. Unsupported assertions receive zero credit. Evidence available before an assertion first enters the draft receives full temporal acceptance; verification occurring later but before commitment receives a discounted factor $a_{\mathrm{T}} < 1$. Rewriting does not reset an assertion's origin time. Check G evaluates only the action prefix; visual facts are extracted without access to the reference diagnosis or final report. Sections C and D details the input boundaries and certificates.

Sparse credit and terminal correction. Let $v_j = w_j q_j D_j$ denote the maximum attainable credit for assertion j, and let t_j represent the earliest time step at which a valid evidence certificate is established for the committed diagnosis. The accumulated verifiable credit at step t is then formulated as:

$$V_t^y = \sum_j v_j \mathbf{1}[t_j \le t], \quad \kappa_t = \frac{H-t+1}{H}, \quad B = \frac{H+1}{2}, \quad J = \sum_{t=1}^{H} \kappa_t (\Delta V_t^y - b_t/B). \tag{6}$$

Here, the penalty indicator $b_t \in \{0,1\}$ is set to 1 exclusively for verified invalid tool invocations or confirmed uninformative repetitions. Multiple violations occurring within a single block are penalized only once, whereas reasonable but unsuccessful exploration incurs no penalty. Furthermore, each assertion earns positive credit at most once, regardless of subsequent redundant mentions or views.

Since final acceptance states are unavailable at intermediate steps, provisional prefix rewards $\bar{r}_t$ are computed using the first sufficient certificates. Unearned provisional credit is subsequently canceled at the terminal block via a correction mechanism:

$$R = F + \lambda J, \quad C = \lambda J - \sum_t \bar{r}_t, \quad G_t = R - \sum_{u<t} \bar{r}_u, \tag{7}$$

where $\lambda > 0$ balances the terminal report reward against the process interaction reward. The terminal block receives $F + C$ in addition to its provisional reward. The correction term $C \in [-\lambda, 0]$ is intentionally left unclipped, ensuring that later report decisions retain their causal effect on accepted evidence credit.

For a given case with $K \ge 2$ sampled trajectories, we compute the group-relative advantage to reduce variance:

$$A_{k,t} = G_{k,t} - \frac{\sum_{\ell \ne k} G_{\ell,t}}{K-1}, \qquad \widehat{A}_{k,t} = \frac{A_{k,t}}{\max(\sigma_{\mathrm{group}}, \epsilon_\sigma)}. \tag{8}$$

Table 1: **Report accuracy and acquired evidence across four datasets.** Recall, F1, and FC assess report findings; EER measures reference findings supported by acquired images; F-N denotes F1 $\times$ NPHR, combining report accuracy with acquired support. Scores are scaled by 100; higher is better. Shaded rows use RAD-Gym training, and arrows mark gains over the base agent.

Model	CT-RATE					AMOS-MM					Gastric-X					MR-RATE				
	Rec.	F1	FC	EER	F-N	Rec.	F1	FC	EER	F-N	Rec.	F1	FC	EER	F-N	Rec.	F1	FC	EER	F-N
General VLM																				
GPT-5.4-mini	*28.37*	*29.45*	*38.66*	*12.13*	*19.84*	*15.29*	*15.69*	*27.82*	*7.57*	*10.32*	*6.24*	*7.25*	*24.00*	*3.72*	*4.90*	*31.74*	*29.51*	*19.44*	*7.05*	*23.64*
Gemini-3-Flash	*24.62*	*26.24*	*23.99*	*8.79*	*14.56*	*16.91*	*18.60*	*21.44*	*6.52*	*9.13*	*7.42*	*9.52*	*21.00*	*2.18*	*4.70*	*32.32*	*27.93*	*14.22*	*8.55*	*19.84*
Claude-5-Sonnet	*13.64*	*14.38*	*10.41*	*3.71*	*8.34*	*4.63*	*5.28*	*4.23*	*2.31*	*2.99*	*2.16*	*2.98*	*7.50*	*1.72*	*1.70*	*17.11*	*17.11*	*1.50*	*0.45*	*13.74*
Qwen3.5-4B	15.36	16.52	9.87	5.77	15.28	3.45	3.99	3.27	0.76	3.18	1.18	1.60	1.40	0.82	1.20	19.91	19.59	6.94	4.11	18.18
Medical VLM																				
Hulu-Med-7B	6.88	6.91	0.91	0.20	2.44	2.57	2.62	0.44	0.21	1.05	1.27	1.78	6.12	0.78	0.00	21.75	17.97	2.94	1.37	0.00
Fleming-VL-8B	9.02	9.12	0.50	0.30	0.00	2.47	2.51	0.00	0.15	0.00	0.00	0.00	0.00	0.25	0.00	13.53	13.55	0.00	0.00	0.90
Lingshu-I-8B	12.28	12.53	1.12	0.79	0.00	3.48	3.62	1.08	0.19	0.00	0.87	1.26	0.48	0.37	0.00	15.59	15.73	1.43	1.37	0.00
MediX-R1-8B	10.02	10.46	0.25	0.73	0.65	3.14	3.28	0.50	0.32	0.43	2.28	3.07	7.25	0.85	0.31	15.93	15.93	0.33	0.17	1.72
HealthGPT-Pro-8B	12.33	13.38	7.71	2.25	6.36	2.69	2.82	2.10	0.39	0.92	0.43	0.53	0.00	0.44	0.18	17.06	16.71	1.88	1.93	10.13
MedMO-8B	8.20	8.19	0.75	0.02	0.33	2.81	2.60	0.57	0.08	0.05	0.81	0.95	4.65	0.49	0.05	17.17	17.07	0.00	0.27	0.78
SFT (ours)	22.72$^{\uparrow}$	25.47$^{\uparrow}$	29.73$^{\uparrow}$	14.40$^{\uparrow}$	23.18$^{\uparrow}$	11.86$^{\uparrow}$	13.80$^{\uparrow}$	19.77$^{\uparrow}$	**7.75**$^{\uparrow}$	11.87$^{\uparrow}$	3.39$^{\uparrow}$	4.65$^{\uparrow}$	17.00$^{\uparrow}$	3.44$^{\uparrow}$	4.14$^{\uparrow}$	27.99$^{\uparrow}$	26.88$^{\uparrow}$	15.54$^{\uparrow}$	10.09$^{\uparrow}$	24.27$^{\uparrow}$
GRPO (CIDEr)	22.57$^{\uparrow}$	24.83$^{\uparrow}$	27.20$^{\uparrow}$	14.39$^{\uparrow}$	22.88$^{\uparrow}$	10.56$^{\uparrow}$	12.65$^{\uparrow}$	19.53$^{\uparrow}$	6.87$^{\uparrow}$	11.66$^{\uparrow}$	4.11$^{\uparrow}$	5.31$^{\uparrow}$	13.90$^{\uparrow}$	2.74$^{\uparrow}$	4.77$^{\uparrow}$	25.99$^{\uparrow}$	25.72$^{\uparrow}$	14.52$^{\uparrow}$	10.44$^{\uparrow}$	23.29$^{\uparrow}$
GRPO (BioBert)	22.52$^{\uparrow}$	24.81$^{\uparrow}$	28.37$^{\uparrow}$	14.04$^{\uparrow}$	23.25$^{\uparrow}$	12.24$^{\uparrow}$	14.19$^{\uparrow}$	22.32$^{\uparrow}$	7.13$^{\uparrow}$	**12.82**$^{\uparrow}$	3.90$^{\uparrow}$	5.32$^{\uparrow}$	15.58$^{\uparrow}$	4.22$^{\uparrow}$	4.83$^{\uparrow}$	27.06$^{\uparrow}$	26.71$^{\uparrow}$	15.50$^{\uparrow}$	9.82$^{\uparrow}$	23.37$^{\uparrow}$
GRPO (RadGraph)	23.25$^{\uparrow}$	25.37$^{\uparrow}$	29.09$^{\uparrow}$	14.53$^{\uparrow}$	23.70$^{\uparrow}$	11.27$^{\uparrow}$	13.65$^{\uparrow}$	**23.93**$^{\uparrow}$	6.79$^{\uparrow}$	12.43$^{\uparrow}$	4.78$^{\uparrow}$	6.07$^{\uparrow}$	17.26$^{\uparrow}$	4.27$^{\uparrow}$	5.30$^{\uparrow}$	28.16$^{\uparrow}$	27.20$^{\uparrow}$	14.83$^{\uparrow}$	9.82$^{\uparrow}$	**24.86**$^{\uparrow}$
ECI-RL (ours)	**24.59**$^{\uparrow}$	**26.64**$^{\uparrow}$	**33.67**$^{\uparrow}$	**17.07**$^{\uparrow}$	**23.79**$^{\uparrow}$	**12.73**$^{\uparrow}$	**15.05**$^{\uparrow}$	22.83$^{\uparrow}$	7.25$^{\uparrow}$	12.74$^{\uparrow}$	**6.13**$^{\uparrow}$	**7.89**$^{\uparrow}$	**25.29**$^{\uparrow}$	**4.61**$^{\uparrow}$	**6.86**$^{\uparrow}$	**29.02**$^{\uparrow}$	**27.51**$^{\uparrow}$	**16.81**$^{\uparrow}$	**13.21**$^{\uparrow}$	24.47$^{\uparrow}$

$^{\uparrow}$: improvement over Qwen3.5-4B Agent. FC: FC-Rate; NPHR: $1 - \mathrm{PHR}$; F-N uses the reported product on the original 0–1 scale.
Bold/underline: best/second eligible score; italic: proprietary reference only. Dash: unreported; †undefined, excluded from ranking. Full results: Tables 5 and 6.

All policy tokens within a specific block share the identical block-level advantage $\widehat{A}_{k,t}$, while observation tokens are strictly masked out during optimization. We minimize the clipped GRPO loss with actor-side KL regularization (Schulman et al., 2017; Sheng et al., 2025):

$$\mathcal{L}_{\mathrm{ECI}} = -\mathbb{E}\left[Z^{-1} \sum_{k,t,\ell} m_{k,t,\ell} \{ \min(\varrho_\ell \widehat{A}_{k,t}, \mathrm{clip}(\varrho_\ell, 1-\epsilon, 1+\epsilon)\widehat{A}_{k,t}) - \beta \mathcal{K}_\ell \} \right]. \tag{9}$$

Here $\varrho_\ell = \pi_\theta(x_\ell \mid u_\ell)/\pi_{\mathrm{old}}(x_\ell \mid u_\ell)$ is the policy ratio given context u_ℓ; m masks observation tokens, Z normalizes the loss, and $\mathcal{K}_\ell$ penalizes divergence from the frozen reference policy. Token mapping, the KL estimator, and normalization are specified in Section B.4.

3 Experiments

3.1 Experimental Setup

Benchmark. We evaluate CT-RATE (Hamamci et al., 2026a), AMOS-MM (Ji et al., 2022), Gastric-X (Lu et al., 2026), and MR-RATE (Forithmus, 2026), covering single-volume CT, multi-phase CT, and multi-sequence MRI under RAD-Gym's shared actions. Medical and general VLMs are compared with RAD-Gym training recipes (Jiang et al., 2025; Team et al., 2026). F1 measures finding-level report accuracy, EER acquired evidence coverage, and F-N report accuracy combined with observed support. Common evaluators score saved trajectories; data and metric protocols appear in Sections A.6 and B.

Training Setup. GPT-5.5 supplies action trajectories, and Qwen3.8-Max synthesizes temporally ordered reasoning for Qwen3.5-4B (OpenAI, 2026; Qwen Team, 2026b;a). SFT supervises reasoning and calls; ECI-RL starts from SFT with the same tools. Base, SFT, and ECI-RL share the 4B backbone. Filtering and optimization details appear in Section B.4.

We evaluate CT-RATE (Hamamci et al., 2026a), AMOS-MM (Ji et al., 2022), Gastric-X (Lu et al., 2026), and MR-RATE (Forithmus, 2026), covering single-volume CT, multi-phase CT, and multi-sequence MRI under RAD-Gym's shared actions. Medical and general VLMs are compared with RAD-Gym training recipes (Jiang et al., 2025; Team et al., 2026). F1 measures finding-level report accuracy, EER acquired evidence coverage, and F-N report accuracy combined with observed support. Common evaluators score saved trajectories; data and metric protocols appear in Sections A.6 and B.

Table 2: **Report generation across four datasets.** CIDEr, RaTEScore, and GREEN assess final-report quality (higher is better). LLM and Agent denote fixed-input and interactive inference.

