EviAgent: Evidence-Driven Agent for Radiology Report Generation

Anonymous ACL submission

Abstract

Automated radiology report generation holds immense potential to alleviate the heavy workload of radiologists. Despite the formidable vision-language capabilities of recent Multimodal Large Language Models (MLLMs), their clinical deployment is severely constrained by inherent limitations: their "black-box" decision-making renders the generated reports untraceable due to the lack of explicit visual evidence to support the diagnosis, and they struggle to access external domain knowledge. To address these challenges, we propose the Evidence-driven Radiology Report Generation Agent (EviAgent). Unlike opaque end-to-end paradigms, EviAgent coordinates a transparent reasoning trajectory by breaking down the complex generation process into granular operational units. We integrate multi-dimensional visual experts and retrieval mechanisms as external support modules, endowing the system with explicit visual evidence and high-quality clinical priors. Extensive experiments on MIMIC-CXR, CheXpert Plus, and IU-Xray datasets demonstrate that EviAgent outperforms both large-scale generalist models and specialized medical models, providing a robust and trustworthy solution for automated radiology report generation.

1 Introduction

Radiology reports play a crucial role in the medical diagnosis and treatment process. However, interpreting radiology images is highly time-consuming and dependent on expert experience, which has catalyzed research into automated radiology report generation. Early approaches based on Encoder-Decoder architectures (Chen et al., 2020; Endo et al., 2021; Tanida et al., 2023) made initial strides but were fundamentally limited by their reliance on statistical image-text correlations rather than region-specific visual evidence. Consequently, despite achieving high natural language generation metrics, these methods often suffer from severe factual errors in their generated clinical descriptions (Tanida et al., 2023).

Recent advancements have been primarily propelled by Multimodal Large Language Models (MLLMs). Generalist models (OpenAI, 2025b; Google DeepMind, 2025; Anthropic, 2025b) have demonstrated formidable vision-language capabilities. Meanwhile, specialized open-source models like MedGemma (Sellergren et al., 2025) and Lingshu (Xu et al., 2025) have achieved performance comparable to generalist models with significantly fewer parameters (<10B) via high-quality biomedical instruction tuning. However, these end-to-end paradigms suffer from inherent limitations: First, they rely primarily on internal parametric reasoning, lacking the capability to acquire external knowledge; Second, their decision-making process is a "black box", rendering the generated reports untraceable due to the lack of explicit visual evidence to support the diagnosis. These issues severely constrain the performance of MLLMs in radiology report generation and hinder their potential for clinical deployment.

The emergence of agents in the medical domain (Kim et al., 2024; Fallahpour et al., 2025) has demonstrated that tool integration is an effective pathway to elevate model performance boundaries. However, most existing medical agents rely on close-source large models via cloud-based APIs as core planners. This data exfiltration paradigm inevitably raises severe concerns regarding medical data privacy and compliance.

To overcome these challenges, we propose the Evidence-driven Radiology Report Generation Agent (EviAgent) without additional training. We prioritize data privacy by building a fully local system where the core planner utilizes the open-source Qwen3-VL-8B (Bai et al., 2025a) model and all integrated tools run entirely on-premise. Furthermore, by utilizing vLLM (Kwon et al., 2023), a

high-throughput and memory-efficient inference engine, our framework ensures strict privacy compliance and faster inference speeds, making the system both secure and efficient for clinical deployment.

We equip the planner with specialized local experts. We integrate discriminative perception tools to address the visual black box by explicitly localizing pathologies, ensuring diagnostic claims are physically anchored in the image. Furthermore, we incorporate a retrieval mechanism that functions analogously to a clinician's accumulated experience, enabling the system to acquire external knowledge beyond its internal parameters.

Our main contributions are as follows.

- We construct a dynamic multi-expert collaboration framework, breaking down the complex report generation process into a sequence of granular operational units. This design establishes a traceable decision trajectory that explicitly maps diagnostic findings back to their supporting visual evidence, thereby significantly enhancing the reliability of the final reports.
- We integrate multi-dimensional visual expert models and knowledge bases as external support modules for evidence acquisition. This mechanism effectively overcomes the inherent limitation of MLLMs in retrieving external domain knowledge, endowing the system with explicit visual evidence and high-quality priors.
- We conduct extensive experiments on the MIMIC-CXR (Johnson et al., 2019), CheXpert Plus (Chambon et al., 2024), and IU-Xray (Demner-Fushman et al., 2015) datasets. The metrics demonstrate that our method outperforms both large-scale generalist models and specialized medical models, providing a robust and trustworthy solution for automated radiology report generation.

2 Related Work

Current advancements in automated radiology report generation are primarily categorized into two paradigms. The first involves Multimodal Large Language Models (MLLMs), which address radiological tasks via end-to-end foundation models. The second comprises medical agent systems, which enhance performance by either coordinating collaborative sub-agents or orchestrating specialized tools.

2.1 MLLMs in Radiology

Radiological image analysis encompasses a wide variety of clinical tasks, ranging from anatomical segmentation and lesion detection to Visual Question Answering (VQA) and report generation (Hosny et al., 2018; Najjar, 2023). Recently, MLLMs have demonstrated significant potential in addressing these diverse challenges. Early works such as LLaVA-Med (Li et al., 2023) and Med-Flamingo (Moor et al., 2023) explored aligning visual encoders with large language models to interpret radiological scans. More recently, advanced models like HuatuoGPT-V (Chen et al., 2024a), Lingshu (Xu et al., 2025), and MedGemma (Sellergren et al., 2025) have further enhanced comprehensive performance across these visual understanding tasks through training on large-scale multimodal data. In parallel, specialized models (Chen et al., 2024c; Bannur et al., 2024; Zambrano Chaves et al., 2025) have been developed specifically for report generation, which are explicitly trained on domain-specific image-report datasets to focus exclusively on this task. However, a critical limitation of these models is their inability to access external knowledge.

2.2 Medical Agents

Medical agent systems have emerged to surmount the rigidity of monolithic models by orchestrating specialized components. Contemporary research bifurcates into two streams. The first prioritizes collaborative reasoning, where frameworks like MedAgents (Wang et al., 2025) and MDAgents (Kim et al., 2024) simulate medical consultations through role-playing and iterative debates to mitigate bias. The second stream focuses on multimodal tool integration. Systems such as MMedAgent (Li et al., 2024) and MedAgent-Pro (Wang et al., 2025) employ central planners to dispatch specialized tools, addressing various medical tasks.

By leveraging collaboration or tool integration, these paradigms have pushed performance boundaries beyond MLLMs. However, they suffer from a persistent deficiency: a weak or absent connection between diagnostic conclusions and supporting visual evidence (Lou et al., 2025). Meanwhile, they are not specifically designed for radiology report generation. Consequently, an evidence-driven radiology report generation agent that grounds diagno-

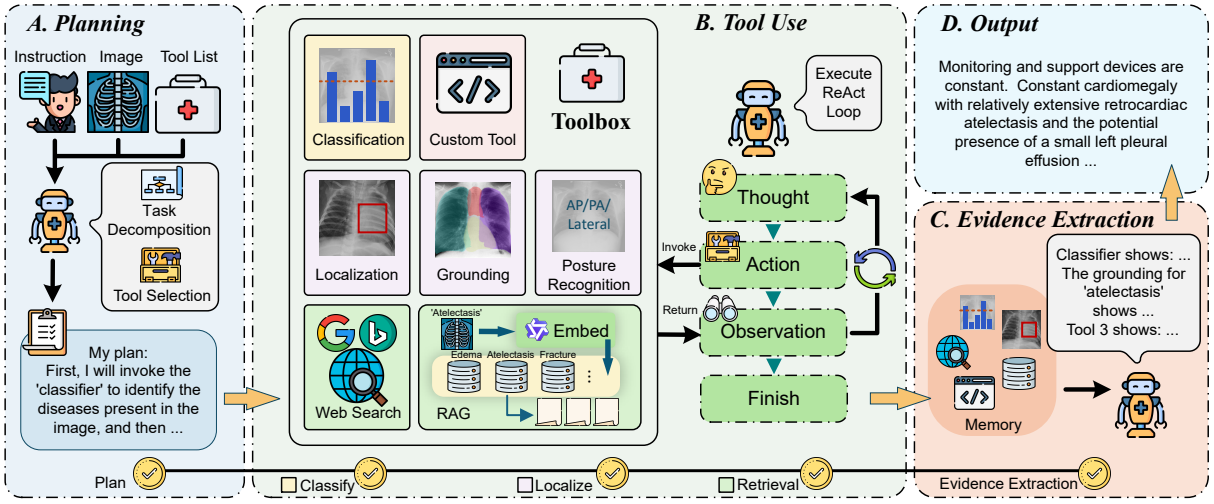

Figure 1: Overview of the EviAgent. The agent operates via four sequential stages: (A) *Planning*, which decomposes the task into granular operational units; (B) *Tool Use*, which dynamically invokes tools using the ReAct loop; (C) *Evidence Extraction*, which consolidates discrete observations into traceable proofs; and (D) *Output*, which synthesizes the final report strictly based on the accumulated evidence.

sis in explicit visual findings is urgently needed.

