Aligning Findings with Diagnosis: A Self-Consistent Reinforcement Learning Framework for Trustworthy Radiology Reporting

Anonymous ACL submission

Abstract

Multimodal Large Language Models (MLLMs) have shown strong potential for Radiology Report Generation, yet their clinical translation is hindered by architectural heterogeneity and the prevalence of factual hallucinations. Standard supervised fine-tuning often fails to strictly align linguistic outputs with visual evidence, while existing Reinforcement Learning approaches struggle with either prohibitive computational costs or limited exploration. To address these challenges, we propose a comprehensive framework for self-consistent radiology report generation. First, we conduct a systematic evaluation to identify optimal vision encoder and LLM backbone configurations for medical imaging. Building on this foundation, we introduce a novel *"Reason-then-Summarize"* architecture optimized via Group Relative Policy Optimization (GRPO). This framework restructures generation into two distinct components: a `<think>` block for detailed findings and an `<answer>` block for structured disease labels. By utilizing a multi-dimensional composite reward function, we explicitly penalize logical discrepancies between the generated narrative and the final diagnosis. Extensive experiments on the MIMIC-CXR benchmark demonstrate that our method achieves State-of-the-Art performance in clinical efficacy metrics and significantly reduces hallucinations compared to strong supervised baselines.

1 Introduction

Recent years have witnessed remarkable progress in Multimodal Large Language Models (MLLMs) (Alayrac et al., 2022; Bai et al., 2025; Li et al., 2023; Liu et al., 2023; Wang et al., 2025b). By integrating a large language model (LLM) backbone, a vision encoder, and a multimodal connector, MLLMs enable joint reasoning over visual and textual inputs and have achieved strong performance across a range of vision–language tasks, including image captioning (Rotstein et al., 2024; Chen et al., 2024c), visual question answering (VQA) (Xu et al., 2024; Kuang et al., 2025), and medical image understanding (Liu et al., 2025a; Wang and Zhang, 2024; Zhao et al., 2024, 2025; Tang et al., 2022; Gu et al., 2025). Among these applications, Radiology Report Generation (RRG) has emerged as a particularly challenging and important testbed for MLLMs, as it requires accurate visual grounding and internal consistency between textual findings and diagnostic conclusions (Yi et al., 2025). RRG aims to automatically generate free-text clinical findings and diagnostic impressions from radiological images (e.g., chest X-rays), with the potential to substantially reduce radiologists' workload while maintaining diagnostic quality.

Despite substantial advances driven by MLLMs (Chen et al., 2024c; Hein et al., 2025; Sellergren et al., 2025a; Zambrano Chaves et al., 2025), the clinical translation of automated RRG remains impeded by two fundamental barriers: architectural design choices and the reliability of generation. First, the optimal integration between vision encoders and LLM backbones for radiological reasoning remains insufficiently characterized. While general-domain studies have underscored the importance of strong visual representations (Tong et al., 2024), systematic investigations of encoder-LLM pairings in medical imaging contexts are still lacking. As a result, existing MLLM-based RRG systems rely on heuristic architectural configurations, potentially limiting their diagnostic acuity and robustness.

Second, even with carefully selected architectures, standard Supervised Fine-Tuning (SFT) frequently leads to factual hallucination. Under this paradigm, model outputs are often dominated by linguistic priors rather than image-grounded evidence, resulting in clinically plausible yet incorrect assertions, such as fabricating pleural effusions (Liu et al., 2025b; Chen et al., 2024a; Wang et al., 2025a). To align generation with clinical

ground truth, recent work has explored Reinforcement Learning (RL) as an alternative optimization paradigm. Unlike SFT, which is tightly coupled to surface-level phrasing, RL enables direct optimization of clinically motivated efficacy metrics, thereby emphasizing image-grounded evidence and essential diagnostic content rather than exact lexical matching.

Selecting an efficient reinforcement learning (RL) algorithm remains challenging. Proximal Policy Optimization (PPO) enables direct optimization of clinical metrics (Zhou et al., 2024; Yang et al., 2025a) but incurs substantial computational overhead due to its reliance on a large critic network. In contrast, Direct Preference Optimization (DPO) is more efficient but limited by its offline nature, restricting active exploration and error correction (Liu et al., 2025b; Hein et al., 2025). Group Relative Policy Optimization (GRPO) (Shao et al., 2024) provides a promising compromise by supporting online exploration without a critic network. However, adapting GRPO to the medical domain remains non-trivial: prior applications exhibit logical inconsistencies between intermediate reasoning and final conclusions (Lai et al., 2025; Pan et al., 2025), and radiology reports do not naturally conform to the explicit "thought–answer" structure assumed by standard reasoning frameworks.