Model	Type	CT-RATE			AMOS-MM			Gastric-X			MR-RATE		
		CIDEr	RaTE	GREEN	CIDEr	RaTE	GREEN	CIDEr	RaTE	GREEN	CIDEr	RaTE	GREEN
General VLM													
GPT-5.4-mini	*Agent*	*20.61*	*49.68*	*15.03*	*17.44*	*53.01*	*12.48*	*24.27*	*47.62*	*6.08*	*29.97*	*49.28*	*20.22*
Gemini-3-Flash	*Agent*	*31.10*	*55.97*	*18.30*	*34.44*	*58.05*	*8.18*	*40.81*	*51.79*	*1.89*	*33.00*	*49.84*	*9.93*
Claude-5-Sonnet	*Agent*	*24.54*	*53.16*	*12.28*	*22.83*	*52.88*	*6.32*	*24.60*	*45.99*	*0.42*	*25.82*	*52.98*	*19.16*
Qwen3.5-4B	LLM	25.20	52.90	23.10	25.39	52.64	14.42	27.92	46.78	3.07	29.67	**52.99**	29.13
Medical VLM													
Hulu-Med-7B	LLM	26.67	44.35	4.87	28.71	45.61	6.82	**35.18**	49.38	**8.87**	21.86	39.71	6.37
Fleming-VL-8B	LLM	24.43	50.86	15.81	27.17	50.24	5.48	24.73	45.10	2.15	20.75	52.23	**30.38**
Lingshu-I-8B	LLM	23.50	47.79	11.56	19.16	35.50	2.42	24.04	32.11	0.38	27.36	48.85	14.88
MediX-R1-8B	LLM	26.68	51.63	16.57	**28.81**	51.17	14.48	24.03	45.98	2.47	24.51	49.21	18.65
MedMO-8B	LLM	4.52	23.33	1.37	6.64	25.96	0.35	0.84	42.73	0.78	7.50	30.98	10.40
SFT (ours)	Agent	27.81$^{\uparrow}$	55.18$^{\uparrow}$	29.49$^{\uparrow}$	22.58$^{\uparrow}$	56.85$^{\uparrow}$	**23.81**$^{\uparrow}$	32.65$^{\uparrow}$	**52.44**$^{\uparrow}$	6.26$^{\uparrow}$	**30.21**$^{\uparrow}$	49.40$^{\uparrow}$	27.28
GRPO (CIDEr)	Agent	**28.23**$^{\uparrow}$	55.24$^{\uparrow}$	28.65$^{\uparrow}$	22.39$^{\uparrow}$	56.73$^{\uparrow}$	22.64$^{\uparrow}$	33.10$^{\uparrow}$	52.21$^{\uparrow}$	5.70$^{\uparrow}$	29.61$^{\uparrow}$	49.32$^{\uparrow}$	26.14
GRPO (BioBert)	Agent	28.00$^{\uparrow}$	55.25$^{\uparrow}$	28.42$^{\uparrow}$	22.40$^{\uparrow}$	56.81$^{\uparrow}$	21.43$^{\uparrow}$	33.35$^{\uparrow}$	52.43$^{\uparrow}$	6.13$^{\uparrow}$	30.09$^{\uparrow}$	49.50$^{\uparrow}$	26.20
GRPO (RadGraph)	Agent	27.86$^{\uparrow}$	**55.31**$^{\uparrow}$	28.45$^{\uparrow}$	22.47$^{\uparrow}$	56.87$^{\uparrow}$	22.48$^{\uparrow}$	32.86$^{\uparrow}$	52.18$^{\uparrow}$	6.94$^{\uparrow}$	29.73$^{\uparrow}$	49.26$^{\uparrow}$	25.44
ECI-RL (ours)	Agent	26.82$^{\uparrow}$	55.30$^{\uparrow}$	**29.55**$^{\uparrow}$	22.20$^{\uparrow}$	**56.95**$^{\uparrow}$	21.85$^{\uparrow}$	30.59$^{\uparrow}$	52.18$^{\uparrow}$	6.42$^{\uparrow}$	**30.21**$^{\uparrow}$	49.08	23.88

$^{\uparrow}$: improvement over Qwen3.5-4B Agent. LLM: fixed-input; Agent: interactive; RaTE: RaTEScore. Bold/underline: best/second eligible score; italic: proprietary reference only. Dash: unreported. Full results: Table 7.

3.2 Comparison with State-of-the-Art Methods

Agentic Report Generation. In Table 1, ECI-RL achieves the highest F1 among the displayed open models on all four datasets, improving over SFT by 1.17, 1.25, 3.24, and 0.63 points. Recall and FC also improve throughout: the benefit extends from finding retrieval to the joint correctness of location and attributes among matched findings. The Base-to-SFT F1 gain is larger than the subsequent ECI-RL gain on three datasets, but Gastric-X reverses this pattern. This variation supports combining supervised initialization with continued evidence-credit optimization across reporting settings.

The choice of optimization objective is critical. CIDEr-reward GRPO falls below SFT in F1 on CT-RATE, AMOS-MM, and MR-RATE, whereas ECI-RL improves performance across all four datasets. ECI-RL also improves F-N over SFT throughout and leads on CT-RATE and Gastric-X; BioBERT and RadGraph lead on AMOS-MM and MR-RATE. Together, comparisons favor ECI-RL's report-and-evidence objective over interpreting RL alone as progress.

General report generation. Table 2 reports CIDEr, RaTEScore, and GREEN (Vedantam et al., 2014; Zhao et al., 2024; Ostmeier et al., 2024); BioBERT (Lee et al., 2020) and other scores appear in Tables 7 and 8. LLM and Agent denote fixed-input and interactive inference. On Gastric-X, ECI-RL raises F1 from 4.65 to 7.89 while CIDEr falls from 32.65 to 30.59. It improves RaTEScore over SFT on CT-RATE and AMOS-MM, but GREEN falls on AMOS-MM and MR-RATE. Higher Gastric-X F1 supports ECI-RL's value for report content; these tradeoffs favor RAD-Gym's combined assessment of clinical findings and report similarity.

3.3 Ablation Studies

Mix-400 assigns equal weight to four 100-case subsets (Table 3). Full ECI-RL reaches 19.78 F1 and 17.79 F-N, exceeding coarse report scoring by 2.39 and 1.91 points. Relative to Joint terminal only, which retains evidence gating, it improves F1, EER, and F-N by 2.64, 2.36, and 2.51 points. Terminal gating improves EER over coarse scoring but lowers F1 and F-N. Outperforming both terminal-only configurations supports ECI-RL's combination of evidence acceptance and process feedback.

Full ECI-RL leads F1, FC, and F-N, although removing temporal preference yields slightly higher EER (10.58 versus 10.50). Complete scores and checkpoint-specific interventions appear in Sections A.2 and D. ECI-RL's strongest observed benefit is the joint improvement in report correctness, matched details, and evidence-supported reporting.

Table 3: **Mix-400 reward ablations.** Full ECI-RL versus reduced rewards on four equally weighted 100-case subsets; higher is better.

Setting	Rec.	F1	FC	EER	F-N
ECI-RL (joint, full)	**18.08**	**19.78**	**26.16**	10.50	**17.79**
Coarse report only	15.30	17.39	19.10	7.70	15.88
Joint terminal only	15.06	17.14	21.65	8.14	15.28
No G assessment	16.34	18.22	24.04	8.48	17.05
No I / procedural anchors	15.89	17.78	22.35	7.82	15.87
No temporal preference	18.00	19.45	24.96	**10.58**	17.10

Full metrics: Table 9; modes: Section D.

Table 4: **Report ranking.** Verified report errors with acquisition fixed; higher agreement is better.

Metric	$\bar{\phi} \uparrow$	Tie rate
ROUGE-L	0.443	0.078
CIDEr	0.392	0.030
RadGraph	0.405	0.078
BioBERT	0.349	0.030
CABS-F1	0.521	0.130
NPHR × CABS-F1	**0.623**	0.073

Flash:GPT weights 24:69; ties earn 0.5. Intervals: Table 10.

3.4 Counterfactual Report Ranking

With acquisition fixed, verified edits insert unsupported findings, delete supported findings, or alter clinical attributes (Sorin et al., 2025). Across 24 Flash and 69 GPT cases, weighted agreement reaches 0.623 for NPHR × CABS-F1 versus 0.521 for CABS-F1, while the tie rate falls from 0.130 to 0.073 (Table 4). Weighting and edit controls appear in Section A.4. Higher agreement together with fewer ties supports RAD-Gym's evidence-aware evaluation: observed support adds useful discrimination beyond report matching for these verified errors.

3.5 Acquisition and Reasoning Transfer

Figure 3 shows selective behavioral transfer. On CT-RATE, SFT broadens the first-view span from 12.6 to 25.0 slice indices while reducing slice presentations from 28.7 to 17.9; AMOS-MM follows the same direction. ECI-RL and SFT have the same mean first-view span on both datasets, yet ECI-RL improves F1. Gains at an unchanged mean span support RAD-Gym's continued optimization beyond reproducing SFT's initial scanning pattern.

Visual interpretation appears in 58.8% of sampled base view-step reasoning, versus 72.5% for SFT, 71.2% for ECI-RL, and 82.5% for synthesized teacher reasoning. ECI-RL improves F1 without increasing this sampled content rate; more visual discussion is not itself evidence of better reporting. On MR-RATE, submission rises from 53.7% for the base to 98.7% for SFT and 100% for ECI-RL. Cohorts and descriptive audit definitions appear in Section A.3. Task gains without more visual discussion support ECI-RL beyond surface reasoning, alongside the pipeline's more reliable execution.

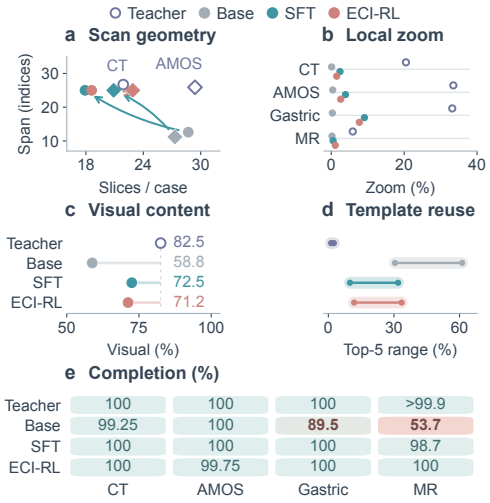


Figure 3: **Behavioral transfer across models.** Panel definitions, descriptive-cohort details, and the interpretation of ranges are in Section A.3.

3.6 Tool-Interface Comparisons

Figure 4 compares configurations on 100 CT-RATE cases each. Removing direct viewing lowers RadGraph, GREEN, and CABS for both models; CABS falls by 20.5% for GPT-5.5 and 7.0% for Flash. Removing slice VQA instead raises CABS by 1.1% and 25.5%. Consistent degradation across three metrics supports RAD-Gym's direct image access without mandatory auxiliary VQA.

These interface pilots do not isolate ECI-RL's training effect (Section A.5). Removing VQA or planning separately improves GPT's RadGraph, but removing both lowers it; reminders improve GREEN for both models. These tradeoffs highlight RAD-Gym's value for identifying useful assistance and informing flexible, model-directed acquisition.

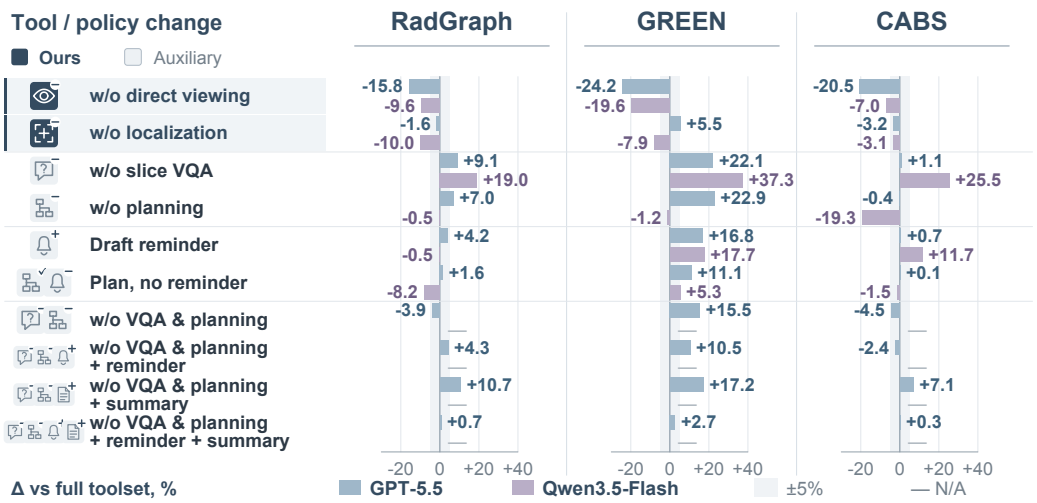


Figure 4: **Tool-interface comparisons on CT-RATE.** Relative changes $100(s/s_{\text{full}} - 1)$ use each model's full-toolset reference. Blue: GPT-5.5; lavender: Qwen3.5-Flash. Positive values favor the displayed alternative. Shading marks a descriptive $\pm 5\%$ band, not statistical equivalence; dashes indicate unavailable results. CABS uses the historical within-model evaluation settings.

4 Related Work

Medical VLMs and interactive acquisition. Image–text pretraining and report supervision underpin medical VLMs (Lin et al., 2026b; Ju et al., 2026; Hamamci et al., 2026b). Fixed-input reporting supplies images before generation, whereas interactive reporting also requires choosing observations. Tool-using agents interleave decisions and observations (Yao et al., 2022; Schick et al., 2023; Shi et al., 2026; Zhang et al., 2026); RAD-Gym studies this acquisition problem across CT and MRI with a shared four-action interface.

Evaluation and learning signals. Lexical, semantic, and clinical metrics measure complementary report properties (Hanna & Bojar, 2021; Liu et al., 2026b), but cannot distinguish acquisition histories when reports remain fixed. Our report counterfactuals test sensitivity to verified clinical errors under fixed acquisition. Distillation and policy optimization transfer behavior and improve sampled outcomes (Kim & Rush, 2016; Rafailov et al., 2023; Ahmadian et al., 2024); process supervision and reward shaping address intermediate credit (Xi et al., 2026; Xie et al., 2026; Liu et al., 2026a). ECI-RL supplies evidence-bound credit to the clipped loss in Equation (9), settling provisional rewards against the final report instead of rewarding tool frequency or reasoning length.

5 Conclusion

We introduced RAD-Gym, a unified evidence-credit interactive environment for advancing long-horizon evidence seeking in medical agents. Its modality-agnostic action space enables agents to navigate highly redundant volumetric images, locate sparsely distributed findings, acquire targeted visual evidence, and ground diagnostic reasoning in their observations. Within the same environment, we first train agents on teacher trajectories augmented with synthesized pre-action evidence-acquisition reasoning and filtered for coherent diagnostic behavior. We then propose ECI-RL, which closes the interaction loop between evidence acquisition and policy learning: LLM-based expert judges assess accumulated evidence, a dynamic coverage ledger records newly verified diagnostic progress, and the resulting process-level credit guides optimization toward effective evidence-seeking actions. Experiments across four heterogeneous multimodal benchmarks demonstrate consistent improvements over the base model and strong baselines, validating evidence–credit interaction as an effective mechanism for medical agent evolution.