3 The EviAgent

We present EviAgent, an Evidence-driven Radiology Report Generation Agent. As illustrated in Figure 1, the agent coordinates the generation process through a structured agentic framework.

3.1 The Evidence-based Framework

The core of EviAgent is a progressive reasoning framework that transforms the "black-box" generation process into an evidence-based trajectory. The detailed inference process is formally described in Algorithm 1 (see Appendix A). Unlike end-to-end models that map pixels directly to text, our system coordinates the diagnosis through a structured "Plan-Act-Report" paradigm, where each diagnostic conclusion is rigorously derived from accumulated evidence.

Planning and Decomposition. Upon receiving a radiology image I and a diagnostic instruction Q, the Planner $\mathcal{P}_{LLM}$ initiates the workflow by analyzing the global context. Instead of attempting to generate the report immediately, the model formulates a structured execution plan. The complex diagnostic goal is decomposed into a sequence of granular operational units (e.g., “Detect lesions in the image” and “Localize diseases output by the classifier”). This explicit planning phase serves as a roadmap, ensuring that subsequent tool invocations are goal-oriented rather than random explorations.

Tool-Augmented ReAct Loop. Guided by the initial execution plan, the agent enters a ReAct (Yao et al., 2022) loop. It is important to note that the plan generated in *Planning* stage is inherently coarse-grained. It merely defines the high-level diagnostic scope but implies no knowledge of concrete pathological findings until the tools are actually executed. Therefore, this stage functions as a process of dynamic refinement. At each time step t, the Planner generates a thought trace based on the current state and dynamically selects the appropriate tool from the toolbox $\mathcal{T}$ to substantiate the preliminary plan with fine-grained evidence.

During execution, the agent parses the thought trace to determine the next action and its arguments. The selected tool is then executed, returning a structured result—such as a specific classification result or bounding box coordinates—which is immediately appended to the Evidence Memory $\mathcal{M}$. This continuous feedback loop empowers the agent with self-correction and investigative depth. For instance, if the initial screening tool detects atelectasis, the agent does not merely stop; instead, it spontaneously adapts its trajectory to trigger follow-up actions like “localize the atelectasis” thereby mimicking the iterative investigative process of a human radiologist.

Evidence Extraction. To ensure the integrity of the final report, the agent performs extraction, analysis, and integration of the accumulated tool outputs stored in Memory $\mathcal{M}$. This stage specifically targets valid execution results, including classification probabilities, localized visual features, and

retrieved expert knowledge, extracting them from the verbose interaction history. By filtering out procedural artifacts and consolidating these findings, the system constructs a structured evidence chain $\mathcal{E}$. This chain functions as a purified set of clinical facts for the subsequent generation.

Output. Finally, the agent generates the final radiology report R by strictly conditioning on the extracted evidence $\mathcal{E}$. By grounding the generation in these verifiable facts rather than purely parametric knowledge, EviAgent ensures that the output is not only linguistically coherent but also clinically accurate and evidence-based.

3.2 Experts & Knowledge Integration

To support the tool-augmented ReAct loop, we construct a comprehensive toolbox tailored for radiological reasoning. By integrating a multi-dimensional set of specialized experts, we empower the Planner with professional visual perception and external domain knowledge. These experts are categorized into three distinct levels: Perception, Knowledge, and Customization.

3.2.1 Discriminative Perception Experts

This module functions as the visual perception unit, providing precise visual interpretation capabilities ranging from global screening to pixel-level grounding.

Classification tool. For initial disease screening, we employ a Swin-Transformer (Bu et al., 2024) trained on MIMIC-CXR (Johnson et al., 2019) as the core classifier. This model takes the raw chest X-ray image as input and performs multi-label classification to output a specific list of detected pathologies (e.g., "Pneumonia", "Cardiomegaly"). These classification results serve as the pivotal diagnostic anchors; the planner subsequently centers its entire reasoning trajectory, including localization and knowledge retrieval, specifically around these identified diseases.

Localization tools. To enable spatial localization and detailed anatomical analysis, we integrate three specialized models for fine-grained perception:

- Posture Recognition: Correctly identifying the projection view is a prerequisite for diagnosis. We employ CheXagent (Chen et al., 2024c) for posture recognition.
- Visual Grounding: We utilize MAIRA-2 (Bannur et al., 2024) to handle grounding tasks. Given a specific disease query, it outputs precise bounding boxes (bbox), allowing the agent to spatially verify diagnoses.
- Anatomical Segmentation: For pixel-level precision, we incorporate MedSAM (Ma et al., 2024) to segment anatomical structures and lesion boundaries.

3.2.2 Retrieval-Augmented Knowledge

Inspired by the clinical practice where radiologists rely on accumulated medical knowledge and experience to compose reports, we design this module to supplement the agent's generation process with external domain expertise. By retrieving historical cases relevant to the current visual findings, it provides the system with essential clinical priors.

To construct the external repository, we derive data from the MIMIC-CXR training set, partitioning it into $N = 14$ pathology-specific knowledge bases corresponding to CheXpert (Irvin et al., 2019) standards. For each pathology c, we explicitly select $M = 50$ high-quality triplet samples (I_j, R_j, L_j), comprising the reference image, clinical report, and disease label respectively. We utilize the GME-Qwen2-VL (Zhang et al., 2024b) model, denoted as $E_{mm}(\cdot)$, to compute multimodal embeddings. The offline embedding vector $\mathbf{v}_{c,j}$ is calculated as:

$$\mathbf{v}_{c,j} = E_{mm}(I_j \oplus L_j) \tag{1}$$

where $\oplus$ denotes concatenation. Consequently, each knowledge base $\mathcal{B}_c$ stores the paired mapping of embedding vectors and raw reports: $\{(\mathbf{v}_{c,j}, R_j)\}_{j=1}^{M}$.

During online inference, the Planner analyzes the context and specifies a set of suspected pathology labels $\mathcal{C}_{query}$. Based on these labels, the RAG module routes the request to the corresponding knowledge bases. It generates queries for each target pathology and aggregates all candidates. The query generation, similarity scoring $s_{c,j}$, and final global retrieval $\mathcal{C}_{ret}$ are formulated as:

$$\mathbf{q}_c = E_{mm}(I_{query} \oplus c), \quad \forall c \in \mathcal{C}_{query} \tag{2}$$

$$s_{c,j} = \frac{\mathbf{q}_c \cdot \mathbf{v}_{c,j}}{\|\mathbf{q}_c\|\|\mathbf{v}_{c,j}\|} \tag{3}$$

$$\mathcal{C}_{ret} = \bigcup_{c \in \mathcal{C}_{query}} \{R_j \mid \text{rank}(s_{c,j}) \leq k\} \tag{4}$$

where $\text{rank}(\cdot)$ returns the descending rank of the similarity score among the candidate pool, ensuring only the top-k most relevant reports are selected.

Additionally, we integrate a web search tool (via Bing or Google APIs). Serving as a dynamic knowledge extension, this module is designed to mitigate the inherent knowledge limitations of the foundation model, empowering the Planner to verify ambiguous concepts or acquire real-time medical information beyond the scope of its pre-trained parametric memory.

3.2.3 Extensible Customization via Configuration

We adopt the Model Context Protocol (MCP) as the unified interface layer, which provides a "Plug-and-Play" extension mechanism. Users can integrate local proprietary tools by simply modifying a JSON configuration file. By defining the tool's description, API endpoint, and argument schema, the Planner can automatically recognize and invoke these custom tools without any model fine-tuning. This design ensures that EviAgent can seamlessly integrate into diverse hospital information ecosystems, fostering a flexible and scalable deployment.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our framework on three public benchmarks. MIMIC-CXR (Johnson et al., 2019), a widely-used dataset developed by the Beth Israel Deaconess Medical Center. We utilize the official split, where the test set comprises 14 disease labels, 2,347 reports and 3,858 images. IU-Xray (Demner-Fushman et al., 2015), a comprehensive chest X-ray dataset released by Indiana University. Following the established 7:1:2 partition from previous works (Wang et al., 2023; Liu et al., 2021), the test set includes 590 reports and 1,180 images. Additionally, we incorporate CheXpert Plus (Chambon et al., 2024), a dataset containing 234 reports and 234 images in its official test split. For all three datasets, only the official test partitions are used in our experiments.