To address these challenges, we propose a comprehensive study and a novel framework for the RRG task. We first conduct a systematic investigation to identify the optimal combination of vision encoder and LLM backbone for RRG. Building upon the best-performing configuration, we introduce a novel RL framework for self-consistent RRG. Unlike traditional approaches that treat report generation and disease classification as separate tasks, our framework integrates them into a rigorous *"Reason-then-Summarize"* workflow that reflects the clinical reporting process. Specifically, the model is guided to structure outputs in two distinct components: a `<think><\think>` block that generates detailed free-text clinical findings, followed by an `<answer><\answer>` block that summarizes these findings into structured, machine-readable JSON disease labels.

To leverage this structure, we design a multi-dimensional composite reward function optimized via RL. Beyond surface-level lexical overlap, this function explicitly enforces internal consistency and clinical accuracy through: (1) a consistency reward ($R_{\text{consistency}}$) that penalizes discrepancies between generated findings in `<think>` and the corresponding structured diagnostic outputs in `<answer>`; (2) diagnostic accuracy rewards (R_{accuracy}) that align both free-text reports and extracted labels with ground-truth clinical annotations; and (3) format & semantic rewards that guarantee adherence to the structured output protocol and semantic fidelity to reference reports. Optimizing this holistic objective enables the model to generate reports that are not only linguistically fluent but also logically consistent and clinically verifiable. Our main contributions are summarized as follows:

- We systematically study the impact of various vision encoder and LLM pairings on RRG performance, showing that domain-adapted vision encoders combined with reasoning-capable LLMs substantially enhance report quality.
- We propose a *"Reason-then-Summarize"* architecture that redefines RRG as a structured two-stage generation task, enabling explicit alignment between textual evidence and structured diagnostic outputs.
- We introduce a self-consistent RL Framework with a composite reward function that explicitly optimizes the internal consistency of the model's reasoning chain and clinical correctness, significantly reducing hallucinations.
- Extensive experiments on the MIMIC-CXR (Johnson et al., 2024) benchmark demonstrate that our method achieves State-of-the-Art (SOTA) performance on clinical efficacy metrics (e.g., CheXbert F1) compared to strong supervised baselines.

2 Related Work

2.1 Medical MLLMs

The proliferation of large-scale MLLMs has catalyzed the development of domain-specific healthcare adaptations, with systems such as LLaVA-Med (Li et al., 2023) and HuatuoGPT-Vision (Chen et al., 2024b) demonstrating remarkable proficiency in radiology report generation and diagnostic tasks. Despite these advances, the predominant paradigm for adapting these architectures to the medical domain remains SFT on final-answer labels (Zhang et al., 2023; Xu et al., 2024; Liu et al., 2025c). This paradigm relies heavily on massive, high-quality image-text corpora—ranging from 660K (Li et al.,

2023) to 32M (Wu et al., 2025) samples— which are widely recognized as costly to curate and subject to label noise and privacy constraints in medical settings. Moreover, prior studies indicate that outcome-level supervision provides limited visibility into intermediate model behavior, raising concerns regarding interpretability and clinician trust (Liu et al., 2025b; Chen et al., 2024a; Wang et al., 2025a). Consistent with these observations, SFT-based models have been reported to exhibit factual hallucinations when generated reports lack strong visual grounding.

2.2 RL for MLLMs

To mitigate the issue of hallucination in MLLMs, RL has been increasingly adopted to align model behavior with human preferences. Within this landscape, GRPO has demonstrated favorable architectural compatibility with LLMs compared to conventional approaches such as PPO and DPO. Recent studies have begun to explore GRPO in general multimodal settings (Shen et al., 2025; Yang et al., 2025b; Zheng et al., 2025); however, its extension to the medical domain remains challenging. For example, Lai et al. (2025) and Pan et al. (2025) apply GRPO to medical VQA tasks, but report critical logical inconsistencies between the model's intermediate reasoning traces and its final answers. Moreover, applying GRPO to RRG is further complicated by the intrinsic structure of clinical narratives, which do not naturally exhibit the explicit "thought–answer" decomposition assumed by standard reasoning-oriented frameworks.