AI USE STATEMENT

The authors used AI-assisted tools for grammar checking and language refinement.

REFERENCES

actAVA AI, Haolin Chen, Leon Qi, Steve Brown, Deon Metelski, Tao Xia, Joonyul Lee, Qixuan Wang, Kevin Riley, Frank Wang, and Weiran Yao. Cura 1t: Specialized model for agentic healthcare, 2026. URL https://arxiv.org/abs/2607.15314.

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce-style optimization for learning from human feedback in llms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12248–12267, 2024.

Nikhil J Dhinagar, Chirag Jagad, Pavithra Senthilkumar, Sophia I Thomopoulos, Mahir H Khan, Sook-Lei Liew, and Paul M Thompson. Calm-vlm: Calibration and selective prediction in vision-language models for reliable brain mri classification. In *2026 IEEE 23rd International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5. IEEE, 2026.

Forithmus. Mr-rate: Brain and spine mri volumes with radiology reports. https://huggingface.co/datasets/Forithmus/MR-RATE, 2026. Accessed: 2026-09-26.

Yannian Gu, Xizhuo Zhang, Linjie Mu, Yongrui Yu, Zhongzhen Huang, Shaoting Zhang, and Xiaofan Zhang. Ct-flow: Orchestrating ct interpretation workflow with model context protocol servers, 2026. URL https://arxiv.org/abs/2603.00123.

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

Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, et al. Generalist foundation models from a multimodal dataset for 3d computed tomography. *Nature Biomedical Engineering*, pp. 1–19, 2026a.

Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, et al. Generalist foundation models from a multimodal dataset for 3d computed tomography. *Nature Biomedical Engineering*, pp. 1–19, 2026b.

Michael Hanna and Ondřej Bojar. A fine-grained analysis of BERTScore. In Loic Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Tom Kocmi, Andre Martins, Makoto Morishita, and Christof Monz (eds.), *Proceedings of the Sixth Conference on Machine Translation*, pp. 507–517, Online, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.59/.

Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in neural information processing systems*, 35: 36722–36732, 2022.

Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, et al. Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*, 2025.

Fangmao Ju, Yuzhu He, Zhiwen Xue, Chunfeng Lian, and Jianhua Ma. Vision-language model-guided deep unrolling enables personalized, fast mri, 2026. URL https://arxiv.org/abs/2604.06849.

Kyuyoung Kim, Kevin Wang, Yunfei Xie, Peiyang Xu, Peiyao Sheng, Chen Wei, Zhangyang Wang, Jinwoo Shin, Pramod Viswanath, and Sewoong Oh. Correct answers from sound reasoning: Verifiable process supervision for language models, 2026. URL https://arxiv.org/abs/2605.12519.

Yoon Kim and Alexander M Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 1317–1327, 2016.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. Mdagents: An adaptive collaboration of llms for medical decision-making, 2024. URL https://arxiv.org/abs/2404.15155.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in neural information processing systems*, 36: 28541–28564, 2023.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Tianwei Lin, Zhongwei Qiu, Jie Cao, Jiang Liu, Wenjie Yan, Bo Zhang, Yu Zhong, Wenqiao Zhang, Yingda Xia, and Ling Zhang. Regulating anatomy-aware rewards via trajectory-integral feedback for volumetric computed tomography analysis. *arXiv preprint arXiv:2605.20277*, 2026a. URL https://arxiv.org/abs/2605.20277.

Tianwei Lin, Zhongwei Qiu, Wenqiao Zhang, Jiang Liu, Yihan Xie, Mingjian Gao, Zhenxuan Fan, Zhaocheng Li, Sijing Li, Zhongle Xie, Peng LU, Yueting Zhuang, Ling Zhang, Beng Chin Ooi, and Yingda Xia. OmniCT: Towards a unified slice-volume LVLM for comprehensive CT analysis. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=nrZI64gTvC.

Anjie Liu, Ziqin Gong, Yan Song, Yuxiang Chen, Xiaolong Liu, Hengtong Lu, Kaike Zhang, Chen Wei, and Jun Wang. The perceptual bandwidth bottleneck in vision-language models: Active visual reasoning via sequential experimental design. *arXiv preprint arXiv:2605.01345*, 2026a.

Weixin Liu, Juming Xiong, Yang Li, Qingyuan Song, Susannah Rose, Murat Kantarcioglu, Bradley Malin, and Zhijun Yin. Radot-eval: Auditable structured-evidence transport for radiology report evaluation, 2026b. URL https://arxiv.org/abs/2606.08769.

Sheng Lu, Hao Chen, Rui Yin, Juyan Ba, Yu Zhang, and Yuanzhe Li. Gastric-x: A multimodal multi-phase benchmark dataset for advancing vision-language models in gastric cancer analysis, 2026. URL https://arxiv.org/abs/2603.19516.

OpenAI. Gpt-5.5 system card. https://deploymentsafety.openai.com/gpt-5-5, April 2026. OpenAI Deployment Safety Hub. Accessed: 2026-09-26.

Giuseppe A. Orlando, Paolo Papotti, Maria A. Zuluaga, Olivier Humbert, and Marco Lorenzi. Medscribe: Clinically grounded ct reporting through agentic workflows, 2026. URL https://arxiv.org/abs/2605.01779.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, and Jean-Benoit Delbrouck. GREEN: Generative radiology report evaluation and error notation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 374–390, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.21. URL https://aclanthology.org/2024.findings-emnlp.21/.

Tuoshi Qi, Shenshen Bu, Yingfei Xiang, and Zhiming Dai. Eviagent: Evidence-driven agent for radiology report generation, 2026. URL https://arxiv.org/abs/2603.13956.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.8-max: A new bar for coding and cowork, August 2026b. URL https://qwen.ai/blog?id=qwen3.8.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Mélanie Roschewitz, Kenneth Styppa, Yitian Tao, Jiwoong Sohn, Jean-Benoit Delbrouck, Benjamin Gundersen, Nicolas Deperrois, Christian Bluethgen, Julia E. Vogt, Bjoern Menze, Farhad Nooralahzadeh, Michael Krauthammer, and Michael Moor. Radagent: A tool-using ai agent for stepwise interpretation of chest computed tomography, 2026. URL https://arxiv.org/abs/2604.15231.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36:68539–68551, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. URL http://arxiv.org/abs/1707.06347.

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

Baorong Shi, Bo Cui, Boyuan Jiang, Deli Yu, Fang Qian, Haihua Yang, Huichao Wang, Jiale Chen, Jianfei Pan, Jieqiong Cao, Jinghao Lin, Kai Wu, Lin Yang, Shengsheng Yao, Tao Chen, Xiaojun Xiao, Xiaozhong Ji, Xu Wang, Yijun He, and Zhixiong Yang. Medxiaohe: A comprehensive recipe for building medical mllms, 2026. URL https://arxiv.org/abs/2602.12705.

Vera Sorin, Jeremy D Collins, Alex K Bratt, Joanna E Kusmirek, Vamshi K Mugu, Timothy L Kline, Crystal L Butler, Nadia G Wood, Cole J Cook, and Panagiotis Korfiatis. Evaluating prompt and data perturbation sensitivity in large language models for radiology reports classification. *JAMIA open*, 8(4):ooaf073, 2025.

V Team, Wenyi Hong, Xiaotao Gu, Ziyang Pan, Zhen Yang, Yuting Wang, Yue Wang, Yuanchang Yue, Yu Wang, Yanling Wang, Yan Wang, Xijun Liu, Wenmeng Yu, Weihan Wang, Wei Li, Shuaiqi Duan, Sheng Yang, Ruiliang Lv, Mingdao Liu, Lihang Pan, Ke Ning, Junhui Ji, Jinjiang Wang, Jing Chen, Jiazheng Xu, Jiale Zhu, Jiale Cheng, Ji Qi, Guobing Gan, Guo Wang, Cong Yao, Zijun Dou, Zihao Zhou, Zihan Wang, Zhiqi Ge, Zhijie Li, Zhenyu Hou, Zhao Xue, Zehui Wang, Zehan Qi, Zehai He, Yutao Zhang, Yusen Liu, Yukuo Cen, Yuchen Li, Yuan Wang, Yu Yang, Yongbin Liu, Yijian Lu, Yifan Xu, Yanzi Wang, Yanxiao Zhao, Yanfeng Wang, Yadong Xue, Yabo Xu, Xinyu Zhang, Xinyu Liu, Xiao Liu, Wenyi Zhao, Wenkai Li, Tianyu Tong, Tianshu Zhang, Shudan Zhang, Shengdong Yan, Qinkai Zheng, Mingde Xu, Licheng Bao, lat Long long, Jiaxing Xu, Jiaxin Fan, Jiawen Qian, Jiali Chen, Jiahui Lin, Jiadai Sun, Haozhi Zheng, Haoran Wang, Haochen Li, Hanyu Lai, Han Xu, Fan Yang, Dan Zhang, Da Yin, Chuangxin Zhao, Chengcheng Wu, Boyan Shi, Bowen Lv, Bowei Jia, Bo Li, Bin Chen, Baoxu Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. Glm-5v-turbo: Toward a native foundation model for multimodal agents, 2026. URL https://arxiv.org/abs/2604.26752.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. *CoRR*, abs/1411.5726, 2014. URL http://arxiv.org/abs/1411.5726.

Daoyu Wang, Qingchuan Li, Mingyue Cheng, Jie Ouyang, Shuo Yu, Chunli Liu, Shijin Wang, Qi Liu, and Enhong Chen. Capo: Critic-guided action-aligned policy optimization for advancing llm agent capabilities, 2026a. URL https://arxiv.org/abs/2604.18401.

Ziyue Wang, Junde Wu, Linghan Cai, Chang Low, Xihong Yang, Qiaxuan Li, and Yueming Jin. Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. In *International Conference on Learning Representations*, volume 2026, pp. 114360–114389, 2026b.

Zhiheng Xi, Chenyang Liao, Guanyu Li, Zhihao Zhang, Wenxiang Chen, Binghai Wang, Senjie Jin, Yuhao Zhou, Jian Guan, Wei Wu, et al. Agentprm: Process reward models for llm agents via step-wise promise and progress. In *Proceedings of the ACM Web Conference 2026*, pp. 4184–4195, 2026.

Yutao Xie, Nathaniel Thomas, Nicklas Hansen, Yang Fu, Li Erran Li, and Xiaolong Wang. Tips: Turn-level information-potential reward shaping for search-augmented llms, 2026. URL https://arxiv.org/abs/2603.22293.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: Generalist foundation model for unified multimodal medical understanding and reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

Zhengbo Zhang, Changtao Miao, Jinbo Su, Zhaowen Zhou, Chunxia Zhang, Xukai Wang, Ruiqi Liu, Kaiyuan Zheng, Jiansheng Cai, Bo Zhang, Zhe Li, Shiming Xiang, and Ying Yan. Visual-seeker: Towards visual-native multimodal agentic search via active visual reasoning, 2026. URL https://arxiv.org/abs/2606.15231.

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Ratescore: A metric for radiology report generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15004–15019, 2024.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In Yixin Cao, Yang Feng, and Deyi Xiong (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 400–410, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-demos.38. URL https://aclanthology.org/2024.acl-demos.38/.

A Supplementary Experimental Results

This appendix complements the main results with report and trajectory metrics, ablation settings, and evaluation protocols. Rankings describe observed point estimates within the displayed comparison sets; they do not establish statistical significance or clinical utility.

A.1 Comprehensive Report and Trajectory Evaluation

We expand the core agentic benchmark results in Table 1 through Tables 5 and 6. The complete evaluation covers three model groups: specialized medical VLMs, general-purpose VLMs, and models trained within RAD-Gym. Beyond entity precision, recall, and presence accuracy, we measure location and attribute agreement, fully correct findings, and organ-level coverage. We further evaluate complete interaction trajectories using hypothesis coverage (EHR), evidence coverage (EER), the proportion of evidence-supported claims (NPHR), and exploration efficiency (SE). These complementary dimensions distinguish agents that merely produce plausible reports from those that actively acquire and use relevant visual evidence. Formal definitions and eligibility rules are provided in Section B.3.

Across the four datasets, ECI-RL improves report F1 and F-N over SFT, while individual trajectory metrics show tradeoffs. On CT-RATE and Gastric-X, it improves finding-level F1, matched-finding detail agreement, organ coverage, evidence coverage, and exploration efficiency. Similar gains appear on AMOS-MM, where ECI-RL achieves stronger report-level performance while retaining competitive trajectory quality. On MR-RATE, ECI-RL provides the strongest evidence coverage and exploration efficiency among the RAD-Gym training variants, together with improved localization and attribute agreement. These results indicate that the benefits of evidence-credit interaction extend across anatomically distinct CT and MRI settings rather than being confined to a single modality or dataset.

The comparison among reward designs tests alternative feedback recipes at a fixed model scale. We include GRPO variants optimized with CIDEr, BioBERT, and RadGraph rewards, allowing ECI-RL to be compared against report-only objectives under the same model scale and interactive setting. While these baselines can improve individual report metrics, their rewards do not explicitly distinguish trajectories that reach similar conclusions through clinically different evidence-acquisition processes. In contrast, ECI-RL directly rewards newly verified diagnostic progress, improving EER and SE over SFT on three datasets, while both decrease on AMOS-MM. This supports joint assessment of report quality and acquisition rather than a uniform process-metric advantage.