Metrics. Following prior studies (Bannur et al., 2024; Zhang et al., 2024a; Xu et al., 2025), we employ three advanced domain-specific metrics. RaTEScore (Zhao et al., 2024) is an entity-aware metric specifically designed to emphasize critical medical entities, including diagnostic findings and anatomical details. It demonstrates robustness to complex medical synonyms and high sensitivity to negation expressions. SembScore (Smit et al., 2020) assesses clinical content alignment by calculating the cosine similarity between vectors of 14 pathological indicators, which are automatically extracted by the CheXbert labeler from both generated and ground-truth reports. Finally, we report RadCliQ^{-1}, the inverse of RadCliQ-v1 (Yu et al., 2023). This composite metric integrates BLEU for lexical precision, BERTScore for semantic consistency, and RadGraph-F1 for graph-based clinical relation matching, thereby providing a holistic assessment of report quality. We utilize the inverse form to ensure that higher scores consistently indicate better performance across all evaluations.

Baselines. To strictly evaluate the performance of our proposed framework, we compare EviAgent against a comprehensive set of state-of-the-art baselines categorized into three distinct groups:

(1) Close-Source Generalist MLLMs, representing the general reasoning capabilities, including GPT-5.1 (OpenAI, 2025b), Claude 4.5 Sonnet (Anthropic, 2025b), and Gemini-2.5-Flash (Google DeepMind, 2025).

(2) Medical Specialized MLLMs, which are explicitly optimized for the medical domain via continued pre-training or instruction tuning. Representative models include LLaVA-Med-7B (Li et al., 2023), HuatuoGPT-V-7B (Chen et al., 2024a), BiMediX2-8B (Mullappilly et al., 2024), MedGemma-4B-IT (Sellergren et al., 2025), and Lingshu-7B (Xu et al., 2025).

(3) Open-Source Generalist MLLMs, serving as foundation baselines, including the InternVL series (Chen et al., 2024b; Zhu et al., 2025) and Qwen-VL series (Bai et al., 2025b,a).

Settings. We employ Qwen3-VL-8B-Instruct (Qwen3-VL-8B) (Bai et al., 2025a) as the unified backbone, sequentially functioning as the planner, tool user, evidence extractor, and report generator. For the *Tool Use* stage, the maximum number of interaction rounds is set to $T_{max} = 10$ to prevent infinite loops while allowing sufficient exploration. In the RAG module, we retrieve the Top-$k = 4$ reference reports. All experiments are run on the NVIDIA H100 GPU. The model is deployed via the vLLM library to accelerate inference.

4.2 Performance on Multiple Metrics

Table 1 demonstrates that EviAgent achieves the best performance across three datasets on almost all metrics. On the large-scale MIMIC-CXR dataset, our method establishes a significant lead, surpassing close-source giants like GPT-5.1 by 3.1 in RaTE and outperforming specialized medical

Model	MIMIC-CXR			CheXpert Plus			IU-Xray		
	RaTE	Semb	RadCliQ^{-1}	RaTE	Semb	RadCliQ^{-1}	RaTE	Semb	RadCliQ^{-1}
GPT-5.1	49.5	28.0	65.6	<u>48.0</u>	27.5	**49.2**	56.8	<u>51.5</u>	85.7
Claude 4.5 Sonnet	49.1	23.9	64.7	47.4	22.5	48.1	<u>57.8</u>	51.4	90.9
Gemini-2.5-Flash	50.3	29.7	59.4	44.3	27.4	44.0	55.6	50.9	91.6
LLaVA-Med-7B	12.8	18.3	52.9	38.8	23.5	44.0	40.9	16.0	58.1
HuatuoGPT-V-7B	48.9	20.0	48.2	44.2	19.3	39.4	52.9	40.7	63.6
BiMediX2-8B	44.4	17.7	53.0	40.8	21.6	43.3	40.1	11.6	53.8
MedGemma-4B-IT	<u>52.4</u>	29.2	62.9	47.2	<u>29.3</u>	46.6	57.0	46.8	86.7
Lingshu-7B	52.1	<u>30.0</u>	<u>69.2</u>	45.4	26.8	47.3	57.6	48.4	<u>108.1</u>
InternVL2.5-8B	47.0	21.0	56.2	43.1	19.7	42.7	51.1	36.7	67.0
InternVL3-8B	48.2	21.5	55.1	44.3	25.2	43.7	51.2	31.3	59.9
Qwen2.5-VL-7B	47.0	18.4	55.1	41.0	17.2	43.1	48.4	36.3	66.1
Qwen3-VL-8B	48.9	26.1	64.2	45.9	27.3	44.6	50.3	46.5	74.7
EviAgent (Ours)	**52.6**	**43.6**	**76.6**	**49.8**	**30.4**	<u>48.8</u>	**60.5**	**52.2**	**110.2**

Table 1: Results on automatic metrics. All scores are scaled by a factor of 100 to enhance clarity and comprehension. The best results are highlighted in **bold**, and the second-best results are <u>underlined</u>. The data for medical MLLMs in this table are taken from the Lingshu paper (Xu et al., 2025).

MLLMs such as Lingshu-7B by 13.6 in Semb. Furthermore, on the IU-Xray and CheXpert Plus benchmarks, EviAgent consistently exceeds the baselines, achieving a RadCliQ^{-1} of 110.2 on IU-Xray. This observation indicates that the proposed framework achieves better performance compared to standard MLLMs.

The efficacy of our framework is most evident when comparing EviAgent with its backbone, Qwen3-VL-8B. Across all three datasets, equipping the base model with our agentic workflow yields a substantial performance leap. For instance, Semb scores on MIMIC-CXR nearly double from 26.1 to 43.6, and RaTE on IU-Xray improves by 10.2 points. GPT-5.1 shows slight linguistic advantages on the RadCliQ^{-1} metric for CheXpert Plus due to extensive general pre-training. These results validate the effectiveness of our proposed method, demonstrating that the collaborative synergy of multiple tools within the agentic architecture enables the system to significantly outperform the traditional standalone inference baseline.

4.3 Clinical Value Evaluation via LLM-as-a-Judge

To further assess the granular medical details and holistic clinical validity required in real-world diagnosis, we conducted a comprehensive evaluation on the entire test set of each benchmark using DeepSeek-V3.2 (DeepSeek-AI et al., 2025) as an impartial judge. The judge scored the generated reports on a scale of 0 to 10 across four specific dimensions: Accuracy, which measures the correctness of findings; Localization, evaluating the precision of anatomical descriptions; Professionalism, assessing the use of standardized radiological terminology; and Clinical Admissibility, a holistic metric indicating whether the report is robust enough for clinical workflows without major modification. As shown in Table 2, EviAgent achieves the best performance across almost all metrics. While Gemini-2.5-Flash exhibits a marginal lead in Professionalism for MIMIC-CXR due to the inherent linguistic advantages of large-scale models, our framework consistently dominates in all other critical diagnostic dimensions.

4.4 Ablation Study

To validate individual module contributions, we evaluated variants by removing the **Classification**, **Localization**, and **Retrieval** components (denoted as "w/o"). Table 3 summarizes the results.

Removing the classification expert causes the most severe degradation. Notably, the Semb score on MIMIC-CXR drops to 23.1, even lower than the vanilla backbone. We attribute this to error propagation: without reliable triage, the planner's initial hallucinations misguide subsequent tool invocations, causing the reasoning trajectory to deviate progressively further than standard end-to-end generation.

The absence of localization tools leads to a decline in RaTE across all datasets, confirming that visual grounding is essential for verifying fine-

Model	MIMIC-CXR				CheXpert Plus				IU-Xray			
	Acc	Loc	Prof	Adm	Acc	Loc	Prof	Adm	Acc	Loc	Prof	Adm
GPT-5.1	4.91	4.98	8.67	5.62	4.85	4.23	8.17	5.55	7.42	5.51	8.81	7.08
Claude 4.5 Sonnet	3.51	3.79	8.59	4.40	3.41	3.14	8.30	4.33	7.12	5.60	9.15	7.20
Gemini-2.5-Flash	5.74	6.19	**9.06**	6.48	4.28	4.70	7.88	5.05	6.67	5.73	8.61	6.70
LLaVA-Med-7B	1.74	2.88	4.28	2.23	1.71	2.85	3.86	2.06	2.07	3.37	3.86	2.01
HuatuoGPT-V-7B	2.20	5.07	7.33	3.17	1.94	4.27	5.65	2.84	1.21	5.50	5.95	2.29
BiMediX2-8B	1.41	2.76	3.84	1.86	1.22	2.35	3.69	1.68	0.51	2.77	4.27	1.14
MedGemma-4B-IT	5.44	5.61	8.16	5.97	4.16	3.82	7.68	4.80	7.24	6.28	8.38	7.41
Lingshu-7B	5.88	5.87	8.66	6.37	4.60	3.91	7.75	5.20	7.39	6.24	8.88	7.67
InternVL2.5-8B	2.41	3.44	7.33	3.41	2.55	2.86	7.06	3.52	6.81	5.87	8.00	6.58
InternVL3-8B	3.07	4.71	7.55	4.05	2.82	3.88	7.29	3.85	3.68	6.36	7.72	4.48
Qwen2.5-VL-7B	2.30	3.49	7.30	3.43	2.21	2.93	7.13	3.33	5.79	6.04	7.86	5.91
Qwen3-VL-8B	3.94	4.66	7.51	4.75	3.72	3.78	7.48	4.58	5.74	4.98	7.63	5.64
EviAgent (Ours)	**6.04**	**6.32**	8.70	**6.61**	**4.91**	**6.66**	**8.45**	**5.72**	**7.48**	**6.55**	**9.29**	**7.72**

Table 2: Clinical value evaluation via LLM-as-a-Judge. We evaluate the generated reports across four dimensions: Acc (Accuracy), Loc (Localization), Prof (Professionalism), and Adm (Clinical Admissibility).