3 Methodology

In this section, we present a comprehensive framework for self-consistent radiology report generation. Our approach is grounded in a two-phases methodology: (i) a systematic study to identify the optimal vision-language backbone (Sec. 3.1), and (ii) a novel self-consistent reinforcement learning framework that enforces logical consistency and clinical accuracy (Sec. 3.2). An overview of our proposed framework is shown in Figure 1.

3.1 Phase I: Systematic Architectural Optimization

The performance of an MLLM for RRG critically depends on both visual representation quality and the reasoning capacity of the language backbone. To provide a strong foundation for our RL framework, we conduct a systematic study to disentangle the respective contributions of the vision encoder and the LLM. Specifically, we define the architectural **search space** as $\mathcal{S} = \mathcal{V} \times \mathcal{L}$, where $\mathcal{V}$ and $\mathcal{L}$ denote the candidate vision encoder set and LLM backbone set, respectively. For the vision encoders ($\mathcal{V}$), we evaluate diverse inductive biases and pre-training paradigms, including CLIP, SigLIP, Xray-CLIP, and Dinov2-Xray-pretrained ViTs. This design enables a direct comparison between general-domain scaling and domain-specific medical pre-training. For the language models ($\mathcal{L}$), we benchmark both general-purpose reasoning models (e.g., Llama-3-8B) and medically adapted models (e.g., Med-Llama). Across all configurations, we adopt a unified connector strategy based on a multi-layer perceptron (MLP) projection to interface the vision encoder with the LLM.

We evaluate all pairs $(v, l) \in \mathcal{S}$ on the MIMIC-CXR dataset using a holistic metric suite. The optimal pair (v^*, l^*) is identified as the backbone policy for the subsequent RL optimization.

3.2 Phase II: Self-Consistent Reinforcement Learning Framework

With the optimal backbone fixed, we focus on improving hallucination control and internal consistency in report generation. Rather than modeling RRG as a pattern-matching problem, we formulate it as a structured reasoning process.

3.2.1 "Reason-then-Summarize" Architecture

Let $\mathcal{D} = \{(x^{(i)}, y^{(i)}_{\text{text}}, y^{(i)}_{\text{label}})\}_{i=1}^{N}$ denote a dataset of radiology studies, where $x^{(i)}$ denotes a chest X-ray image, $y^{(i)}_{\text{text}}$ is the corresponding free-text report (i.e., the findings section), and $y^{(i)}_{\text{label}}$ represents the associated structured diagnostic label vector (e.g., the fourteen CheXpert pathologies). We aim to learn a policy π_θ, parameterized by an MLLM, that generates a sequence of tokens Y. To align with the hierarchical formulation of RL, we conceptualize the generation of the "findings" section as the model's reasoning process and the diagnostic labels as the final answer. Accordingly, the MLLM output is structured into two components: (i) `<think><\think>` block containing a detailed, free-text narrative of radiological observations (e.g., "Opacity is observed in the right lower lobe."). (ii) `<answer><\answer>` block including a structured JSON object derived solely from the preceding narrative, mapping the findings to discrete clinical labels (e.g., "Pneumonia": 1.0, "No Finding": 0.0).

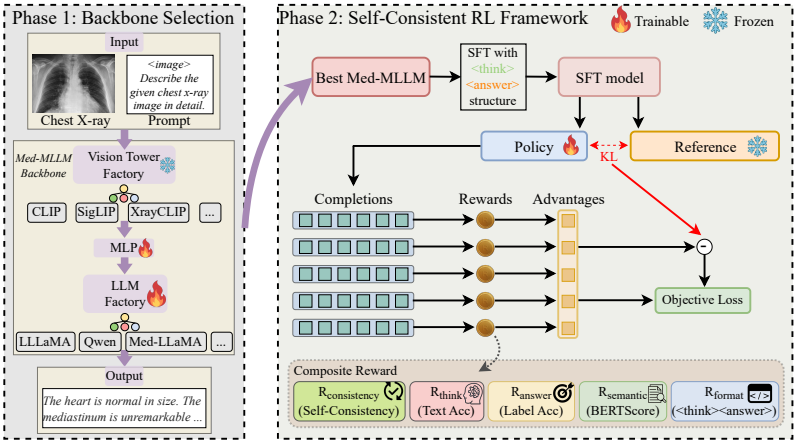


Figure 1: Overview of the proposed framework. Phase 1 performs a systematic exploration to identify the optimal vision–language backbone, while Phase 2 introduces a self-consistent reinforcement learning framework to enforce logical rigor and clinical accuracy.