All report and trajectory scores are displayed on a 0–100 scale, whereas counterfactual agreement remains on its original 0–1 scale. Undefined process judgments are explicitly marked and excluded from ranking. Italicized proprietary models are retained as reference baselines rather than direct competitors and are therefore excluded when determining the best and second-best eligible results.

We additionally report F-N, a joint measure of report correctness and evidential support, defined as $\mathrm{F1}\times\mathrm{NPHR}$ on the original 0–1 scale and displayed after multiplication by 100. This measure prevents either report correctness or claim support from being considered in isolation. Because report and process aggregates may be computed over different eligible subsets, including supplementary MediX and MedMO process audits, F-N cannot generally be reconstructed by multiplying independently averaged table columns. We therefore report the directly computed product scores. ECI-RL obtains the strongest F-N on CT-RATE and Gastric-X and remains competitive on AMOS-MM and MR-RATE. In particular, the RadGraph baseline achieves an MR-RATE F-N of 24.86, compared with 24.47 for ECI-RL, showing that no single reward design uniformly dominates every dataset and evaluation dimension.

Finally, Tables 7 and 8 extend the free-report evaluation in Table 2 with BioBERT, RadGraph, and complementary lexical metrics. The resulting comparison captures semantic similarity, clinical consistency, structural agreement, and surface-level overlap. Among the RAD-Gym training variants, ECI-RL remains competitive across these complementary measures and achieves particularly strong clinical and lexical performance on several datasets. Together with the trajectory results, these findings suggest that optimizing evidence acquisition can improve grounded reporting without reducing evaluation to a single report-level score.

Table 5: **Complete agentic benchmark.** Report quality and evidence-seeking performance across four datasets. F-N denotes F1 × NPHR; all metrics are higher-is-better. CIDEr, BioBERT, and RadGraph rewards match the three GRPO baselines in the main tables.

Model	Size	Entity				Clinical			Organ		Trajectory				
		Prec.	Recall	F1	Acc	Loc.	Attr.	FC	Or.	FMOr.	EHR	EER	NPHR	SE	F-N
A CT-RATE															
Medical VLM															
Hulu-Med-7B	7B	7.12	6.88	6.91	33.64	0.91	0.91	0.91	9.41	6.48	0.50	0.20	35.38	0.03	2.44
Fleming-VL-8B	8B	9.45	9.02	9.12	13.38	1.84	0.50	0.50	9.92	8.62	0.34	0.30	0.00	0.01	0.00
Lingshu-I-8B	8B	13.37	12.28	12.53	22.47	3.37	2.25	1.12	14.34	12.34	0.63	0.79	38.89	0.04	0.00
MediX-R1-8B	8B	11.77	10.02	10.46	30.08	4.47	0.71	0.25	15.50	10.06	0.91	0.73	6.22	0.12	0.65
HealthGPT-Pro-8B	8B	16.74	12.33	13.38	30.25	14.33	7.79	7.71	15.82	14.54	3.19	2.25	47.50	0.43	6.36
MedMO-8B	8B	8.24	8.20	8.19	64.00	0.75	0.75	0.75	12.13	1.57	0.16	0.02	3.97	0.01	0.33
General VLM															
Qwen3.5-4B	4B	20.79	15.36	16.52	52.64	20.36	10.87	9.87	27.88	20.18	6.90	5.77	92.48	0.74	15.28
GPT-5.4-mini	–	*34.08*	*28.37*	*29.45*	*89.75*	*49.75*	*40.42*	*38.66*	*54.31*	*31.42*	*7.10*	*12.13*	*67.36*	*2.52*	*19.84*
Gemini-3-Flash	–	*32.19*	*24.62*	*26.24*	*78.50*	*40.58*	*25.37*	*23.99*	*42.83*	*28.62*	*8.50*	*8.79*	*55.47*	*1.15*	*14.56*
Claude-5-Sonnet	–	*17.16*	*13.64*	*14.38*	*68.25*	*19.25*	*10.82*	*10.41*	*24.22*	*14.34*	*4.29*	*3.71*	*58.00*	*0.38*	*8.34*
RAD-Gym Related															
SFT (ours)	4B	**33.89**	22.72	25.47	68.75	40.94	30.83	29.73	30.41	27.25	17.44	14.40	91.02	2.44	23.18
GRPO (CIDEr)	4B	31.71	22.57	24.83	70.43	37.82	28.57	27.20	32.24	27.16	16.29	14.39	92.16	2.48	22.88
GRPO (BioBert)	4B	32.68	22.52	24.81	71.50	40.10	31.25	28.37	32.79	28.60	**18.12**	14.04	**93.72**	2.42	23.25
GRPO (RadGraph)	4B	32.10	23.25	25.37	71.00	37.67	30.50	29.09	31.90	28.89	16.06	14.53	93.43	2.52	23.70
ECI-RL (ours)	4B	33.03	**24.59**	**26.64**	**73.75**	**44.80**	**34.88**	**33.67**	**38.26**	**31.48**	16.17	**17.07**	89.29	**2.83**	**23.79**
B AMOS-MM															
Medical VLM															
Hulu-Med-7B	7B	3.02	2.57	2.62	32.75	0.87	0.44	0.44	9.64	2.76	0.26	0.21	40.00	0.05	1.05
Fleming-VL-8B	8B	2.64	2.47	2.51	5.07	0.30	0.00	0.00	2.99	2.61	0.16	0.15	0.00	0.00	0.00
Lingshu-I-8B	8B	4.12	3.48	3.62	8.60	2.15	1.08	1.08	4.30	3.67	1.29	0.19	60.00	0.00	0.00
MediX-R1-8B	8B	3.80	3.14	3.28	21.36	2.43	0.50	0.50	6.44	3.15	0.18	0.32	13.73	0.05	0.43
HealthGPT-Pro-8B	8B	3.34	2.69	2.82	10.10	2.27	2.10	2.10	4.78	3.18	0.80	0.39	32.80	0.08	0.92
MedMO-8B	8B	2.83	2.81	2.60	43.34	0.57	1.13	0.57	15.16	2.11	0.26	0.08	1.99	0.02	0.05
General VLM															
Qwen3.5-4B	4B	5.98	3.45	3.99	17.09	6.28	3.89	3.27	5.43	4.48	1.05	0.76	79.66	0.09	3.18
GPT-5.4-mini	–	*18.65*	*15.29*	*15.69*	*96.23*	*34.23*	*29.01*	*27.82*	*44.79*	*20.40*	*4.21*	*7.57*	*65.80*	*1.57*	*10.32*
Gemini-3-Flash	–	*25.01*	*16.91*	*18.60*	*91.46*	*35.30*	*25.48*	*21.44*	*40.06*	*23.05*	*6.98*	*6.52*	*49.07*	*0.84*	*9.13*
Claude-5-Sonnet	–	*7.22*	*4.63*	*5.28*	*59.05*	*6.78*	*5.36*	*4.23*	*12.78*	*4.51*	*2.93*	*2.31*	*56.62*	*0.25*	*2.99*
RAD-Gym Related															
SFT (ours)	4B	20.65	11.86	13.80	86.68	26.07	21.90	19.77	23.81	15.35	**7.06**	**7.75**	86.03	**1.19**	11.87
GRPO (CIDEr)	4B	20.14	10.56	12.65	85.64	25.90	22.25	19.53	24.22	14.66	5.72	6.87	**92.18**	1.06	11.66
GRPO (BioBert)	4B	21.64	12.24	14.19	87.94	**30.15**	24.83	22.32	24.88	16.35	5.88	7.13	90.38	1.12	**12.82**
GRPO (RadGraph)	4B	22.26	11.27	13.65	88.92	29.70	25.07	**23.93**	23.92	14.86	6.06	6.79	91.09	1.06	12.43
ECI-RL (ours)	4B	**23.08**	**12.73**	**15.05**	**92.19**	28.40	**25.23**	22.83	**27.08**	**17.34**	6.57	7.25	84.65	1.05	12.74

Table 6: **Complete agentic benchmark (continued).** Report quality and evidence-seeking performance across four datasets. F-N denotes F1 $\times$ NPHR; all metrics are higher-is-better. CIDEr, BioBERT, and RadGraph rewards match the three GRPO baselines in the main tables.

Model	Size	Entity				Clinical			Organ		Trajectory				
		Prec.	Recall	F1	Acc	Loc.	Attr.	FC	Or.	FMOr.	EHR	EER	NPHR	SE	F-N
C Gastric-X															
Medical VLM															
Hulu-Med-7B	7B	3.35	1.27	1.78	57.14	9.18	6.12	6.12	8.96	2.38	0.70	0.78	22.17	0.10	0.00
Fleming-VL-8B	8B	0.00	0.00	0.00	6.25	0.00	0.00	0.00	0.77	0.00	0.00	0.25	0.00	0.01	0.00
Lingshu-I-8B	8B	2.33	0.87	1.26	15.24	3.81	1.43	0.48	2.12	1.21	1.05	0.37	18.75	0.03	0.00
MediX-R1-8B	8B	5.29	2.28	3.07	92.00	9.25	11.50	7.25	16.26	4.21	1.15	0.85	10.30	0.08	0.31
HealthGPT-Pro-8B	8B	1.08	0.43	0.53	19.46	0.54	0.00	0.00	3.21	0.66	0.69	0.44	21.21	0.05	0.18
MedMO-8B	8B	1.31	0.81	0.95	75.00	5.23	4.65	4.65	17.96	1.41	1.90	0.49	5.68	0.08	0.05
General VLM															
Qwen3.5-4B	4B	2.87	1.18	1.60	39.11	3.91	2.51	1.40	5.76	1.57	1.09	0.82	74.76	0.08	1.20
GPT-5.4-mini	–	*9.50*	*6.24*	*7.25*	*100.00*	*30.25*	*25.50*	*24.00*	*33.45*	*10.81*	*1.74*	*3.72*	*67.59*	*0.73*	*4.90*
Gemini-3-Flash	–	*15.22*	*7.42*	*9.52*	*98.00*	*29.33*	*26.75*	*21.00*	*29.55*	*12.91*	*2.07*	*2.18*	*49.42*	*0.16*	*4.70*
Claude-5-Sonnet	–	*5.72*	*2.16*	*2.98*	*67.00*	*9.50*	*9.00*	*7.50*	*14.18*	*3.81*	*1.98*	*1.72*	*57.21*	*0.21*	*1.70*
RAD-Gym Related															
SFT (ours)	4B	9.13	3.39	4.65	96.00	20.75	18.75	17.00	30.74	6.22	**7.97**	3.44	89.08	0.29	4.14
GRPO (CIDEr)	4B	9.55	4.11	5.31	93.97	18.34	14.41	13.90	31.39	6.73	6.11	2.74	89.92	0.23	4.77
GRPO (BioBert)	4B	9.51	3.90	5.32	96.48	19.60	18.09	15.58	31.86	6.45	6.91	4.22	**90.84**	0.35	4.83
GRPO (RadGraph)	4B	10.27	4.78	6.07	93.94	22.14	18.18	17.26	30.60	7.72	7.17	4.27	87.29	0.36	5.30
ECI-RL (ours)	4B	**13.27**	**6.13**	**7.89**	**99.00**	**30.04**	**27.54**	**25.29**	**39.16**	**11.52**	7.56	**4.61**	86.98	**0.39**	**6.86**
D MR-RATE															
Medical VLM															
Hulu-Med-7B	7B	17.11	21.75	17.97	60.00	4.12	7.06	2.94	57.65	12.94	0.68	1.37	35.75	0.17	0.00
Fleming-VL-8B	8B	13.59	13.53	13.55	20.51	0.00	0.00	0.00	19.87	12.82	0.25	0.00	6.67	0.00	0.90
Lingshu-I-8B	8B	16.01	15.59	15.73	19.29	1.43	1.43	1.43	19.29	16.43	3.08	1.37	100.00†	0.07	0.00
MediX-R1-8B	8B	15.96	15.93	15.93	30.67	0.83	0.50	0.33	29.83	16.00	0.00	0.17	10.00	0.02	1.72
HealthGPT-Pro-8B	8B	16.67	17.06	16.71	34.28	4.24	2.24	1.88	33.22	15.02	1.93	1.93	60.39	0.22	10.13
MedMO-8B	8B	17.14	17.17	17.07	**75.63**	0.00	0.84	0.00	**69.75**	1.68	0.00	0.27	4.59	0.06	0.78
General VLM															
Qwen3.5-4B	4B	19.89	19.91	19.59	40.37	6.94	7.35	6.94	39.75	23.29	4.41	4.11	**92.78**	0.39	18.18
GPT-5.4-mini	–	*29.22*	*31.74*	*29.51*	*80.33*	*25.11*	*21.69*	*19.44*	*76.83*	*39.00*	*3.09*	*7.05*	*80.12*	*1.42*	*23.64*
Gemini-3-Flash	–	*26.54*	*32.32*	*27.93*	*82.00*	*20.19*	*16.31*	*14.22*	*78.83*	*31.50*	*7.55*	*8.55*	*71.02*	*0.81*	*19.84*
Claude-5-Sonnet	–	*17.59*	*17.11*	*17.11*	*35.33*	*3.61*	*2.00*	*1.50*	*34.50*	*19.83*	*1.25*	*0.45*	*80.30*	*0.06*	*13.74*
RAD-Gym Related															
SFT (ours)	4B	27.66	27.99	26.88	63.51	18.19	18.58	15.54	61.32	32.94	10.10	10.09	90.30	0.87	24.27
GRPO (CIDEr)	4B	27.17	25.99	25.72	62.54	18.31	16.08	14.52	60.37	31.94	12.45	10.44	90.55	0.95	23.29
GRPO (BioBert)	4B	**28.21**	27.06	26.71	62.33	19.61	17.72	15.50	59.50	32.17	10.62	9.82	87.49	0.87	23.37
GRPO (RadGraph)	4B	28.14	28.16	27.20	59.33	19.69	16.94	14.83	56.83	32.17	10.28	9.82	91.40	0.88	**24.86**
ECI-RL (ours)	4B	28.07	**29.02**	**27.51**	62.33	**21.97**	**19.00**	**16.81**	59.83	**33.83**	**12.77**	**13.21**	88.96	**1.20**	24.47

Table 7: **Complete report generation benchmark.** Clinical report quality across four datasets, organized by model family and training recipe.