Dataset	Metric	Full	w/o (Remove)			Base
			Cls.	Loc.	Ret.	
MIMIC-CXR	RaTE	**52.6**	50.6	51.8	51.7	48.9
	Semb	**43.6**	23.1	43.0	43.4	26.1
	RadCliQ^{-1}	**76.6**	65.1	75.3	76.2	64.2
CheXpert Plus	RaTE	**49.8**	47.7	49.2	48.5	45.9
	Semb	**30.4**	27.5	29.9	28.4	27.3
	RadCliQ^{-1}	**48.8**	47.4	48.2	47.9	44.6
IU-Xray	RaTE	**60.5**	54.7	60.1	59.0	50.3
	Semb	**52.2**	43.1	51.5	50.4	46.5
	RadCliQ^{-1}	**110.2**	83.9	104.6	91.8	74.7

Table 3: Ablation study of module contributions. We report the performance drop when removing specific components. Full represents the complete EviAgent framework, while Base denotes the vanilla Qwen3-VL-8B performing end-to-end generation. Cls.: Classification, Loc.: Localization, Ret.: Retrieval.

grained anatomical entities. Similarly, removing the retrieval tools consistently lowers RadCliQ^{-1}, indicating that external knowledge is vital for maintaining professional stylistic standards and minimizing linguistic hallucinations.

4.5 Qualitative Analysis

To qualitatively validate the efficacy and interpretability of our framework, we present a comparative analysis of two cases in Figure 2.

Precision in Complex Clinical Scenarios. The top panel illustrates a complex ICU scenario. The ground truth confirms extensive bilateral opacities, bilateral pleural effusions, cardiomegaly, and basal atelectasis. In this case, the generalist model GPT-5.1 exhibits significant failures: it misses the diagnosis of atelectasis and incorrectly characterizes the effusions as small. Furthermore, it fails to provide a definitive conclusion, stating that effusions or consolidation cannot be excluded. In contrast, EviAgent successfully identifies all key pathologies. The tools correctly capture the bilateral nature of the effusions and opacities, as well as the presence of cardiomegaly and atelectasis. This structured evidence guides the planner to generate a precise report that accurately reflects the extent and spatial distribution of the disease.

Traceability of Diagnostic Discrepancies. The bottom panel demonstrates the error traceability of our architecture. In this case, the ground truth describes bibasilar opacities and mild congestion. However, our agent reports a left-sided opacity and pulmonary edema. Crucially, this discrepancy is not a generative fabrication. The evidence log explicitly shows that the grounding tool restricted the lung opacity to the left field and the classifier outputted edema. The planner faithfully aggregated these specific tool outputs into the final report. This confirms that the errors—mislocalization and clinical over-diagnosis—stem directly from the perception module's limitations rather than the reasoning

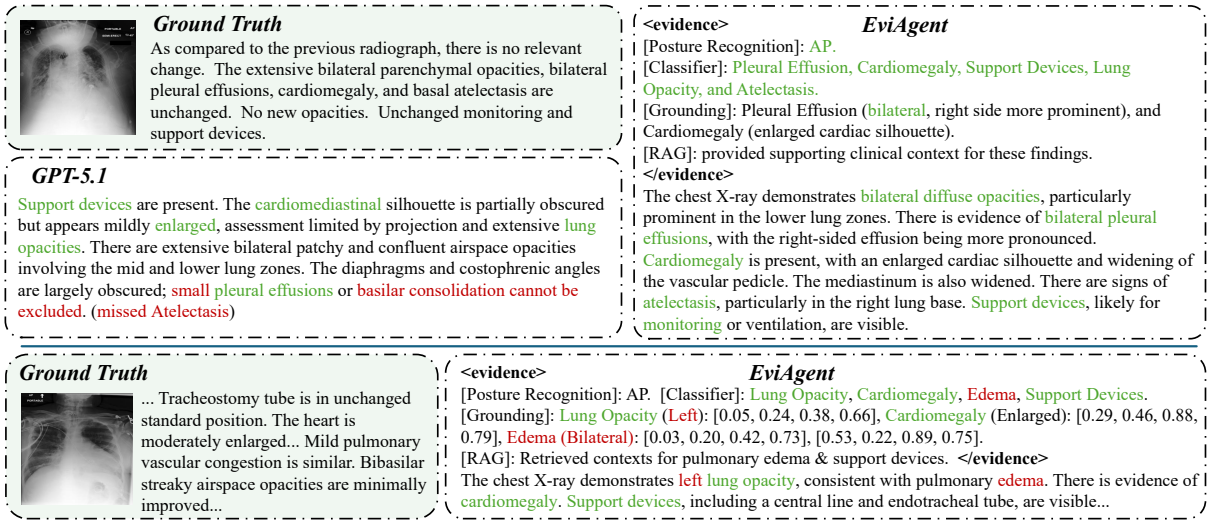

Figure 2: Qualitative analysis. Top: A complex case (Study 56122911) where EviAgent achieves accurate diagnosis and localization. Bottom: A case demonstrating error traceability (Study 50239281). Text in green indicates consistency with the ground truth, while text in red denotes discrepancies.

Backbone Planner	VR	Tool Call	FER	QS
GPT-5.1	98.9%	4.40	4.2%	6.95
Claude 4.5 Sonnet	98.6%	4.48	4.5%	6.93
InternVL3-8B	97.2%	5.13	10.3%	6.72
Qwen3-VL-8B	98.2%	4.70	8.8%	6.89

Table 4: Scalability analysis of different backbone planners on the mixed evaluation subset.

engine. Crucially, this validates the integrity of our evidence-driven paradigm: the planner faithfully executes based on the provided visual signals rather than hallucinating content. Consequently, the system maintains full traceability, allowing diagnostic discrepancies to be logically isolated to specific visual experts.

4.6 Agentic Capability and Extensibility Analysis

To evaluate the extensibility of our framework and the impact of different backbone planners, we conducted a test using a mixed evaluation subset of 1,000 instances: 500 cases from MIMIC-CXR to test report generation capabilities and 500 questions from MIMIC-CXR-VQA (Bae et al., 2023).

We employ four agent-specific metrics: (1) Valid Rate (VR): The probability of generating a valid final response within the maximum limit of tool invocations T_{max}; (2) Tool Call: The average number of tool executions per episode; (3) Format Error Rate (FER): The frequency of parsing failures or hallucinating tools; and (4) Quality Score (QS): The generation quality ranging from 0 to 10 by DeepSeek-V3.2 based on the ground truth.

As shown in Table 4, although Qwen3-VL-8B incurs higher average tool calls than GPT-5.1 due to retry loops triggered by formatting errors, its final Quality Score is only marginally lower (6.89 vs. 6.95). Our experiment consistently demonstrate the stability and robustness of the proposed agent framework regardless of the underlying model choice.

5 Conclusion

In this work, we presented EviAgent, an evidence-driven agent designed to fundamentally address the twin challenges of "black-box" opacity and the inability to access external knowledge in existing MLLMs. By shifting from static parametric reasoning to active evidence acquisition, our system ensures that every diagnostic conclusion is rigorously grounded in explicit visual findings and retrieved domain knowledge. Extensive experiments demonstrate that EviAgent achieves superior performance in clinical accuracy. Notably, our approach significantly outperforms open-source models of similar scale and even surpasses top-tier proprietary giants like GPT-5.1, validating the efficacy of our evidence-driven agentic framework in medical domains.

Limitations

While EviAgent demonstrates superior clinical accuracy and interpretability, it entails a trade-off

regarding inference latency. Unlike monolithic end-to-end MLLMs that generate reports in a single forward pass, our framework relies on a multi-round evidence-driven agentic workflow. The frequent invocation of external tools and the iterative reasoning process inevitably introduce additional computational overhead, resulting in longer generation times compared to direct generation approaches.

Future optimizations in tool execution and infrastructure are expected to mitigate this latency, bridging the efficiency gap without compromising clinical rigor.