Unlike conventional approaches that directly map $x \to y_{\text{text}}$, we formulate the generation process as a two-step causal chain:

$$\begin{aligned} P(Y|x) &= P(Y_{\text{think}}, Y_{\text{answer}}|x) \\ &= \underbrace{P(Y_{\text{think}}|x)}_{\text{Reasoning}} \cdot \underbrace{P(Y_{\text{answer}}|x, Y_{\text{think}})}_{\text{Summarization}}, \end{aligned} \quad (1)$$

where Y_{think} denotes the descriptive findings (reasoning) and Y_{answer} indicates the structured diagnostic labels (answer) conditioned on those findings.

3.2.2 Composite Reward Function Design

We design a composite reward function consisting of five complementary components to jointly enforce internal consistency, clinical accuracy, and format compliance during reinforcement learning.
(i) Consistency Reward ($R_1 = R_{\text{consistency}}$) measures the consistency between the model's generated narrative in `<think>` and its diagnostic conclusion in `<answer>` by using a "Self-Check" mechanism, which constitutes our core innovation:

$$R_1 = \text{CFS}(\mathcal{E}(\hat{Y}_{\text{think}}), \hat{Y}_{\text{answer}}) \quad (2)$$

Here, $\mathcal{E}$ denotes the CheXbert labeler used to map free-text reports to structured pathology labels. To quantify diagnostic alignment, we define the *Clinical F1 Score* (CFS) as the arithmetic mean of weighted rewards across pathology categories. Given a set of diagnostic labels $\mathcal{C}$ (e.g., $|\mathcal{C}| = 14$ pathologies), CFS is defined as

$$\text{CFS}(\hat{Y}, Y) = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \mathcal{S}(\hat{y}_c, y_c) \quad (3)$$

where $\mathcal{S}$ is an asymmetric scoring function. Specifically, positive ground-truth cases receive a maximal reward of $+2.0$ to emphasize sensitivity, negative cases receive $+1.0$ to emphasize specificity, and uncertain labels are assigned a neutral reward of $+0.5$. Definitive diagnostic errors (false positives or false negatives) incur a mild penalty of -0.3. The normalization ensures that the resulting score remains stable with respect to the number of evaluated pathology categories.
(ii) Reasoning Accuracy Reward ($R_2 = R_{\text{think-acc}}$) measures the diagnostic accuracy of the free-text report against ground-truth labels by:

$$R_2 = \text{CFS}(\mathcal{E}(\hat{Y}_{\text{think}}), Y_{\text{answer}}). \quad (4)$$

(iii) Diagnostic Accuracy Reward ($R_3 = R_{\text{answer-acc}}$) measures the accuracy of the final structured output against ground truth labels by:

$$R_3 = \text{CFS}(\hat{Y}_{\text{answer}}, Y_{\text{answer}}). \quad (5)$$

(iv) Semantic Fidelity Reward ($R_4 = R_{\text{semantic}}$) ensures that the linguistic quality and semantic meaning of the generated report align with the radiologist's reference report. We use BERTScore to capture semantic similarity beyond n-gram overlap.

$$R_4 = \text{BERTScore}(\hat{Y}_{\text{think}}, Y_{\text{think}}). \quad (6)$$

(v) Format Compliance Reward ($R_5 = R_{\text{format}}$) enforces structural constraints using a binary cumulative function. The reward is defined as:

$$R_5 = 0.5 \cdot \mathbb{I}_{\text{tags}} + 0.5 \cdot \mathbb{I}_{\text{json}}, \quad (7)$$

where $\mathbb{I}_{\text{tags}}$ returns 1 if both `<think>` and `<answer>` delimiters are present, and $\mathbb{I}_{\text{json}}$ returns 1 if the content within `<answer>` is a valid JSON object. This yields a maximum reward of 1.0 only for fully compliant outputs.

The final reward is computed as a weighted sum of R_1–R_5:

$$R_{\text{total}} = \sum_{i=1}^{5} \lambda_i R_i, \tag{8}$$

where λ_1–λ_5 are hyperparameters controlling the relative importance of each reward component.