Model	Type	CT-RATE				AMOS-MM			
		CIDEr	RaTEScore	GREEN	BioBert	CIDEr	RaTEScore	GREEN	BioBert
Medical VLM									
Hulu-Med-7B	LLM	26.67	44.35	4.87	67.63	28.71	45.61	6.82	70.53
Fleming-VL-8B	LLM	24.43	50.86	15.81	69.04	27.17	50.24	5.48	71.44
Lingshu-I-8B	LLM	23.50	47.79	11.56	69.15	19.16	35.50	2.42	66.98
MediX-R1-8B	LLM	26.68	51.63	16.57	69.55	**28.81**	51.17	14.48	**71.96**
MedMO-8B	LLM	4.52	23.33	1.37	57.44	6.64	25.96	0.35	59.02
General VLM									
Qwen3.5-4B	LLM	25.20	52.90	23.10	67.72	25.39	52.64	14.42	69.07
Qwen3.5-4B	Agent	24.38	53.98	22.69	69.16	19.79	53.93	13.88	69.41
GPT-5.4-mini	*Agent*	*20.61*	*49.68*	*15.03*	*69.03*	*17.44*	*53.01*	*12.48*	*69.77*
Gemini-3-Flash	*Agent*	*31.10*	*55.97*	*18.30*	*70.62*	*34.44*	*58.05*	*8.18*	*71.91*
Claude-5-Sonnet	*Agent*	*24.54*	*53.16*	*12.28*	*69.59*	*22.83*	*52.88*	*6.32*	*69.92*
RAD-Gym Related									
SFT (ours)	Agent	27.81	55.18	29.49	71.02	22.58	56.85	**23.81**	70.63
GRPO (CIDEr)	Agent	**28.23**	55.24	28.65	**71.07**	22.39	56.73	22.64	70.63
GRPO (BioBert)	Agent	28.00	55.25	28.42	71.01	22.40	56.81	21.43	70.67
GRPO (RadGraph)	Agent	27.86	**55.31**	28.45	71.01	22.47	56.87	22.48	70.66
ECI-RL (ours)	Agent	26.82	55.30	**29.55**	70.87	22.20	**56.95**	21.85	70.55

Model	Type	Gastric-X				MR-RATE			
		CIDEr	RaTEScore	GREEN	BioBert	CIDEr	RaTEScore	GREEN	BioBert
Medical VLM									
Hulu-Med-7B	LLM	**35.18**	49.38	**8.87**	70.96	21.86	39.71	6.37	65.69
Fleming-VL-8B	LLM	24.73	45.10	2.15	69.55	20.75	52.23	**30.38**	67.14
Lingshu-I-8B	LLM	24.04	32.11	0.38	67.12	27.36	48.85	14.88	68.56
MediX-R1-8B	LLM	24.03	45.98	2.47	69.44	24.51	49.21	18.65	68.79
HealthGPT-Pro-8B	LLM	33.82	49.74	4.76	**71.51**	22.65	45.58	21.42	**69.67**
MedMO-8B	LLM	0.84	42.73	0.78	62.35	7.50	30.98	10.40	59.91
General VLM									
Qwen3.5-4B	LLM	27.92	46.78	3.07	67.31	29.67	**52.99**	29.13	67.77
Qwen3.5-4B	Agent	23.12	46.55	4.84	68.60	21.39	49.24	30.03	67.89
GPT-5.4-mini	*Agent*	*24.27*	*47.62*	*6.08*	*70.15*	*29.97*	*49.28*	*20.22*	*69.51*
Gemini-3-Flash	*Agent*	*40.81*	*51.79*	*1.89*	*71.23*	*33.00*	*49.84*	*9.93*	*69.87*
Claude-5-Sonnet	*Agent*	*24.60*	*45.99*	*0.42*	*69.05*	*25.82*	*52.98*	*19.16*	*69.69*
RAD-Gym Related									
SFT (ours)	Agent	32.65	**52.44**	6.26	70.54	**30.21**	49.40	27.28	68.93
GRPO (CIDEr)	Agent	33.10	52.21	5.70	70.57	29.61	49.32	26.14	68.88
GRPO (BioBert)	Agent	33.35	52.43	6.13	70.72	30.09	49.50	26.20	68.90
GRPO (RadGraph)	Agent	32.86	52.18	6.94	70.54	29.73	49.26	25.44	68.85
ECI-RL (ours)	Agent	30.59	52.18	6.42	70.28	**30.21**	49.08	23.88	68.97

LLM: single-turn reporting; Agent: tool-augmented reporting. **Bold**: best; underline: second best eligible result within each dataset. *Italic*: proprietary general VLM, shown for reference and excluded from ranking. —: unreported.

Table 8: Additional report-generation scores.

Model	Type	RadGraph	BLEU-1	BLEU-4	ROUGE-L	METEOR
CT-RATE						
Qwen3.5-4B	LLM	11.04	19.80	2.20	14.00	18.86
Fleming-VL-8B	LLM	8.59	16.71	1.66	13.97	17.14
Qwen3.5-4B	Agent	10.91	14.63	1.69	11.39	14.50
SFT (ours)	Agent	10.44	12.61	1.60	11.70	13.38
GRPO (CIDEr)	Agent	10.51	12.61	1.64	11.73	13.39
GRPO (BioBert)	Agent	10.46	12.50	1.62	11.66	13.31
GRPO (RadGraph)	Agent	10.41	12.48	1.59	11.66	13.28
ECI-RL (ours)	Agent	11.21	12.92	1.62	11.29	13.37
Hulu-Med-7B	LLM	5.43	12.17	1.50	11.99	13.69
HealthGPT-Pro-8B	LLM	27.84	38.11	22.89	37.17	43.50
GPT-5.4-mini	*Agent*	*6.43*	*8.33*	*1.12*	*9.67*	*8.76*
Claude-5-Sonnet	*Agent*	*10.66*	*19.99*	*1.87*	*12.44*	*24.39*
Gemini-3-Flash	*Agent*	*11.36*	*22.88*	*2.57*	*14.94*	*22.02*
MediX-R1-8B	LLM	7.91	20.36	2.41	14.19	19.70
MedMO-8B	LLM	0.22	1.71	0.21	4.62	3.97
AMOS-MM						
Qwen3.5-4B	LLM	11.65	23.80	2.67	17.33	24.44
Fleming-VL-8B	LLM	9.12	21.24	2.68	18.29	21.10
Qwen3.5-4B	Agent	10.03	16.18	1.87	12.86	15.10
SFT (ours)	Agent	11.09	15.58	2.02	12.90	14.36
GRPO (CIDEr)	Agent	11.23	15.69	1.99	12.81	14.53
GRPO (BioBert)	Agent	11.27	15.70	1.97	12.84	14.41
GRPO (RadGraph)	Agent	11.21	15.74	2.00	12.87	14.46
ECI-RL (ours)	Agent	11.53	16.63	2.09	12.86	15.19
Hulu-Med-7B	LLM	7.76	12.74	1.76	14.66	14.78
HealthGPT-Pro-8B	LLM	5.41	3.83	0.57	10.63	7.59
GPT-5.4-mini	*Agent*	*7.52*	*11.74*	*1.44*	*11.07*	*10.77*
Claude-5-Sonnet	*Agent*	*10.38*	*20.29*	*2.02*	*13.81*	*26.44*
Gemini-3-Flash	*Agent*	*17.10*	*29.01*	*3.87*	*19.82*	*33.33*
MediX-R1-8B	LLM	10.85	24.66	2.91	18.04	24.12
MedMO-8B	LLM	1.69	3.00	0.30	7.42	6.66
Gastric-X						
Qwen3.5-4B	LLM	9.07	25.11	3.64	15.83	25.04
Fleming-VL-8B	LLM	7.80	22.04	2.87	17.32	19.91
Qwen3.5-4B	Agent	9.66	19.06	2.50	13.08	16.36
SFT (ours)	Agent	11.58	20.55	4.02	14.38	17.71
GRPO (CIDEr)	Agent	11.53	21.10	4.22	14.39	18.12
GRPO (BioBert)	Agent	12.15	20.82	4.22	14.48	17.92
GRPO (RadGraph)	Agent	11.72	20.65	4.10	14.30	17.59
ECI-RL (ours)	Agent	11.80	21.83	4.15	14.02	18.76
Hulu-Med-7B	LLM	5.97	7.11	1.16	12.62	12.34
HealthGPT-Pro-8B	LLM	9.44	12.84	3.06	16.14	16.54
GPT-5.4-mini	*Agent*	*9.08*	*20.33*	*2.49*	*13.33*	*16.38*
Claude-5-Sonnet	*Agent*	*9.04*	*19.76*	*2.07*	*13.90*	*28.03*
Gemini-3-Flash	*Agent*	*12.60*	*28.52*	*5.86*	*19.12*	*33.03*
MediX-R1-8B	LLM	6.20	23.90	3.15	16.18	22.38
MedMO-8B	LLM	2.55	2.83	0.30	4.72	8.01
MR-RATE						
Qwen3.5-4B	LLM	13.14	20.99	2.61	14.00	22.01
Fleming-VL-8B	LLM	9.76	16.70	1.95	14.40	18.76
Qwen3.5-4B	Agent	7.24	14.43	1.88	10.77	13.35
SFT (ours)	Agent	9.27	18.28	2.42	12.47	16.75
GRPO (CIDEr)	Agent	9.14	17.93	2.36	12.43	16.39
GRPO (BioBert)	Agent	9.25	18.00	2.41	12.52	16.48
GRPO (RadGraph)	Agent	9.23	17.98	2.38	12.41	16.56
ECI-RL (ours)	Agent	9.32	18.94	2.46	12.68	17.40
Hulu-Med-7B	LLM	3.72	11.02	1.46	10.47	10.83
HealthGPT-Pro-8B	LLM	8.42	9.26	1.22	10.92	11.17
GPT-5.4-mini	*Agent*	*8.79*	*14.63*	*2.04*	*12.06*	*13.70*
Claude-5-Sonnet	*Agent*	*13.96*	*17.27*	*1.73*	*11.73*	*26.91*
Gemini-3-Flash	*Agent*	*12.26*	*22.73*	*2.94*	*14.35*	*25.23*
MediX-R1-8B	LLM	5.96	16.03	1.89	11.63	15.10
MedMO-8B	LLM	0.85	3.69	0.55	6.03	5.41

Scores are multiplied by 100. Italic proprietary models are reference only. Dash: unreported.

A.2 Extended Reward Evaluation on Mixed Datasets

Table 9 expands Table 3. Mix-400 contains 100 selected test cases from each dataset, with equal dataset weights. Report means use available evaluations without imputing missing scores. Full/coarse-only runs average 9.0/8.7 rounds and 26.2/24.3 slice presentations per case: this comparison does not isolate the contributions of additional viewing and action choice. The No-G, No-I, joint-terminal-only, and no-temporal-preference results extend the comparison. Joint terminal only retains evidence gating but removes the additional process reward; No-I uses update 25. The full reward achieves the highest F1 (19.78), FC-Rate (26.16), and F-N (17.79). No temporal preference, evaluated at update 10, has slightly higher EER (10.58 versus 10.50) and SE (1.38 versus 1.31), but lower F1 and F-N. No-G has higher NPHR (92.06 versus 88.03), illustrating why a higher supported-claim fraction alone does not imply better joint report quality. These are comparisons of the reported checkpoints, not estimates of isolated component effects under equal optimization progress.

Table 9: **Extended Mix-400 reward ablations.** Report and process metrics for the settings in Table 3, averaged equally across four 100-case subsets. F-N denotes F1 $\times$ NPHR. All scores are scaled by 100; higher is better.

Setting	Prec.	Recall	F1	Loc.	Attr.	FC	Or.	FMOr.
ECI-RL (joint, full)	**25.61**	**18.08**	**19.78**	**32.87**	**27.87**	**26.16**	**41.78**	**24.69**
Coarse report only	24.00	15.30	17.39	27.81	21.05	19.10	35.80	21.62
Joint terminal only	23.41	15.06	17.14	29.31	22.65	21.65	36.77	19.24
No G assessment	24.01	16.34	18.22	28.67	25.79	24.04	37.13	20.92
No I / procedural anchors	24.65	15.89	17.78	29.04	24.10	22.35	36.51	20.35
No temporal preference	25.38	18.00	19.45	31.53	26.05	24.96	40.41	22.15

Setting	EHR	EER	NPHR	SE	F-N
ECI-RL (joint, full)	**11.07**	10.50	88.03	1.31	**17.79**
Coarse report only	8.95	7.70	91.19	1.01	15.88
Joint terminal only	7.78	8.14	87.83	1.13	15.28
No G assessment	9.23	8.48	**92.06**	1.12	17.05
No I / procedural anchors	8.56	7.82	87.35	1.04	15.87
No temporal preference	10.92	**10.58**	88.38	**1.38**	17.10

Each dataset contributes 100 selected cases and one quarter of the aggregate weight. F-N uses the reported aggregate product, not the product of displayed aggregate F1 and NPHR. Bold marks the best observed score. The temporal ablation uses the reported update-10 result; No-I uses update 25.