Ethical Considerations

This research is conducted using the MIMIC-CXR (Johnson et al., 2019), MIMIC-CXR-VQA (Bae et al., 2023), IU-Xray (Demner-Fushman et al., 2015), and CheXpert Plus (Chambon et al., 2024) datasets, all of which are publicly accessible and have undergone automatic de-identification to mitigate privacy risks.

While our framework ensures that the generated findings are explicitly traceable to specific tool outputs, we emphasize that such traceability does not guarantee absolute accuracy. The integrated tools or the reasoning process may still yield erroneous evidence or omissions. Therefore, these outputs must not be used as a substitute for expert medical judgment. We strongly advocate for mandatory validation by qualified radiologists or healthcare professionals before any clinical or diagnostic application.

References

Anthropic. 2025a. Introducing Claude 4. Accessed: 2025-12-27.

Anthropic. 2025b. Introducing Claude Sonnet 4.5. Accessed: 2025-12-27.

Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei Ji, Eric I-Chao Chang, Tackeun Kim, and Edward Choi. 2023. Ehrxqa: A multi-modal question answering dataset for electronic health records with chest x-ray images. *Preprint*, arXiv:2310.18652.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025b. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, Mercy Ranjit, Shaury Srivastav, Julia Gong, Noel C. F. Codella, Fabian Falck, Ozan Oktay, Matthew P. Lungren, Maria Teodora Wetscherek, and 2 others. 2024. Maira-2: Grounded radiology report generation. *Preprint*, arXiv:2406.04449.

Shenshen Bu, Taiji Li, Yuedong Yang, and Zhiming Dai. 2024. Instance-level expert knowledge and aggregate discriminative attention for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14194–14204.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P Langlotz. 2024. Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. *arXiv preprint arXiv:2405.19538*.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, and 1 others. 2024a. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, and 1 others. 2024c. Chexagent: Towards a foundation model for chest x-ray interpretation. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 245 others. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *Preprint*, arXiv:2512.02556.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2015. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Mark Endo, Rayan Krishnan, Viswesh Krishna, Andrew Y Ng, and Pranav Rajpurkar. 2021. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In *Machine Learning for Health*, pages 209–219. PMLR.

Adibvafa Fallahpour, Jun Ma, Alif Munim, Hongwei Lyu, and Bo Wang. 2025. Medrax: Medical reasoning agent for chest x-ray. *arXiv preprint arXiv:2502.02673*.

Google DeepMind. 2025. Gemini 2.5: Our most intelligent ai model. Accessed: 2025-12-27.

Sunan He, Yuxiang Nie, Hongmei Wang, Shu Yang, Yihui Wang, Zhiyuan Cai, Zhixuan Chen, Yingxue Xu, Luyang Luo, Huiling Xiang, Xi Lin, Mingxiang Wu, Yifan Peng, George Shih, Ziyang Xu, Xian Wu, Qiong Wang, Ronald Cheong Kin Chan, Varut Vardhanabhuti, and 5 others. 2024. Gsco: Towards generalizable ai in medicine via generalist-specialist collaboration. *Preprint*, arXiv:2404.15127.

Ahmed Hosny, Chintan Parmar, John Quackenbush, Lawrence H Schwartz, and Hugo JWL Aerts. 2018. Artificial intelligence in radiology. *Nature Reviews Cancer*, 18(8):500–510.

Zixuan Huang, Yikun Ban, Lean Fu, Xiaojie Li, Zhongxiang Dai, Jianxin Li, and Deqing Wang. 2025. Adaptive batch-wise sample scheduling for direct preference optimization. *Preprint*, arXiv:2506.17252.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, and 1 others. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik S Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae W Park. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems*, 37:79410–79452.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626.

Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, Yuheng Li, Konstantinos Psounis, and Xiaofeng Yang. 2025. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *arXiv preprint arXiv:2503.13939*.

Binxu Li, Tiankai Yan, Yuanting Pan, Jie Luo, Ruiyang Ji, Jiayuan Ding, Zhe Xu, Shilong Liu, Haoyu Dong, Zihao Lin, and 1 others. 2024. Mmedagent: Learning to use medical tools with multi-modal agent. *arXiv preprint arXiv:2407.02483*.

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

Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wanggui He, Hao Jiang, Mengze Li, Xiaohui Song, and 1 others. 2025. Healthgpt: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. *arXiv preprint arXiv:2502.09838*.

Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Ping Zhang, and Xu Sun. 2021. Contrastive attention for automatic chest x-ray report generation. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*, pages 269–280.

Jinhui Lou, Yan Yang, Zhou Yu, Zhenqi Fu, Weidong Han, Qingming Huang, and Jun Yu. 2025. Cxragent: Director-orchestrated multi-stage reasoning for chest x-ray interpretation. *arXiv preprint arXiv:2510.21324*.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. 2024. Segment anything in medical images. *Nature Communications*, 15(1):654.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR.

Sahal Shaji Mullappilly, Mohammed Irfan Kurpath, Sara Pieri, Saeed Yahya Alseiari, Shanavas Cholakkal, Khaled Aldahmani, Fahad Khan, Rao Anwer, Salman Khan, Timothy Baldwin, and 1 others. 2024. Bimedix2: Bio-medical expert lmm for diverse medical modalities. *arXiv preprint arXiv:2412.07769*.

Reabal Najjar. 2023. Redefining radiology: a review of artificial intelligence integration in medical imaging. *Diagnostics*, 13(17):2760.

OpenAI. 2025a. Introducing GPT-4.1 in the API. Accessed: 2025-12-27.

OpenAI. 2025b. Introducing GPT-5.1. Accessed: 2025-12-27.

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. 2025. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 337–347. Springer.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. 2020. Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*.

Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. 2023. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7433–7442.

5 Team, Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, Lucen Zhong, Mingdao Liu, Rui Lu, Shulin Cao, Xiaohan Zhang, Xuancheng Huang, Yao Wei, and 152 others. 2025a. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *Preprint*, arXiv:2508.06471.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, Zhuofu Chen, Jialei Cui, Hao Ding, Mengnan Dong, Angang Du, Chenzhuang Du, Dikang Du, Yulun Du, Yu Fan, and 150 others. 2025b. Kimi k2: Open agentic intelligence. *Preprint*, arXiv:2507.20534.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2023. Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11558–11567.

Ziyue Wang, Junde Wu, Linghan Cai, Chang Han Low, Xihong Yang, Qiaxuan Li, and Yueming Jin. 2025. Medagent-pro: Towards evidence-based multi-modal medical diagnosis via reasoning agentic workflow. *arXiv preprint arXiv:2503.18968*.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, and 1 others. 2025. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y Ng, and 1 others. 2023. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9).

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, Hany Awadalla, Julia Gong, Houdong Hu, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Yu Gu, Cliff Wong, Mu Wei, and 8 others. 2025. A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings. *Nature Communications*, 16(1):3108.

Xiaoman Zhang, Hong-Yu Zhou, Xiaoli Yang, Oishi Banerjee, Julián N Acosta, Josh Miller, Ouwen Huang, and Pranav Rajpurkar. 2024a. Rexrank: A public leaderboard for ai-powered radiology report generation. *arXiv preprint arXiv:2411.15122*.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2024b. Gme: Improving universal multimodal retrieval by multimodal llms. *arXiv preprint arXiv:2412.16855*.

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Ratescore: A metric for radiology report generation. *arXiv preprint arXiv:2406.16845*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Implementation Details of EviAgent

In this section, we provide granular details regarding the inference process, prompt engineering, tool configuration, and knowledge base construction to ensure the reproducibility of our agent.

A.1 Formal Inference Process

To provide a precise mathematical description of the "Plan-Act-Report" paradigm discussed in Section 3.1, Algorithm 1 outlines the complete inference loop. It explicates how the Planner $\mathcal{P}_{LLM}$

decomposes the user instruction Q into an initial $Plan$, iteratively updates the Evidence Memory $\mathcal{M}$ through dynamic tool invocation ($Acts_t$), and finally synthesizes the report R based on the structured evidence $\mathcal{E}$.

Algorithm 1 Process of EviAgent

Input: Q: User Instruction
I: Input Image
$\mathcal{T}$: Toolbox $\{\mathcal{T}_{cls}, \mathcal{T}_{loc}, \mathcal{T}_{ret}, \mathcal{T}_{custom}\}$
T_{max}: Maximum allowed rounds
$\mathcal{P}_{LLM}$: Planner Agent
Output: R: Final Radiology Report
Initialize:
$Plan \leftarrow \mathcal{P}_{LLM}(I, Q, \text{description}(\mathcal{T}))$
$\mathcal{M} \leftarrow \emptyset, t \leftarrow 0$
while Unfinished($Plan$) **and** $t < T_{max}$ **do**
$Thought_t \leftarrow \mathcal{P}_{LLM}(I, \mathcal{M}, Plan)$
$Actions_t, Args_t \leftarrow \text{Parse}(Thought_t)$
if $Acts_t \in \mathcal{T}$ **then**
$Observation_t \leftarrow InvokeTools(Acts_t, Args_t, \mathcal{T})$
$\mathcal{M} \leftarrow \mathcal{M} \cup \{Observation_t\}$
else
break
end if
$t \leftarrow t + 1$
end while
$\mathcal{E} \leftarrow \text{EvidenceExtraction}(\mathcal{M})$
$R \leftarrow \mathcal{P}_{LLM}(I, \mathcal{E})$
return R

A.2 Detailed Prompt Designs

To ensure reproducibility, we provide the full system prompts used in the different stages of EviAgent.