3.2.3 Loss Function

The training procedure consists of three steps: (i) *Sampling*. For each input v, we sample G candidate outputs $\{o_i\}_{i=1}^{G}$ from the reference policy $\pi_{\theta_{\text{old}}}$. (ii) *Advantage Estimation*. We compute a relative advantage A_i for each candidate by normalizing its reward r_i with respect to the group statistics:

$$A_i = \frac{r_i - \text{mean}(\{r_1, \ldots, r_G\})}{\text{std}(\{r_1, \ldots, r_G\}) + \epsilon}. \tag{9}$$

This formulation encourages trajectories that outperform the group average. (iii) *Policy Optimization*. The policy π_θ is updated by maximizing the GRPO objective $\mathcal{J}_{\text{GRPO}}$, which employs a clipped surrogate loss to regularize updates and ensure training stability:

$$\begin{aligned}\mathcal{J}_{\text{GRPO}}(\theta) = \ & \mathbb{E}_{v\sim P(V)}[\mathbb{E}_{\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|v)}[\\ & \frac{1}{G}\sum_{i=1}^{G}(\min(r_i^{\text{ratio}}A_i,\ \text{clip}(r_i^{\text{ratio}}, 1\pm\epsilon)A_i) \\ & - \beta\,\mathbb{D}_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}}))]],\end{aligned} \tag{10}$$

where

$$r_i^{\text{ratio}} = \frac{\pi_\theta(o_i \mid v)}{\pi_{\theta_{\text{old}}}(o_i \mid v)}. \tag{11}$$

An additional Kullback–Leibler (KL) term $\mathbb{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})$ penalizes divergence from the reference policy π_{ref}. The coefficients $\epsilon, \beta \in \mathbb{R}_{\geq 0}$ control the regularization strengths.

4 Experimental Setup

4.1 Datasets and Protocols

We evaluate our method on two widely used radiology report generation datasets. **MIMIC-CXR** (Johnson et al., 2024): The largest publicly available chest X-ray dataset, comprising 377,110 images and 227,835 radiology reports. We follow the official data split (222,758 training, 1,808 validation, and 3,269 testing samples). The `<think>` component is extracted from the "Findings" section of each report, while the `<answer>` component is derived by applying the CheXbert labeler (Smit et al., 2020) to extract 14 structured disease labels, which are formatted as JSON outputs. **IU X-Ray** (Demner-Fushman et al., 2015): A smaller dataset collected by Indiana University, containing 7,470 images and 3,955 radiology reports.

4.2 Evaluation Metrics

We evaluate model performance along both linguistic quality and clinical diagnostic accuracy using a multi-dimensional evaluation framework comprising three categories: natural language generation (NLG) metrics, clinical efficacy (CE) metrics, and LLM-based evaluation measures. Linguistic performance is assessed using BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE-L (Lin, 2004), and BERTScore (Zhang et al., 2019). Clinical alignment is evaluated with RadGraph (Jain et al., 2021), RadCliQ (Yu et al., 2022), and F1 scores, complemented by the GREEN metric (Ostmeier et al., 2024) for LLM-based semantic assessment.

To particularly measure the internal consistency of the *"Reason–then–Summarize"* paradigm, we introduce a novel Self-Consistency Score (SCS), which quantifies the agreement between the model's intermediate reasoning trace (`<think>`) and its final diagnostic summary (`<answer>`):

$$SCS = \text{F1}\left(\mathcal{E}(\hat{Y}_{\text{think}}), \hat{Y}_{\text{answer}}\right) \tag{12}$$

A higher SCS indicates stronger agreement between the generated descriptive findings and the corresponding diagnostic outputs, reflecting improved internal consistency in the model's report generation process.

4.3 Baselines

We benchmark our framework against comprehensive SOTA baselines, grouped into two categories:

Specialized Medical MLLMs: We evaluate multiple domain-adapted architectures, including HuatuoGPT-Vision (Chen et al., 2024b), CheXagent (Chen et al., 2024c), LLaVA-Rad (Chaves et al., 2024), MedGemma (Sellergren et al., 2025b) and DeepMedix-R1 (Lin et al., 2025). These models are extensively pre-trained on large-scale biomedical corpora and constitute the current standard for RRG tasks, serving as primary baselines for evaluating clinical alignment between generated narratives and ground-truth annotations.