A.3 Acquisition and Reasoning Audit

Figure 3 compares the GPT-5.5 teacher, Qwen3.5-4B base, SFT, and ECI-RL. Teacher training cohorts contain 3200/1237/1188/2701 trajectories for CT-RATE/AMOS-MM/Gastric-X/MR-RATE, respectively; student test cohorts contain 400/398/200/300 trajectories for the corresponding datasets. These are descriptive comparisons across unequal cohorts.

Acquisition and completion (a,b,e). The scan geometry metric is evaluated on CT-RATE and AMOS-MM, which are the single-volume datasets. Slice presentations include repeats; the first-view span is an index span, not a normalized anatomical coverage. Arrows connect model means rather than paired cases. Local zoom counts view calls with nontrivial bounding boxes. Submission measures execution reliability, not diagnostic accuracy.

Reasoning content and diversity (c,d). While GPT-5.5 generates the primary trajectories, Qwen3.8-Max synthesizes the teacher reasoning. GPT-5.4-mini labels 576 sampled reasoning steps, with 36 per model-dataset stratum; visual interpretation is evaluated using only the view-step subset. Labels capture textual content, not image-verified correctness. We interpret these annotations descriptively rather than as paired evidence of diagnostic correctness. "Opening reuse" refers to the range of top-five opening-phrase coverage across datasets: teacher 1.3–2.5%, base 30.5–61.2%, SFT 9.9–31.9%, and ECI-RL 11.8–33.5%. These ranges represent observed spreads, not confidence intervals.

A.4 Counterfactual Cohorts and Report Rankings

The report-ranking experiment in Table 4 uses 24 Flash and 69 GPT cases. Acquisition is held fixed while the final report is changed. The cohorts are generated by Qwen3.5-Flash and GPT-5.5, respectively.

Each insertion, deletion, or substitution adds one verifier-approved atomic error while holding acquisition fixed; chains stop when no valid edit remains. A separate paraphrase preserves clinical content. RadGraph uses RadGraph-XL F1; BioBERT uses BERTScore with `dmis-lab/biobert-v1.1`, layer 12. For agreement ϕ and tie rate η, the displayed summaries use fixed included-cohort weights:

$$\bar{\phi} = \frac{24\phi_{\text{Flash}} + 69\phi_{\text{GPT}}}{93}, \qquad \bar{\eta} = \frac{24\eta_{\text{Flash}} + 69\eta_{\text{GPT}}}{93}. \tag{A.1}$$

These fixed cohort weights produce descriptive group summaries, not pooled per-case estimates; eligible cases can differ across edit chains. Half credit is assigned to ties. Table 10 retains the original group intervals; their endpoints are not averaged. ROUGE-L and CIDEr supply lexical comparisons (Lin, 2004; Vedantam et al., 2014).

Table 10: **Report-ranking agreement by source group.** Flash and GPT estimates underlying Table 4, with their reported intervals and tie rates. Agreement measures whether a metric orders verified report corruptions correctly; ties receive half credit.

Metric	Flash			GPT		
	$\phi \uparrow$	Interval	Tie rate	$\phi \uparrow$	Interval	Tie rate
ROUGE-L	0.368	[0.270, 0.464]	0.10	0.469	[0.392, 0.547]	0.07
CIDEr	0.310	[0.206, 0.419]	0.03	0.421	[0.345, 0.499]	0.03
RadGraph	0.291	[0.223, 0.357]	0.10	0.445	[0.364, 0.530]	0.07
BioBERT	0.252	[0.129, 0.382]	0.03	0.383	[0.308, 0.464]	0.03
CABS-F1	0.561	[0.473, 0.650]	0.16	0.507	[0.425, 0.586]	0.12
NPHR $\times$ CABS-F1	0.657	[0.587, 0.719]	0.08	0.611	[0.528, 0.688]	0.07

Group summaries are retained for audit. The main table weights point estimates and tie rates by 24:69. Intervals are reproduced as reported and are not averaged into a combined interval.

The comparison addresses ordering under verified edits, not universal metric superiority or calibration. Paraphrases are a separate control and do not enter the corruption ordering. We do not infer a significance test for the weighted difference from the individual metric intervals.

A.5 Additional Tool-Interface Pilots

These experiments complement Figure 4 with retrieval and workflow comparisons. Each configuration reports 100 cases. The two appendix figures show relative changes, $100(s/s_{\text{ref}} - 1)$, computed from aggregate scores using the same model, dataset, and version's reference. The reference is the full autonomous toolset for retrieval and the autonomous agent with the same toolset for workflow. Zero denotes the reference; positive values indicate higher scores. Both metric panels within each figure share a linear scale, and labels sit outside the bars.

The shaded $\pm 5\%$ band is descriptive, not a confidence or equivalence interval. GASTRIC-X retrieval and CT-RATE workflow use legacy RadGraph entity scores; their absolute values are not comparable with newer RadGraph-XL F1 scores. The original tables provide no CABS for these contrasts. These historical pilots motivate interface design rather than establish the effect of the finalized ECI-RL recipe. The full-toolset reference is the historical pilot configuration, not the final four-action interface. GPT and Flash CABS use different historical judge settings, so their absolute values are not compared across models. Figure 5 shows that removing web search has opposite effects on GPT's two datasets. Retrieval should therefore be assessed for the reporting task rather than assumed to help by default. In Figure 6, guided execution lowers both displayed scores with the same toolset, supporting model-directed acquisition over the tested fixed workflows.

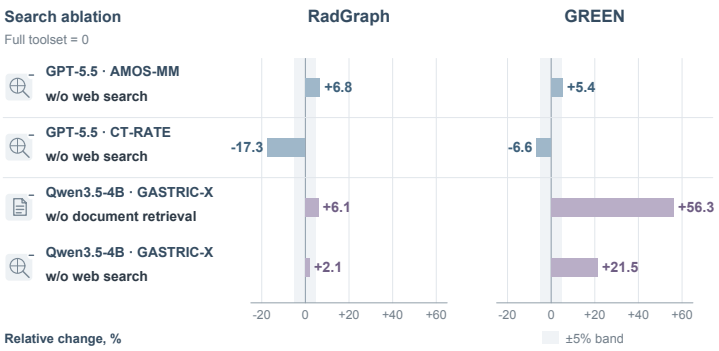


Figure 5: **External retrieval pilots.** RadGraph and GREEN changes relative to the matched full-toolset reference. Blue denotes GPT-5.5; lavender denotes Qwen3.5-4B. Removing web search improves both displayed metrics on AMOS-MM but lowers them on CT-RATE. On GASTRIC-X, removing either retrieval mechanism improves both. Retrieval effects therefore depend on the setting.

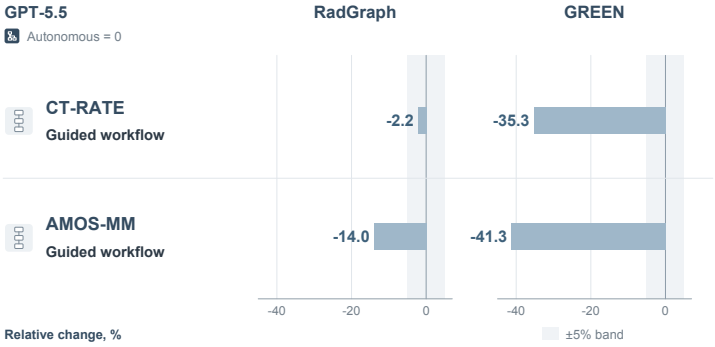


Figure 6: **Autonomous versus guided reporting.** GPT-5.5 with the same toolset, normalized to the matched autonomous agent. Guided execution lowers GREEN by 35.3% on CT-RATE and 41.3% on AMOS-MM; RadGraph falls by 2.2% and 14.0%, respectively. These comparisons favor autonomous action selection over the tested guided policies.

A.6 Evaluation Protocol and Reproducibility

The benchmarking protocol in Section 3 separates policy generation from evaluation. Saved trajectories retain the final report, termination status, ordered actions, and the file, slice, and crop associated with each image. Requested and executed actions are distinguished so that environment corrections are not attributed to the policy. The same trajectories can therefore be rescored under shared evaluator settings.

Comparable evaluation. The comparison uses fixed test lists and common metric definitions. Reported means use eligible evaluations for each model and metric; they are not necessarily paired over identical completed-case subsets. ECI-RL additionally records observation availability and generation-block boundaries for the reward and loss in Equations (7) and (9). Sections B.4 and D supplies the corresponding optimization and credit definitions.

B Implementation Details

B.1 Case Representation and Preprocessing

A case has one reference report shared by its volumes, with modality, phase or sequence, legal slice indices, and localization metadata. Volumes use RAS orientation, 384×384 in-plane arrays, and grayscale values in $[0, 255]$. CT-RATE and AMOS-MM reuse their single-volume preprocessing. Gastric-X provides non-contrast, arterial, venous, and delayed CT, each sampled to 32 slices after

clipping HU to $[-1000, 1000]$. Phases are not registered: matching indices need not identify matching anatomy.

MR-RATE normalizes each sequence independently by its 5th–95th percentiles. Series with at most 128 slices are sampled to 32, and those with 129–256 to 64; longer series and MRA are excluded. MRI intensity values are therefore not directly comparable across sequences. CT localization uses a 117-class anatomical taxonomy, while MRI uses brain/body taxonomies. Localization hints are anatomical rather than diagnostic labels and require image verification.

B.2 Actions and Observation Budgets

The four actions follow Section 2.2. Viewing accepts a case-local file, slice indices, and an optional normalized crop. Requests are bounded to valid indices and the image allowance, with corrections recorded explicitly. Localization returns anatomical ranges and boxes; writing updates the draft; submission requires a valid draft. At the round limit, an existing draft is submitted automatically; otherwise the episode is invalid.

Environment defaults are 20 rounds, six slices per view, 32 images per trajectory, 3,000 report characters, and minimum crop area 0.04. Fixed evaluation lists require reports, images, and localization metadata. Mix-400 selects the first 100 test cases from each dataset. The single-turn baseline samples uniformly for single-volume studies and four slices per volume for multi-volume studies; its observation budget is recorded separately from adaptive acquisition.

B.3 Metric Implementation Details

Clinical Units and One-to-one Matching. Following (Lin et al., 2026a), we represent a clinical finding as

$$f = \langle o, d, \mathcal{U}, \ell, \chi, e \rangle, \tag{B.1}$$

where o is the organ, d the finding, $\mathcal{U}$ its attributes, ℓ its location, χ certainty, and e the report evidence span. The evidence span is textual provenance, distinct from the image evidence required by Equation (3). The benchmark parser retains abnormal findings and excludes normal or negated statements. Organ names are normalized to the project taxonomy before matching.

For reference unit $f_i^{\mathrm{G}} \in \mathcal{G}$ and predicted unit $f_j^{\mathrm{P}} \in \mathcal{P}$, let h_{ij} indicate organ agreement and semantic finding equivalence; let l_{ij} and a_{ij} indicate location and attribute agreement. The default Hungarian matcher selects a maximum-weight one-to-one assignment:

$$\begin{aligned} h_{ij} &= \mathbf{1}[o_i^{\mathrm{G}} = o_j^{\mathrm{P}}]\, \mathbf{1}[d_i^{\mathrm{G}} \simeq d_j^{\mathrm{P}}], \\ S_{ij} &= h_{ij}(l_{ij} + a_{ij}) + \eta h_{ij}, \qquad \eta = 0.01, \\ \mathcal{M}^\star &\in \operatorname*{argmax}_{\mathcal{M} \in \mathfrak{M}} \sum_{(i,j)\in\mathcal{M}} S_{ij}, \\ \mathcal{M}_+ &= \{(i,j) \in \mathcal{M}^\star : h_{ij} = 1\}, \qquad n = |\mathcal{M}_+|. \end{aligned} \tag{B.2}$$

Here $\mathfrak{M}$ contains assignments with no repeated reference or prediction index. The small hit term favors a valid finding match over a nonmatch when both detail scores are zero. Semantic normalization and matcher settings are fixed across systems.

Report Metrics. With $N_{\mathrm{G}} = |\mathcal{G}| > 0$ and $N_{\mathrm{P}} = |\mathcal{P}| > 0$, finding-level precision, recall, and F1 are

$$\mathrm{Precision} = \frac{n}{N_{\mathrm{P}}}, \qquad \mathrm{Recall} = \frac{n}{N_{\mathrm{G}}}, \qquad \mathrm{F1} = \frac{2\,\mathrm{Precision}\,\mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}. \tag{B.3}$$

Detail accuracy is conditional on a matched finding:

$$\begin{aligned} \mathrm{LocAcc} &= \frac{1}{n} \sum_{(i,j)\in\mathcal{M}_+} l_{ij}, \qquad \mathrm{AttrAcc} = \frac{1}{n} \sum_{(i,j)\in\mathcal{M}_+} a_{ij}, \\ \mathrm{FCRate} &= \frac{1}{n} \sum_{(i,j)\in\mathcal{M}_+} l_{ij} a_{ij}. \end{aligned} \tag{B.4}$$

Let $\mathcal{O}_{\mathrm{G}}$ and $\mathcal{O}_{\mathrm{P}}$ be the organs represented in the reference and prediction, and let $\mathcal{O}_{\mathcal{M}}$ be reference organs with at least one matched finding. Organ coverage and finding-matched organ coverage are

$$\mathrm{OrRate} = \frac{|\mathcal{O}_{\mathrm{G}} \cap \mathcal{O}_{\mathrm{P}}|}{|\mathcal{O}_{\mathrm{G}}|}, \qquad \mathrm{FMOrRate} = \frac{|\mathcal{O}_{\mathrm{G}} \cap \mathcal{O}_{\mathcal{M}}|}{|\mathcal{O}_{\mathrm{G}}|}. \tag{B.5}$$

Report-level abnormality-presence accuracy is

$$\mathrm{Acc}_{+} = \mathbf{1}\big[\mathbf{1}[N_{\mathrm{G}}^{+} > 0] = \mathbf{1}[N_{\mathrm{P}}^{+} > 0]\big], \tag{B.6}$$

where N_{G}^{+} and N_{P}^{+} count positive abnormalities. For abnormality-only evaluation they equal N_{G} and N_{P}.