System Prompt for the Planner Agent

You are a professional radiologist AI assistant. Your goal is to write a comprehensive Chest X-Ray report.
No need to fill in the image path when you use tool calls; the system will automatically load the image for you.
Follow this strict workflow:
1. Diagnostic Process
Start by listing a numbered plan:

[Step 1]
[Step 2]
[Step 3]

2. Tool Usage
Execute the diagnostic plan by calling tools in this specific order:
(1) Classification Phase:
- Call the "classify" tool to detect potential pathologies.
- List all positive findings with their confidence scores.
(2) Localization Phase:
- View Identification: Call "chexagent" with the prompt "What is the projection of this chest X-ray (AP, PA, or Lateral)?" to distinguish different views.
- IMPORTANT: Use this information ONLY for internal spatial mapping logic. DO NOT include the view projection (AP/PA/Lateral) in the final report Findings or Impression.
- Spatial Mapping:
- For each positive finding from Phase 1, use the "grounding" tool to locate it.
- Use the "segmentation" tool to outline major anatomical structures (lungs, heart, clavicles) to confirm anatomical relationships.
(3) Retrieval Phase:
-Knowledge Retrieval: You MUST ALWAYS call the "rag_retriever" tool.
-WARNING: The returned content is provided ONLY as a reference for medical terminology and reporting style. DO NOT use them as the final source of information.
-CRITICAL: You MUST use the exact positive findings (the "diseases" or "rag_diseases" list) identified by the "classify" tool as the search queries.
-Do not use generic terms or hallucinate keywords. If the classifier returns "No Finding", search for "No Finding".
-External Validation: Use "web_search" ONLY if you have uncertain knowledge.

3. Evidence Extraction
Extract clinical evidence from tool outputs.

4. Final Output Generation
Generate the final response based on the user's request type:
(A) For Report Generation (Default):
- If the user asks for a "Report" or "Chest X-Ray Report", follow the strict diagnostic workflow above.
- Output valid JSON with these keys:

{
"findings": "Detailed anatomical observations.",
"impression": "Brief summary string. DO NOT mention AP/PA view.",
"evidence": "Detailed evidence. MUST include the VQA view confirmation and coordinate logic."
}

(B) For General Questions (VQA/Other):
- If the user asks a specific question (e.g., "Is there pneumonia?", "Where is the mass?"), you are FREE to choose relevant tools (you do not need to follow the strict report workflow).
- If you have answered the question with the tool, DO NOT call the tool again with the same prompt.
- Output valid JSON with these keys:

{
"output": "Direct answer to the user's question.",
"evidence": "Detailed evidence. MUST include the VQA view confirmation and coordinate logic."
}

IMPORTANT:
- Your response must be in valid JSON format.
- DO NOT wrap the JSON in markdown code blocks. Output RAW JSON only.
- The "classify" tool is the ULTIMATE AUTHORITY on disease diagnosis.

A.3 Knowledge Base Construction Details

To support the retrieval of diverse clinical contexts, we constructed $N = 14$ pathology-specific knowledge bases. The taxonomy aligns strictly with the standard observations defined by the CheXpert labeler (Irvin et al., 2019). The specific categories include: Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Cardiomediastinum, Fracture, Lung Lesion, Lung Opacity, No Finding, Pleural Effusion, Pleural Other, Pneumonia, Pneumothorax, and Support Devices.

Data Selection Pipeline. The reference samples were exclusively curated from the MIMIC-CXR training split to prevent data leakage. For each of the 14 categories, we executed a rigorous selection process to curate $M = 50$ high-quality exemplars:

- **Label Filtering.** We first filtered cases where the specific CheXpert label was strictly positive (1.0), excluding uncertain (0.0) or negative (-1.0) cases to ensure diagnostic clarity.
- **Quality Control.** To ensure the retrieved text provides sufficient stylistic guidance, we filtered out reports with extremely short descriptions.
- **Sampling.** From the filtered pool, we randomly sampled 50 image-report pairs to form the final knowledge base $\mathcal{B}_c$ for each pathology.

For each entry, the image and its corresponding disease text label are jointly encoded using the GME-Qwen2-VL (Zhang et al., 2024b) model and indexed into a localized vector database. During the retrieval phase, the system identifies the Top-k most similar cases based on cosine similarity. Crucially, only the textual reports of these retrieved cases are returned to the planner to serve as stylistic references.

A.4 MCP Tool Configuration

The Model Context Protocol (MCP) serves as a standardized interface layer that decouples the AI planner from specific tool implementations (Huang et al., 2025). By abstracting the communication details, MCP allows for a "Plug-and-Play" architecture where new capabilities can be added without modifying the agent's core codebase.

To illustrate this, we present the actual configuration file used to integrate an external Web Search tool below. Users are only required to define the server type and endpoint URL. Crucially, by pointing to a localhost endpoint, the system ensures that all data interactions remain strictly within the on-premise environment, adhering to privacy standards. The Planner then automatically negotiates with the endpoint to discover the tool's schema and arguments, significantly lowering the barrier for deployment.

user_config.json


```
{
  "mcpServers": {
    "WebSearch": {
      "type": "streamable_http",
      "url": "http://localhost:8080/mcp"
    }
```

```
    }
}
```

A.5 vLLM Settings

To ensure reproducibility, we provide the specific command used to deploy the Qwen3-VL-8B-Instruct model using the vLLM framework. The server was initialized with the following hyperparameters, specifically tuned for tool-use capabilities:

```
python -m vllm.entrypoints.openai.
    api_server \
  --model Qwen/Qwen3-VL-8B-Instruct \
  --trust-remote-code \
  --gpu_memory_utilization 0.87 \
  --max-model-len 19000 \
  --port 8000 \
  --enable-auto-tool-choice \
  --tool-call-parser hermes \
  --served-model-name qwen3-vl-8b \
  --kv-cache-dtype auto
```

B Supplementary Experiments

In this section, we expand our experimental evaluation by including additional baseline models and providing the exact protocols used for the qualitative evaluation.

B.1 Extended Comparison with Large-Scale Models

To provide a holistic view of the current landscape in automated radiology report generation, we extend our evaluation to include a wider array of state-of-the-art models. The performance data for these additional baselines are referenced from the Lingshu benchmark (Xu et al., 2025). We categorize these models into three distinct groups:

- **Closed-Source Generalist MLLMs:** Including high-performing proprietary models such as GPT-4.1 (OpenAI, 2025a) and Claude 4 Sonnet (Anthropic, 2025a).
- **Open-Source Medical MLLMs:** Covering recent domain-specific models like MedR1-2B (Lai et al., 2025), MedVLM-R1-2B (Pan et al., 2025), HealthGPT-14B (Lin et al., 2025), and larger-scale variants like HuatuoGPT-V-34B (Chen et al., 2024a) and MedDr-40B (He et al., 2024).
- **Open-Source Generalist MLLMs:** Incorporating powerful general-purpose vision-language models with varying scales, including InternVL2.5-38B (Chen et al., 2024b), InternVL3-14B/38B (Zhu et al., 2025), and Qwen2.5-VL-32B (Bai et al., 2025b).

Analysis. As presented in Table 5, EviAgent demonstrates remarkable clinical efficacy. Notably, despite utilizing a significantly smaller parameter backbone (8B), our approach consistently outperforms models with much larger parameter scales. For instance, EviAgent surpasses the 40B-parameter MedDr-40B and the 38B-parameter InternVL3-38B across critical clinical metrics (e.g., RadCliQ and Semb). This superior performance highlights a pivotal insight: in high-stakes medical domains, a rigorous, evidence-driven planning mechanism is more critical for diagnostic precision than simply scaling up raw model parameters.

B.2 Prompt Template for Automated Evaluation

To ensure a rigorous and fair assessment, we design a structured prompt that positions the evaluator as a "senior expert radiologist." As shown in the Prompt Template below, we adopt a comparative list-wise evaluation strategy. Instead of scoring each model in isolation, the Ground Truth (GT) report is input simultaneously with a batch of candidate reports (represented by the placeholder {models_text}). This side-by-side presentation enables the judge to capture subtle nuances and strictly differentiate between "Good" and "Great" models based on four critical dimensions: Disease Identification, Lesion Localization, Professionalism, and Total Utility.