Finetuned General-Domain MLLMs: We also benchmark finetuned general-domain models, including LLaVA and Qwen-2.5/3-VL-Instruct. All models are finetuned on the MIMIC-CXR and IU-Xray datasets using a unified multi-task train-

LLM	ViT	B-1	B-2	B-3	B-4	Meteor	R-1	R-2	R-L	BERTScore	RadGraph	RadCliQ-v1(↓)	RadCliQ-v2(↓)	GREEN
Llama	CLIP	0.423	0.291	0.212	0.149	0.428	0.447	0.206	0.422	0.548	0.311	2.511	0.585	0.607
	DINOv2	0.425	0.289	0.207	0.141	0.422	0.448	0.200	0.421	0.546	0.315	2.493	0.588	0.610
	DINOv2-Xray	0.429	0.296	0.214	0.148	0.429	0.453	0.209	0.428	0.553	0.321	2.472	0.570	0.622
	XrayCLIP	0.438	0.307	0.227	**0.164**	0.438	**0.466**	**0.223**	**0.441**	**0.565**	**0.334**	**2.392**	**0.531**	0.623
	SigLIP	0.420	0.289	0.211	0.149	0.424	0.448	0.207	0.423	0.547	0.309	2.555	0.623	0.595
Qwen2	CLIP	0.417	0.286	0.206	0.143	0.424	0.442	0.202	0.418	0.540	0.300	2.582	0.642	0.596
	DINOv2	0.426	0.294	0.214	0.149	0.429	0.451	0.209	0.427	0.552	0.321	2.510	0.607	0.610
	DINOv2-Xray	0.434	0.302	0.221	0.156	0.434	0.461	0.215	0.435	0.562	0.333	2.458	0.556	**0.630**
	XrayCLIP	0.437	0.304	0.223	0.158	0.440	0.462	0.217	0.436	0.562	0.331	2.410	**0.531**	**0.630**
	SigLIP	0.426	0.293	0.212	0.147	0.427	0.449	0.205	0.423	0.547	0.313	2.508	0.592	0.605
Med-Llama	CLIP	0.440	0.300	0.216	0.152	0.439	0.451	0.204	0.425	0.547	0.309	2.566	0.628	0.587
	DINOv2	0.441	0.300	0.216	0.151	0.440	0.450	0.203	0.423	0.547	0.311	2.553	0.613	0.592
	DINOv2-Xray	0.447	0.309	0.225	0.160	0.448	0.458	0.211	0.432	0.556	0.315	2.462	0.568	0.608
	XrayCLIP	**0.452**	**0.312**	**0.228**	0.162	**0.453**	0.464	0.215	0.440	0.561	0.318	2.426	0.550	0.608
	SigLIP	0.433	0.296	0.215	0.153	0.440	0.444	0.203	0.420	0.542	0.300	2.586	0.647	0.578

Table 1: Overall performance on NLP and clinical efficacy (CE) metrics for different combinations of LLM backbones and vision encoders on radiology report generation.

ing objective that requires generating free-text reports, structured JSON outputs, and explicit "think-answer" reasoning chains to align with our proposed paradigm.

4.4 Implementation Details

We utilize TinyLLaVA system for two-phase training and the Transformer Reinforcement Learning (TRL) library to implement the GRPO algorithm. To ensure rigorous comparisons in Phase 2, we do not reuse model parameters finetuned in Phase 1; instead, models are re-initialized from the same base checkpoints pre-trained on standard LLaVA datasets. We design a two-step optimization strategy in Phase 2. (i) Cold Start: Models are first supervised fine-tuned for 2 epochs using the AdamW optimizer with a learning rate of $\eta = 2 \times 10^{-5}$. (ii) Policy Optimization: The model is subsequently optimized with the proposed composite reward function, using weighting coefficients $\lambda_1 = 0.2$, $\lambda_2 = 0.5$, $\lambda_3 = 1.0$, $\lambda_4 = 0.3$, and $\lambda_5 = 0.5$. The KL-divergence regularization coefficient is set to $\beta = 0.03$. All experiments are performed on NVIDIA A100 GPUs with 80GB of memory.

5 Experimental Results

5.1 Comparative Analysis of Vision-Language Backbones

In Phase 1, we investigate optimal encoder–LLM pairings by finetuning MLLMs for standard radiology report generation, without enforcing the *"Reason-then-Summarize"* structure. Table 1 reports a comparative evaluation of multiple vision–language configurations across both NLP and CE metrics. The results reveal a clear performance divergence across metric categories. Llama paired with XrayCLIP achieves the strongest performance on traditional linguistic metrics, including BLEU-4, ROUGE, and BERTScore. In contrast, clinical efficacy metrics favor different architectural choices: DINOv2-Xray combined with Qwen2 yields the best GREEN score, while XrayCLIP-based models consistently perform well on RadGraph F1 and RadCliQ. This discrepancy highlights a trade-off between surface-level lexical similarity and clinical validity, indicating that high n-gram overlap does not necessarily imply improved diagnostic correctness.