Final-report Score R_{CABS}. For reference finding i, take h_i, l_i, a_i from its assigned pair, with all three set to zero if unmatched. The CABS evaluator also defines the following reference-normalized scalar

$$\begin{aligned} r_i^{\mathrm{CABS}} &= h_i \left(l_i a_i + \frac{1}{2}|l_i - a_i| \right), \\ R_{\mathrm{CABS}}(y, y^{\star}) &= \frac{1}{N_{\mathrm{G}}} \sum_{i=1}^{N_{\mathrm{G}}} r_i^{\mathrm{CABS}} \in [0,1], \qquad N_{\mathrm{G}} > 0. \end{aligned} \tag{B.7}$$

A matched finding earns 1 when both location and attributes match, $1/2$ when exactly one matches, and zero otherwise. Unmatched references earn zero. This scalar is documented for metric interpretation; it is not one of the three GRPO reward baselines in the main tables. Unmatched predictions are not directly penalized by this reference-normalized scalar, although they lower precision. Consequently, R_{CABS} and ECI-RL's F in Equation (5) are distinct objectives.

Separation from training reward. The joint ECI-RL reward uses frozen claim slots, a report-only quality judge, and observed-evidence acceptance (Equation (5)), rather than CABS precision/recall. Task-relevant negative claims can enter that training ledger; the abnormality benchmark remains unchanged.

Empty sets and aggregation. When both unit sets are empty, precision, recall, and F1 equal one; when exactly one is empty they all equal zero. F1 is zero when precision and recall sum to zero. With no matched findings, LocAcc, AttrAcc, and FCRate are zero. If $\mathcal{O}_{\mathrm{G}}$ is empty, both organ rates equal one only when $\mathcal{O}_{\mathrm{P}}$ is also empty, and zero otherwise. In our implementation, R_{CABS} evaluates to zero when the reference set is empty (this includes cases where a normal report is correctly predicted). These abnormality-only conventions do not define the joint training score. For a metric m with eligible evaluation cases $\mathcal{D}_m$, we report its case-level macro-average:

$$\overline{m} = \frac{1}{|\mathcal{D}_m|} \sum_{c \in \mathcal{D}_m} m(c). \tag{B.8}$$

Parsing failures are distinct from valid reports with no abnormal findings.

Additional Process Metrics. Let N_a count executed tool calls and $\mathcal{H}(f, z_t) \in \{0,1\}$ indicate an explicit pre-action hypothesis about finding f, judged against its preceding context. Using the same $\mathcal{G}$ and $\mathcal{E}$ as in Equation (3),

$$\mathrm{EHR} = \frac{\sum_{f \in \mathcal{G}} \mathbf{1}[\exists t : \mathcal{H}(f, z_t) = 1]}{|\mathcal{G}|}, \quad \mathrm{SE} = \mathrm{EER}/N_a, \quad \mathrm{PHR} = 1 - \mathrm{NPHR}. \tag{B.9}$$

EHR measures hypothesized reference coverage, SE evidence coverage per executed action, and PHR unsupported predicted claims. Evidence matching requires a real nonempty image observation and a supported association in reasoning or a draft. Same-organ viewing and textual agreement alone do not suffice. Undefined denominators are reported separately from the report-metric empty-set rules.

B.4 Training Details

Teacher-guided Supervision For Equation (4), GPT-5.5 supplies actions and Qwen3.8-Max synthesizes pre-action reasoning with observations revealed in temporal order. The reference report is available only to the terminal supervision audit, never to the policy. Filtering removes invalid

or erroneous trajectories, implausible actions, missing reasoning, excessive repetition, and reports outside one half to twice the reference length or the absolute cap.

Qwen3.5-4B is trained with LLaMA-Factory (Zheng et al., 2024). Assistant reasoning and calls have mask $m = 1$; prompts, images, and observations have $m = 0$. Images are reconstructed from saved acquisition records, and validation uses a fixed held-out list.

Generation-block Policy Optimization ECI-RL uses a verl multi-turn rollout adapter. One generation block comprises a policy response and its calls; all generated tokens share the block advantage in Equation (8). For actual blocks $\mathcal{I} = \{(k,t) : 1 \le k \le K,\ 1 \le t \le T_k\}$,

$$\bar{A} = |\mathcal{I}|^{-1} \sum_{\mathcal{I}} A_{k,t}, \qquad \sigma^2_{\text{group}} = |\mathcal{I}|^{-1} \sum_{\mathcal{I}} (A_{k,t} - \bar{A})^2. \tag{B.10}$$

Missing peer blocks contribute zero to the leave-one-out baseline, but padding is excluded from this variance. In Equation (9), the sampled KL term is $\mathcal{K}_\ell = v_\ell - 1 - \log v_\ell$, where $v_\ell = \pi_{\text{ref}}(x_\ell \mid u_\ell)/\pi_\theta(x_\ell \mid u_\ell)$ and π_{ref} is frozen. Z is the loss normalizer; observation tokens receive no policy loss.

Reported models. SFT uses checkpoint 186, and ECI-RL uses the ATR5 LoRA policy after 20 updates. The latter has 400/397/200/300 evaluated cases on CT-RATE/AMOS-MM/Gastric-X/MR-RATE. Mix-400 is evaluated separately (Section A.2); its full and coarse-only runs use LoRA rank 32 with the same data and training configuration.

C Prompt Templates

We provide the exact prompt strings used for policy execution, reasoning synthesis, and report evaluation. We normalize dash typography strictly for typesetting consistency. The policy receives case-local file descriptions and tool observations directly. We supply the reference report exclusively to the submission audit and evaluators.

C.1 Policy and Tool Instructions

For the interface defined in Section 2.2, we construct the user message by combining the task description, tool instructions, case-specific imaging-file inventory, and termination rules. The inventory enumerates each file's modality, sequence or phase, slice range, and image dimensions. We append a final directive instructing the agent to locate organs, inspect images, and submit the report sequentially.

Policy system role

```
You are an expert radiologist AI agent. You write radiology reports for 3D medical
↪  imaging studies by actively inspecting the images with the provided tools, reasoning
↪  step by step.
```

Policy task description

```
# Task
You are given one radiology case with {n_volumes} 3D imaging file(s). Your job is to
↪  inspect the images and write a clinical radiology report.

Rules:
- You have at most {max_turns} interaction rounds (tool-call turns). Use them wisely.
- The task ends ONLY when you call `submit`, which submits the latest report you wrote via
↪  `write`.
- The final report must be at most {max_report_chars} characters.
- Work like a radiologist: first get an overview (organ locations), then inspect
↪  suspicious slice ranges, then draft and revise the report, then submit.
- Note on image delivery: images produced by tools are returned to you inside a following
↪  user message (each image preceded by a short [file][slice i/D] label). This is a
↪  transport detail of the environment --- treat those images as the tool's visual
↪  output.
```

Policy tool definitions

```
# Tools
1. `view_slice(file_name, slice_indices, bbox=None)` --- view specific 0-based slices of a
↪  file as 2D grayscale images.
   - Do NOT request too many slices at once (at most {max_slices_per_call} per call; at
   ↪  most {max_images_per_traj} images in total for the whole task).
```

Policy tool definitions (continued)

```
   - Prefer to call `locate_organ` first and use its slice ranges to choose slices.
   - `bbox` optionally crops each slice with a normalized box [x1, y1, x2, y2] (values in
   ↪  0-1); the crop is resized back to the original size.
   - The tool message returns only a short text confirmation (which slices were
   ↪  returned); the images themselves arrive in the following user message.
2. `locate_organ(file_name)` --- returns a summary of anatomical structures identified in
↪  a file. Fields per structure:
   - `Class`: organ/structure name;
   - `Slice range`: the slice interval (0-based, end-exclusive style `start--end / D`)
   ↪  containing the structure --- use it to guide `view_slice`;
   - `Bbox (x,y,z)`: 3D bounding box `[x,y,z] -> [x,y,z]` in volume coordinates;
   - `Centroid`: (x, y, z) center of the structure;
   - `Voxel count`: size of the structure (and its percentage of the volume);
   - `Max cross-section`: the slice index where the structure has its largest area.
   - IMPORTANT: this summary is generated by an automatic analysis pipeline. It may be
   ↪  inaccurate or incomplete, and it may occasionally find no structures at all. Treat
   ↪  it strictly as a rough reference for choosing slices --- never as ground truth;
   ↪  always confirm by viewing the images yourself.
3. `write(text)` --- write/overwrite the draft report (each call replaces the previous
↪  version).
4. `submit()` --- submit the LATEST `write` report as the final answer and end the task.
↪  You must call `write` at least once before `submit`; if the draft feels incomplete,
↪  call `write` again to revise it before submitting.{text_tool_note}
```

We apply an optional text-tool suffix for models receiving textual calls instead of a native tool schema.

Optional text-tool suffix

```
To use a tool, output a tool call in your reply in this exact format:
<tool_call>{"name": "<tool_name>", "arguments": {<arguments as JSON>}}</tool_call>
One tool call per reply. Then wait for the tool result.
```

MRI processing note

```
Note: these MRI volumes were normalized per sequence with percentile scaling to 0-255.
↪  Grayscale intensities are NOT comparable across sequences or across cases; judge
↪  signal as high/low only relatively within the same image.
```

Multi-phase CT processing note

```
Note: this case contains multiple CT phases of the same study. The phases are NOT
↪  co-registered; the same slice index in different phases may not correspond
↪  anatomically.
```

Termination rules

```
# Termination
- The task ends when `submit` succeeds.
- If you reach the {max_turns}-round limit without submitting, the latest written report
↪  is submitted automatically; if you never wrote any report, the task is marked
↪  invalid.
- Tool calls that fail return a `[TOOL_ERROR]` message; {max_consecutive_errors}
↪  consecutive tool errors terminate the task.
```

C.2 Reasoning Synthesis

The synthesizer maintains a rolling conversation where each action request strictly precedes its observation reveal. Revealed observations include the actual images. During submission requests, we expose the reference report solely for the terminal audit, preventing it from influencing policy generation.

Synthesizer system role

```
You are reconstructing the reasoning of a radiologist AI agent from its action trajectory.
You will be shown the trajectory STEP BY STEP: at each turn I give you the ACTION the
↪  agent took
(tool name + arguments), and you must write the reasoning the agent plausibly had BEFORE
↪  taking that action.
After each action I reveal that step's OBSERVATION --- including the actual slice images
↪  when the action
was view_slice, so you can see everything the agent has seen up to that point.

Hard rules:
- The reasoning is written BEFORE the action's observation is revealed: it can only use
↪  information
  available up to that point (task setup, previous observations, and previously seen
  ↪  images).
- When earlier slice images are available, the reasoning MUST actively discuss the image
↪  findings
```

Synthesizer system role (continued)

```
  they showed --- which organs/regions were visible, what looked normal or suspicious,
  ↪  and how that
  motivates the current action --- as long as it stays logically consistent with what the
  ↪  agent knew
  at that point. Never mention images that have not been revealed yet.
- Keep the reasoning concise (2-6 sentences), in English.
- Also judge whether this action is COMPLETELY unreasonable given the whole previous
↪  context.

Output STRICT JSON: {"reasoning": "<text>", "action_plausible": <true/false>}
At the FINAL submit turn I will additionally give you the ground-truth report; then
↪  output:
{"reasoning": "<text>", "action_plausible": <true/false>, "traj_plausible": <true/false>}
where traj_plausible=false means the WHOLE trajectory is completely unreasonable.
Output ONLY the JSON object.
```

Action request template

```
ROUND {round} ACTION (tool call {tool_call_index}):
{action_json}

Write the reasoning the agent had BEFORE this action, and judge the action. Output JSON
↪  only.
```

Observation reveal template

```
ROUND {round} OBSERVATION (tool result, already executed):
{observation}

Acknowledge briefly (one sentence, e.g. "Noted.") and wait for the next action.
```

Submission audit template

```
ROUND {round} ACTION (tool call {tool_call_index}):
{action_json}

The submitted final report is:
<report>
{report}
</report>

GROUND-TRUTH report for this case:
<ground_truth>
{gt}
</ground_truth>

Write the reasoning before this submit, judge the action, and judge whether the WHOLE
↪  trajectory
is completely unreasonable (considering all previous steps and how the final report
↪  compares to the
ground truth). Output JSON with keys reasoning, action_plausible, traj_plausible. Output
↪  JSON only.
```

C.3 Report Evaluation and Fixed-Input Baseline

The abnormal-unit parser detailed in Section B.3 extracts explicit report findings. We populate its organ-label field using the configured taxonomy. We present the optional LLM matcher separately from the default algorithmic matching procedure.