LLM-as-Judge Prompt Template

You are a senior expert radiologist and medical AI evaluator.
Your task is to evaluate and compare multiple AI-generated radiology reports against a Ground Truth (GT) report written by a human radiologist.

Ground Truth Report:
{gt_report}

Candidate Reports:
{models_text}

Evaluation Instructions:
Evaluate each candidate report based on the

Model	MIMIC-CXR			CheXpert Plus			IU-Xray		
	RaTE	Semb	RadCliQ^{-1}	RaTE	Semb	RadCliQ^{-1}	RaTE	Semb	RadCliQ^{-1}
Close-Source Generalist MLLMs									
GPT-4.1	51.3	23.9	57.1	45.5	23.2	45.5	51.3	47.5	80.3
GPT-5.1	49.5	28.0	65.6	<u>48.0</u>	27.5	**49.2**	56.8	<u>51.5</u>	85.7
Claude 4 Sonnet	45.6	19.7	53.4	43.5	18.9	43.3	55.4	41.0	72.1
Claude 4.5 Sonnet	49.1	23.9	64.7	47.4	22.5	48.1	<u>57.8</u>	51.4	90.9
Gemini-2.5-Flash	50.3	29.7	59.4	44.3	27.4	44.0	55.6	50.9	91.6
Medical MLLMs									
Med-R1-2B	40.6	14.8	42.4	38.5	17.8	37.6	41.4	12.5	43.6
MedVLM-R1-2B	41.6	14.2	48.3	38.9	15.5	40.9	46.1	22.7	54.3
MedGemma-4B-IT	<u>52.4</u>	29.2	62.9	47.2	<u>29.3</u>	46.6	57.0	46.8	86.7
LLaVA-Med-7B	12.8	18.3	52.9	38.8	23.5	44.0	40.9	16.0	58.1
HealthGPT-14B	48.4	16.5	52.7	44.4	22.7	42.6	50.8	16.6	56.9
HuatuoGPT-V-7B	48.9	20.0	48.2	44.2	19.3	39.4	52.9	40.7	63.6
HuatuoGPT-V-34B	48.5	23.0	47.1	42.9	22.1	39.7	54.4	42.2	59.3
BiMediX2-8B	44.4	17.7	53.0	40.8	21.6	43.3	40.1	11.6	53.8
MedDr-40B	45.2	12.2	47.0	44.7	24.2	44.7	40.3	7.3	48.9
Lingshu-7B	52.1	<u>30.0</u>	<u>69.2</u>	45.4	26.8	47.3	57.6	48.4	<u>108.1</u>
Open-Source Generalist MLLMs									
InternVL2.5-8B	47.0	21.0	56.2	43.1	19.7	42.7	51.1	36.7	67.0
InternVL2.5-38B	47.5	18.2	54.9	42.6	20.3	45.4	53.5	38.5	69.7
InternVL3-8B	48.2	21.5	55.1	44.3	25.2	43.7	51.2	31.3	59.9
InternVL3-14B	48.6	17.4	46.5	44.1	20.7	39.4	55.0	38.7	55.0
InternVL3-38B	47.9	18.1	47.2	43.8	20.2	39.4	53.5	33.1	55.2
Qwen2.5-VL-7B	47.0	18.4	55.1	41.0	17.2	43.1	48.4	36.3	66.1
Qwen2.5-VL-32B	47.5	17.1	45.2	43.4	18.5	40.3	51.3	38.1	54.0
Qwen3-VL-8B	48.9	26.1	64.2	45.9	27.3	44.6	50.3	46.5	74.7
EviAgent (Ours)	**52.6**	**43.6**	**76.6**	**49.8**	**30.4**	<u>48.8</u>	**60.5**	**52.2**	**110.2**

Table 5: Comprehensive Evaluation Results. We categorize baselines into Close-Source Generalist MLLMs, Medical MLLMs, and Open-Source Generalist MLLMs. **Bold** denotes the best result, and <u>underline</u> denotes the second best.

following 4 dimensions.
Your goal is to distinguish between "Good" models and "Great" models.
Scores should be widely distributed. Do not cluster everything at 10 or 0.

1. Disease Identification Accuracy
Focus: Precision and Completeness.
- 10 (Perfect): Captures ALL findings, including subtle ones (e.g., "trace effusion", "tiny nodule") and modifiers (severity, chronicity).
- 8-9 (Excellent): Correctly identifies the main pathology but misses a minor detail or modifier.
- 6-7 (Good): Identifies the general diagnosis but lacks specificity (e.g., "opacity" instead of "consolidation").
- 4-5 (Fair): Misses a secondary finding or is slightly ambiguous.
- 1-3 (Poor): Misses the Primary Diagnosis.
- 0 (Failure): Misses all findings or states normal when pathology exists.

2. Lesion Localization Accuracy
Focus: Anatomical Precision.
- 10 (Perfect): Exact match (e.g., "Right Upper Lobe Apical Segment").
- 8-9 (Strong): Correct side and general region (e.g., "Right Upper Lobe").
- 6-7 (Acceptable): Correct side but vague (e.g., "Right lung").
- 0-5 (Incorrect): Wrong Side, Wrong Lobe, or Wrong Organ.

3. Professionalism & Fluency
Focus: Radiologist-level Style.

- 10 (Professional): Indistinguishable from the GT. Concise, RadLex terms, professional tone.
- 8-9 (Fluent): Good medical language but slightly verbose or less polished than GT.
- 6-7 (Readable): Understandable but uses non-standard phrasing.
- 0-5 (Poor): Repetitive, robotic, or non-medical style.

4. Total Score (Weighted Utility)
Focus: Overall ranking.
- 10 (Exceptional): Reserved for models that are FLAWLESS and professionally written.
- 8-9 (High Quality): Accurate and safe, but maybe not perfect in nuance.
- 6-7 (Usable): Needs minor edits by a doctor.
- 0-5 (Unsatisfactory): Significant errors or useless.

Scoring Guidelines:
1. Differentiate: If Model A is more precise than Model B (e.g., "Left Lower Lobe" vs "Left Lung"), Model A MUST get a higher score (e.g., 10 vs 7).
2. Implied Negatives (The "Smart" Rule): If GT says "No pneumothorax" and Model is silent, assume Model agrees (Normal). Do NOT penalize.
3. Be Strict on Positives: If GT reports a finding, the Model MUST report it accurately to get >8.
4. No Inflation: Do not give 10 easily. 10 means "I would sign this report without editing."

Output Format:
Return a strictly valid JSON object.

```
{
    "evaluations": {
        "model_name": {
            "disease_accuracy_score": <0-10>,
            "localization_score": <0-10>,
            "professionalism_score": <0-10>,
            "total_score": <0-10>,
            "reasoning": "<Compare this model to others. Why is it better/worse?>"
        },
        ...
    }
}
```

B.3 Robustness Analysis via Diverse LLM Judges

To mitigate potential biases inherent in a single LLM evaluator and ensure the objectivity of our results, we extended our evaluation framework to include two additional high-performance models as impartial judges: GLM-4.5 (Team et al., 2025a) and Kimi-K2-Instruct (Team et al., 2025b). We applied the same rigorous scoring criteria across the four dimensions—Accuracy, Localization, Professionalism, and Admissibility—on the full test sets of MIMIC-CXR, CheXpert Plus, and IU-Xray.

The results, presented in Table 6 and Table 7, demonstrate a high degree of consistency with our main findings.

Consistency in Clinical Precision. Under the evaluation of GLM-4.5, EviAgent achieves state-of-the-art performance in Accuracy and Localization across all three datasets. Notably, on the IU-Xray benchmark, our method surpasses the second-best model (GPT-5.1) by a significant margin in Admissibility (7.16 vs. 6.81). Similarly, the evaluation via Kimi-K2-Instruct further corroborates this trend, where EviAgent secures the top rank in Accuracy and Localization on both MIMIC-CXR and CheXpert Plus. This cross-model consensus confirms that our evidence-driven framework effectively anchors generation in visual facts, yielding diagnostic reports that are universally recognized as accurate by diverse generalist reasoners.

Resilience in Stylistic Alignment. While large-scale commercial models like Claude 4.5 Sonnet exhibit strong linguistic capabilities, often leading in the Professionalism metric, EviAgent remains highly competitive. In the Kimi-K2-Instruct evaluation, our model even achieved the highest Professionalism score on MIMIC-CXR (8.63) and IU-Xray (9.35). This indicates that by retrieving high-quality reference cases, EviAgent can match or exceed the stylistic fluency of much larger models while maintaining superior diagnostic fidelity.

In summary, the consistent superiority of EviAgent across different judge models validates that our performance gains are robust and not artifacts of a specific evaluator's preference.