Our results also indicate that the vision encoder emerges as the dominant factor influencing performance. Domain-specific encoders such as XrayCLIP and DINOv2-Xray consistently outperform general-purpose counterparts, a trend most clearly reflected by the GREEN metric. Notably, combining domain-pretrained vision encoders with domain-specific LLMs does not always lead to further gains, suggesting diminishing complementarity between highly specialized components. Considering both linguistic quality and clinical alignment, we select **Llama + XrayCLIP** as the backbone for Phase 2.

5.2 Comparative Analysis of Diagnostic Label Prediction

We evaluate the diagnostic label accuracy of the proposed method by comparing it against two baseline categories summarized in Table 2: *Report-only* models, which derive labels from generated narrative reports, and *Answer-only* models, which perform direct visual classification. On the MIMIC-CXR benchmark, our method consistently outperforms all baselines across evaluation metrics. In particular, it achieves a Macro-F1 of 37.36 and a Micro-F1 of 57.84, surpassing the strongest direct classification baseline (LLaVA-Next-Mistral) by a substantial margin. These results indicate that the proposed approach more effectively captures fine-grained pathological information than conventional vision–language baselines.

Similar trends are observed on the smaller IU-

	MIMIC-CXR		IU-Xray	
	Macro-F1	Micro-F1	Macro-F1	Micro-F1
Report-only				
HuatuoGPT-Vision	22.22	35.32	10.19	18.53
Chexagent	12.75	18.18	13.64	23.49
LLaVA-Rad	31.54	47.53	21.31	37.90
MedGemma	33.47	50.41	21.95	37.56
Llava-1.5 (SFT)	23.46	40.98	12.29	32.24
Llava-Next (Mistral)	26.28	45.64	15.51	31.76
Llava-Next (Vicuna)	22.99	42.12	15.04	39.08
Qwen-3-VL-8B	23.73	42.91	16.04	36.57
Qwen-2.5-VL-8B	20.71	39.53	13.05	25.53
DeepMedix-R1	31.26	51.89	**24.98**	39.05
Answer-only				
Llava-1.5 (SFT)	29.23	50.34	13.43	35.83
Llava-Next (Mistral)	33.21	55.11	15.43	35.71
Llava-Next (Vicuna)	30.81	51.89	16.06	39.35
Qwen-2.5-VL-7B	29.46	49.72	14.43	27.55
Qwen-3-VL-8B	32.00	52.68	16.12	30.54
Ours	**37.36**	**57.84**	16.24	**40.04**

Table 2: Diagnostic classification performance comparison between our proposed method and two baseline categories: Report-only models (labels derived from generated reports) and Answer-only models (direct visual classification), evaluated on the MIMIC-CXR and IU-Xray datasets. We report Macro-F1 and Micro-F1 scores (%) to assess label prediction accuracy. Best results are highlighted in bold.

Xray dataset, with dataset-specific characteristics. While MedGemma attains competitive Macro-F1 scores in this limited-data setting, our method maintains a clear advantage in Micro-F1 (40.04), reflecting improved instance-level diagnostic accuracy. Moreover, the larger performance gap observed on the MIMIC-CXR benchmark suggests superior scalability as dataset size increases. Overall, these results demonstrate that the proposed *"Reason-then-Summarize"* paradigm produces diagnostic labels that outperform those derived from both traditional report-based pipelines and direct visual classification models.

5.3 Effectiveness of Consistency-Aware Reinforcement Learning

In this section, we evaluate the effectiveness of consistency-aware reinforcement learning from three complementary perspectives. We first examine whether our RL-based method improves clinical efficacy over SFT methods under the same structured *think–answer* generation setting. We then analyze whether the explicit consistency reward ($R_{\text{consistency}}$) effectively reduce factual hallucinations and logical conflicts between the generated narrative and the diagnostic labels. Finally, we conduct an ablation study to explain the contribution of each reward term in the composite objective.

(i) Clinical Efficacy. We begin by comparing reinforcement learning with supervised fine-tuning under an identical output format. While the SFT baseline attains slightly higher surface-level fluency (e.g., BLEU-4: 0.130 vs. 0.112), this advantage does not translate into improved clinical accuracy. In contrast, optimizing directly for diagnostic objectives yields consistent gains across clinically meaningful metrics. Specifically, the full RL model improves the Report-Level Macro-F1 from 36.40% to 37.01% and the Answer-Level Macro-F1 from 35.71% to 37.36%. These results suggest that explicit alignment with diagnostic correctness, rather than imitation of reference phrasing, is the primary driver of clinical efficacy.