Parser system role

```
You are a clinical information extraction engine for radiology reports. Output STRICT
↪  JSON only.
```

Parser user template

```
Task. Extract abnormality entities from a single ground-truth CT report and output them
↪  as a structured JSON object with high precision.

Input report. The ground-truth Findings or Impression text is provided below:
<report>
{report}
</report>

Step 1: abnormality definition. An abnormality entity is any medical concept explicitly
↪  described as abnormal, pathological, or non-normal, including diseases (e.g.,
↪  pneumonia, fatty liver), abnormal imaging findings (e.g., ground-glass opacity,
↪  nodule), and structural abnormalities (e.g., effusion, calcification, enlargement,
↪  stenosis).

Do not extract the following.
1. Explicitly negated findings, such as "no", "not seen", or "negative for".
2. Normal anatomical structures or normal imaging findings.
```

Parser user template (continued)

```
3. Content that only describes scanning conditions, image quality, or technical
↪  limitations.
4. Recommendations, follow-up suggestions, or examination plans without an explicit
↪  abnormality statement.

Step 2: extraction rules.
- Evidence-only extraction: extract only abnormalities explicitly stated in the report.
↪  Do not infer, assume, or hallucinate.
- Preserve uncertainty: if the report says "possible", "cannot exclude", or "consider",
↪  this uncertainty must be preserved rather than upgraded to a definite finding.
- Entity merging: if multiple sentences clearly refer to the same abnormality, merge them
↪  into one entity.
- Single-entity principle: if one statement describes a unified abnormality concept, such
↪  as "multiple small nodules in both lungs", treat it as one entity unless the report
↪  explicitly distinguishes separate lesions.
- Negation exclusion: any abnormality that is explicitly negated or ruled out must be
↪  omitted.

Step 3: required fields for each abnormality entity.
- name: the standardized abnormality name only, without anatomic location, size,
↪  severity, modifiers, or attributes. Example values include nodule, ground-glass
↪  opacity, and fatty liver.
- evidence: a verbatim text span copied directly from the report, concise but sufficient
↪  to support the abnormality.
- location: the explicitly stated anatomic location; if absent, use an empty string.
- attributes: imaging or pathological attributes such as size, number, extent, density,
↪  morphology, severity, or temporal change; if absent, use an empty string.
- certainty: must be one of definite or possible, based strictly on the report wording.
- organ: the normalized organ label, restricted to {organ_labels}.

Step 4: report-level field. In addition to the abnormality list, the output must include
↪  report_has_abnormality, which is true if at least one abnormality entity is extracted
↪  and false otherwise.

Final constraints. Do not extract diagnostic interpretations, speculative explanations,
↪  or parenthetical inferences as independent abnormalities. Repeated descriptions of
↪  the same lesions should remain merged. Only one valid JSON object may be returned,
↪  with no explanations or additional text.

Output format:
{{"abnormalities": [ {{"name": ..., "evidence": ..., "location": ..., "attributes": ...,
↪  "certainty": ..., "organ": ...}} ], "report_has_abnormality": true/false}}
Output ONLY the JSON object.
```

Matcher system role

```
You are a clinical radiology report evaluator. Output STRICT JSON only.
```

Matcher user template

```
Task. Compare a ground-truth CT report and a model-predicted CT report in terms of
↪  abnormality detection and description consistency, and output one JSON object only.

Inputs.
- Ground-truth abnormalities, already extracted as structured JSON:
{gt}
- Model-predicted Findings or Impression text:
<pred>
{pred}
</pred>

Step 1: abnormality extraction from the prediction.
- Use the provided ground-truth JSON directly as the reference abnormality set.
- Do not modify, merge, split, infer, or add any new ground-truth entities.
- Extract abnormality entities from the model-predicted report under the same definition
↪  rules.
- Exclude explicitly negated findings, normal findings, examination-condition
↪  descriptions, and interpretive or etiological impressions that are not tied to
↪  explicit imaging abnormalities.
- The name field of each extracted predicted entity must contain only the standardized
↪  abnormality name, without location, severity, modifiers, or attributes.

Step 2: abnormality matching. Match each ground-truth abnormality entity to at most one
↪  predicted abnormality entity based primarily on abnormality semantics. Medically
↪  equivalent or highly similar expressions can be matched, but no broadening,
↪  generalization, or inference beyond explicit textual evidence is allowed. If a
↪  matching abnormality is present, set hit = true; otherwise, set hit = false. Any
↪  predicted abnormality that matches no ground-truth entity must be listed as a false
↪  positive.

Step 3: consistency judgment rules.
- hit: true if a medically equivalent abnormality is explicitly present in the predicted
↪  report; false otherwise.
- location match: true if the predicted anatomic location is medically consistent with
↪  the ground-truth location, or if neither side explicitly contains location
↪  information. It must be false when hit = false.
- attribute match: true if the predicted imaging or pathological attributes are medically
↪  consistent with the ground-truth attributes, or if neither side explicitly contains
↪  attribute information. It must be false when hit = false.
- All judgments must remain evidence-based; missing locations or attributes must not be
↪  inferred.

Step 4: result aggregation.
- Report the matching and consistency results for every ground-truth abnormality entity,
↪  in the SAME ORDER as the input JSON.
- Report all unmatched false-positive abnormality entities found in the predicted report.
```

Matcher user template (continued)

```
Additional strict constraints. Do not infer, assume, hallucinate, or strengthen uncertain
↪  expressions. Do not add abnormalities based on clinical plausibility or typical
↪  associations. Only explicit, text-supported abnormalities may be matched. Negated or
↪  ruled-out findings must not be extracted.

Output format:
{{"abnormalities": [ {{"name": "<ground-truth abnormality name>", "hit": true/false,
↪  "location_match": true/false, "attribute_match": true/false}} ], "false_positive": [
↪  {{"name": "<false positive abnormality name>"}} ]}}
Output ONLY the JSON object.
```

The fixed-input baseline receives uniformly selected images before generating a report.

Baseline system role

```
You are an expert radiologist. Based on the provided slices uniformly sampled from the 3D
↪  imaging study, write a complete clinical radiology report directly.
```

Baseline user template

```
Below are {n_images} slice image(s) uniformly sampled from the following file(s):
{file_lines}

Each image is preceded by a label of the form [file_name][slice i/D].
Write the complete radiology report now (at most {max_report_chars} characters). Output
↪  the report text only.
```

C.4 ECI-RL Expert Contracts

We summarize the expert input boundaries defined in Section D. The report judge reads frozen claims alongside the final report. Check G processes the policy prefix and intended calls. Check I isolates blind visual-fact extraction from assertion matching. Check X assesses applicable differential claims. We compute check T programmatically from image availability timestamps and assertion origins.

Table 11: Joint ECI-RL evaluator contracts.

Evaluator	Inputs and outputs
Report	Frozen slots and final report only; quality, exact report quotes, error severity.
G	Actual policy prefix and intended call; acquisition relevance, pass/partial/fail.
I: visual facts	Actual images, provenance and neutral prefix question; localized observations.
I: support	Visual facts and credited assertion; support, insufficient, contradiction, certificate.
X	Applicable assertion, observed evidence and differential requirement; justified exclusion.
T (program)	Image availability, assertion origin and commitment; visibility and temporal factor.

Expert errors are distinct from policy failures. Report-mode labels are hidden from the report judge; all evaluator state is hidden from the policy.

D ECI-RL Mathematical Details

This section expands the joint acceptance and temporal credit in Section 2.4, and defines the interventions in Table 3.

Frozen Claims and Expert Acceptance. Each case has independently supportable core and detail claims. Let $\mathcal{R}_c$ be its nonempty tiers and Ω_r their slots. With $(\omega_{\text{core}}, \omega_{\text{detail}}) = (0.8, 0.2)$,

$$w_j = \frac{\omega_{r_j}}{|\Omega_{r_j}| \sum_{r \in \mathcal{R}_c} \omega_r}, \qquad F_0 = 6\Big[\sum_j w_j q_j - U\Big]_+. \tag{D.1}$$

Weights and X applicability are frozen before rollout and are never renormalized based on model success or the disabling of experts. Normal cases use task-relevant negative claims; an empty required ledger is treated as a data error. The report judge receives slots and the final report, without trajectory, expert verdict, mode, or checkpoint identity. Positive q_j requires a report quotation. The error map is $U \in \{0, 0.1, 0.25, 0.5\}$ for none/minor/major/severe, accompanied by a quoted error and justification. Absence from the reference alone does not by itself constitute a false claim. These are protocol settings, not fitted optima.

Expert G assesses acquisition relevance from the actual policy prefix; short reasoning is not by itself a failure. A multi-view certificate takes the minimum G factor across indispensable calls. Evaluator I first extracts localized visual facts without the reference or final report, then matches these facts to the credited assertion. Sufficient support, insufficient evidence, contradiction, and scoring error are distinct categories. Expert X checks only frozen applicable claims; otherwise its factor is one. Evaluator failures are excluded from policy-quality judgments.

Visibility, Origins, and Certificates. An image becomes available at the next generation block that actually receives it. Same-block tool outputs cannot justify an already generated write or submit action. Assertion origin is defined as its first occurrence with the same meaning, polarity, scope, and credited attributes; rewording or reinsertion does not refresh it. Correcting a false assertion creates a new assertion.

$$\chi_j = \mathbf{1}[\text{required images available before commitment}],$$
$$a_{\mathrm{T},j} = \begin{cases} 1, & \text{available before assertion origin,} \\ 0.85, & \text{verified after origin but before commitment,} \\ 0, & \chi_j = 0. \end{cases} \tag{D.2}$$

A certificate identifies necessary images, calls, supported content, and completion block. The first prefix certificate time t_j^0 uses only observations returned so far. The final certificate time t_j is the earliest time still valid under relevant later evidence and supporting the final assertion, such that $t_j^0 \le t_j$. Evidence overturned later may create a provisional event, whose credit is revoked at settlement.

Prefix Credit and Settlement. With accepted mass $v_j = w_j q_j D_j$,

$$\bar{V}_t = \sum_j w_j \mathbf{1}[t_j^0 \le t], \quad V_t^y = \sum_j v_j \mathbf{1}[t_j \le t], \quad \bar{r}_t = \lambda \kappa_t (\Delta \bar{V}_t - b_t/B). \tag{D.3}$$

Both ledgers are monotone, with $0 \le V_t^y \le \bar{V}_t \le 1$; each slot creates at most one positive increment per ledger. Provisional evaluation covers all frozen slots, not only those later reported. Visual facts must be extracted from actual observation prefixes, not from a judge shown future images.

There are exactly T real generation-block rewards. After termination, coverage is held fixed and $b_t = 0$ up to the common configured horizon H:

$$r_t = \bar{r}_t + \mathbf{1}[t = T](F + C), \quad C = \lambda J - \sum_{t=1}^{T} \bar{r}_t, \quad \sum_{t=1}^{T} r_t = R, \quad G_t = \sum_{u=t}^{T} r_u. \tag{D.4}$$

Every accepted event has no larger mass and occurs no earlier than its provisional event; hence $-\lambda \le C \le 0$. Negative $F + C$ represents legitimate cancellation and is not clipped to zero. Reward vectors follow generation blocks, not tool counts; automatic submission adds no policy token segment.

Bounds and Temporal Interpretation. Let $e_t = 1 - V_t^y$ and $Q_t = \sum_{u \le t} b_u$. For the linear kernel,

$$\sum_t \kappa_t \Delta V_t^y = H^{-1} \sum_t V_t^y, \qquad J = 1 - H^{-1} \sum_t e_t - (HB)^{-1} \sum_t Q_t. \tag{D.5}$$

The positive weighted increments and normalized violation cost each lie in $[0, 1]$:

$$-1 \le J \le 1, \quad -\lambda \le R \le 6 + \lambda, \quad F_a - F_b > 2\lambda \implies R_a > R_b. \tag{D.6}$$

This terminal margin refers to the joint F, not the report-only F_0. Holding D, F, H and all other events fixed, moving mass v_j from block s to $t < s$ gives

$$\Delta R = \lambda(s - t) v_j / H, \qquad r_{\mathrm{noop}} = -\lambda \kappa_t / B. \tag{D.7}$$

If the move changes temporal acceptance, both D and F must be recomputed. With fixed report and evidence, replacing idle padding with an uninformative penalized action cannot improve reward. Exploration without a new closure is not automatically a no-op.

Ablation Contracts. All modes share report-judge output and frozen weights. **Coarse report only** sets $D_j = 1$, $\lambda = 0$, and $R = F_0$, skipping experts and certificates. **Joint terminal only** sets $\lambda = 0$

while retaining the joint gate. **No G** skips the acquisition-relevance expert and sets its factor to one, not zero.

No temporal preference sets $a_{\mathrm{T}} = 1$, $\kappa_t = 1$, and $B = H$, retaining actual visibility χ; $|J| \leq 1$ still holds. **No imaging verification** replaces I certificates with weaker same-organ, actually visible procedural anchors. It sets $h = a_{\mathrm{I}} = 1$ and removes the I-dependent gate; without an anchor there is no positive temporal event and $a_{\mathrm{G}} = a_{\mathrm{T}} = 0$. Applicable X may read images independently. This changes the evidentiary basis and is not a clean independent-scalar removal.

Action-summary Denominators. For the behavioral comparisons in Figure 3, executed calls are distinct from generation blocks. With $n_{k,a}$ calls of action a in case k,

$$\mathrm{MeanCalls}_a = N^{-1} \sum_k n_{k,a}, \qquad \mathrm{Share}_a = \frac{\sum_k n_{k,a}}{\sum_{k,a'} n_{k,a'}}. \tag{D.8}$$

Pooled call shares differ from mean per-trajectory shares; figure-specific cohorts and denominators are given in Section A.3.

E Limitations and Future Work

Our study is limited to four volumetric imaging benchmarks and a shared four-action interface, which may not fully capture the diversity of real-world clinical workflows. The current environment focuses on retrospective evidence seeking within available scans and does not model additional information sources, such as prior examinations or clinical records. Moreover, evidence credit depends on automated expert judges and may inherit their evaluation biases. Future work will extend RAD-Gym to broader imaging modalities, institutions, and tool interfaces, while examining scalability across model capacities and interaction budgets and validating evidence attribution with expert assessment.

F Ethics Statement

RAD-Gym is a research environment for retrospective medical-image reporting. Its benchmark results do not establish safety for autonomous diagnosis or patient management. Use of the underlying datasets remains subject to their respective access conditions and governance requirements. Any clinical translation would require independent validation, appropriate oversight, and assessment of privacy and deployment risks.