Model	MIMIC-CXR				CheXpert Plus				IU-Xray			
	Acc	Loc	Prof	Adm	Acc	Loc	Prof	Adm	Acc	Loc	Prof	Adm
GPT-5.1	4.12	4.62	8.72	4.52	4.12	4.18	9.01	4.44	6.94	7.21	8.85	6.81
Claude 4.5 Sonnet	2.87	3.32	8.91	3.33	3.26	3.47	**9.36**	3.61	6.91	7.17	8.92	6.89
Gemini-2.5-Flash	4.38	5.08	8.68	4.62	3.89	4.47	8.28	4.16	6.02	6.76	8.66	6.02
LLaVA-Med-7B	1.24	1.57	4.24	1.26	1.31	1.57	4.12	1.15	1.89	2.36	4.27	1.44
HuatuoGPT-V-7B	1.78	3.61	5.66	2.08	1.57	2.97	5.19	1.76	0.85	1.75	5.68	1.16
BiMediX2-8B	1.35	1.58	5.75	1.64	1.21	1.41	5.82	1.51	0.38	0.45	5.97	0.67
MedGemma-4B-IT	4.01	4.77	8.11	4.26	3.25	3.32	7.30	3.26	6.74	7.49	8.21	6.39
Lingshu-7B	4.42	5.12	8.02	4.64	3.38	3.59	7.00	3.41	7.25	7.69	8.40	6.91
InternVL2.5-8B	1.64	2.42	6.92	1.89	2.25	2.69	6.77	2.29	6.59	7.01	7.72	5.94
InternVL3-8B	2.20	3.05	6.87	2.44	2.42	3.13	6.74	2.53	3.21	4.08	7.18	3.09
Qwen2.5-VL-7B	1.40	2.20	7.03	1.62	1.68	2.08	6.97	1.77	5.50	6.14	7.56	5.03
Qwen3-VL-8B	2.71	3.26	7.04	2.95	2.82	3.12	7.25	2.97	5.55	6.03	7.15	4.99
EviAgent (Ours)	**4.62**	**5.22**	**8.96**	**5.15**	**4.35**	**4.62**	8.95	**4.75**	**7.58**	**7.82**	**9.05**	**7.16**

Table 6: Clinical value evaluation via LLM-as-a-Judge using GLM-4.5. We evaluate the generated reports across four dimensions: Acc (Accuracy), Loc (Localization), Prof (Professionalism), and Adm (Clinical Admissibility).

B.4 Detailed Case Study

Figure 3 illustrates the complete, step-by-step inference trajectory for Study 56122911, corresponding to the analysis presented in the Case Study section of the main text. Due to space constraints, we provide the full visualization of the inference process here.

Model	MIMIC-CXR				CheXpert Plus				IU-Xray			
	Acc	Loc	Prof	Adm	Acc	Loc	Prof	Adm	Acc	Loc	Prof	Adm
GPT-5.1	4.16	4.50	8.27	4.82	4.20	3.90	8.28	4.82	6.81	6.41	8.27	6.73
Claude 4.5 Sonnet	3.21	3.38	8.50	3.93	3.51	3.48	**8.78**	4.33	6.64	6.17	8.98	6.83
Gemini-2.5-Flash	4.40	5.16	8.59	5.12	4.12	4.39	8.23	4.79	6.52	6.38	8.62	6.70
LLaVA-Med-7B	1.16	1.55	4.58	1.67	1.33	1.72	4.40	1.73	1.90	2.31	5.07	2.00
HuatuoGPT-V-7B	1.91	4.45	6.13	2.81	1.72	3.91	5.70	2.50	0.99	4.35	6.39	2.01
BiMediX2-8B	1.52	2.05	5.42	2.27	1.67	2.21	5.21	2.38	0.44	0.84	5.84	1.42
MedGemma-4B-IT	4.20	4.57	7.86	4.85	3.49	3.21	7.56	4.01	7.04	6.06	8.50	7.29
Lingshu-7B	5.00	5.03	8.08	5.45	3.90	3.57	7.33	4.33	7.02	6.98	8.62	7.40
InternVL2.5-8B	1.66	2.65	7.49	2.53	2.30	2.53	7.35	3.03	6.68	6.62	8.35	6.63
InternVL3-8B	2.28	3.68	7.44	3.10	2.45	3.34	7.42	3.17	3.51	5.69	7.97	4.08
Qwen2.5-VL-7B	1.54	2.47	7.07	2.38	1.96	2.60	7.20	2.71	5.62	6.29	8.07	5.71
Qwen3-VL-8B	2.63	3.34	7.28	3.32	2.81	2.98	7.50	3.44	5.55	5.25	7.64	5.46
EviAgent (Ours)	**5.02**	**5.27**	**8.63**	**5.71**	**4.28**	**4.88**	8.22	**4.96**	**7.09**	**7.06**	**9.35**	**7.53**

Table 7: Clinical value evaluation via LLM-as-a-Judge using Kimi-K2-Instruct. We evaluate the generated reports across four dimensions: Acc (Accuracy), Loc (Localization), Prof (Professionalism), and Adm (Clinical Admissibility).

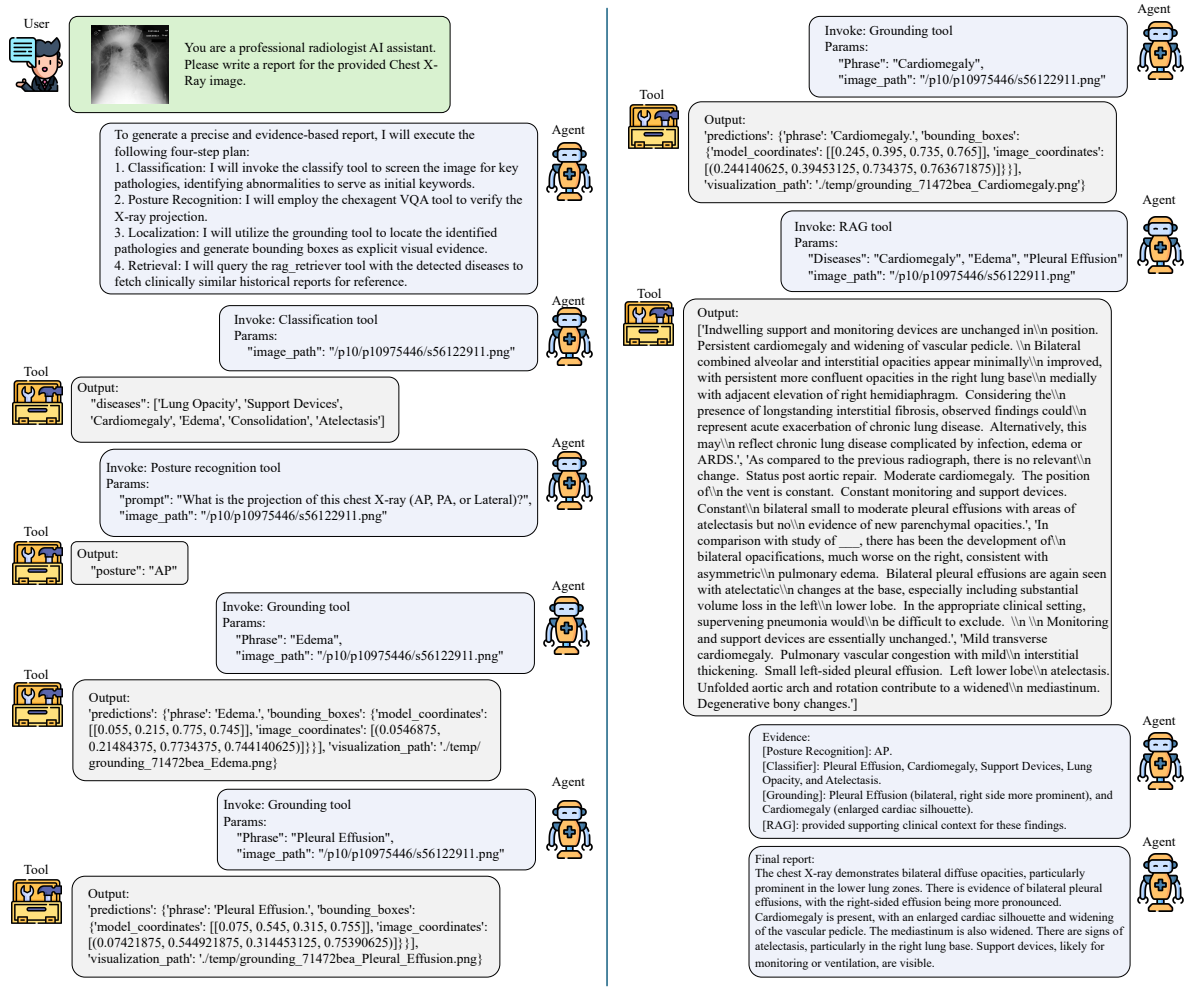

Figure 3: Detailed case study. A step-by-step visualization of how EviAgent processes a sample input.