(ii) Consistency and Hallucination Reduction. We next investigate the role of the consistency reward $R_{\text{consistency}}$. Removing this term leads to a substantial degradation in self-consistency, with the SCS Macro-F1 dropping from 92.16% to 75.68%. This collapse indicates that, in the absence of explicit penalties, the model frequently produces diagnostic labels that contradict its own generated findings. By contrast, the full model maintains consistently high agreement between reasoning and diagnosis (SCS Micro-F1 exceeding 97%), demonstrating that $R_{\text{consistency}}$ is essential for suppressing factual hallucinations and enforcing internal coherence between narrative and labels.

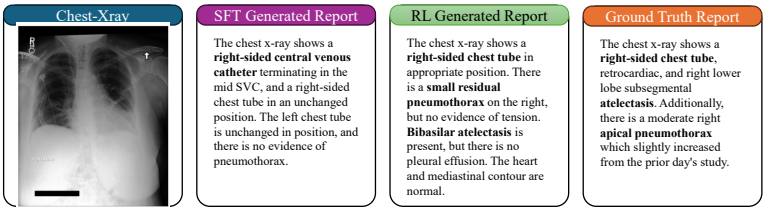
(iii) Component-Wise Ablation Analysis. Finally, we analyze how individual reward components shape model behavior. As expected, removing R_1 results in the most severe degradation in reliability and consistency. However, enforcing consistency alone is insufficient for high-quality report generation. When the reasoning and semantic rewards (R_2 and R_4) are excluded, the model produces outputs that are internally consistent but descriptively impoverished, yielding the lowest semantic fidelity (BERTScore: 0.554). This behavior suggests a tendency toward conservative yet uninformative reports. Overall, the full RL configuration achieves the best balance by combining strict logical constraints (R_1, R_3) with rewards that encourage informative and meaningful descriptions (R_2, R_4).

5.4 Qualitative Analysis

Figure 2 provides a qualitative comparison illustrating the behavioral differences between the SFT baseline and our reinforcement learning framework. The SFT model exhibits notable hallucinations, including the incorrect description of a "right-sided central venous catheter" that is absent in the ground truth. It also demonstrates internal inconsistencies, where the generated narrative con-

	Report-Level						Answer-Level		SCS	
	NLG Metrics				Clinical Efficacy					
	B-4	MTR	RG-L	BERTScore	Macro-F1	Micro-F1	Macro-F1	Micro-F1	Macro-F1	Micro-F1
Llava-1.5	0.113	0.356	0.375	0.614	24.78	44.34	25.57	45.51	90.05	94.93
Llava-Next(vicuna)	0.116	0.360	0.378	0.617	27.30	46.74	27.86	47.72	91.14	94.92
Llava-Next(mistral)	0.117	0.365	0.382	0.625	31.46	50.39	32.38	51.21	90.69	94.38
Qwen-2.5-VL-7B	0.112	0.358	0.375	0.617	27.72	46.77	27.18	45.86	89.76	95.08
Qwen-3-VL-8B	0.118	0.368	0.383	0.626	30.97	40.82	31.48	50.84	91.29	94.72
Ours(SFT)	**0.130**	**0.375**	**0.401**	0.627	36.40	55.70	35.71	56.63	89.06	95.83
Ours(RL (full))	0.112	0.363	0.376	**0.631**	**37.01**	**56.97**	**37.36**	**57.84**	**92.16**	97.39
Ours(RL ($R_3 + R_5$))	0.108	0.307	0.375	0.602	34.75	53.31	34.09	53.57	89.45	97.02
Ours(RL (w/o R_3))	0.091	0.268	0.302	0.544	34.03	52.18	33.23	52.18	89.30	96.12
Ours(RL (w/o R_1))	0.082	0.242	0.298	0.509	27.12	45.96	27.42	43.43	75.68	88.62
Ours(RL (w/o $R_2 + R_4$))	0.104	0.282	0.328	0.554	34.96	53.94	35.03	54.45	92.10	**97.67**

Table 3: Main results on radiology report generation and clinical efficacy. We compare the proposed SFT and RL stages with multiple MLLM baselines across NLG metrics, CE metrics, and the Self-Consistency Score (SCS), which evaluates alignment between generated reports and diagnostic labels. Incorporating the explicit consistency reward ($R_{\text{consistency}}$) in *Ours (RL (full))* leads to improved clinical accuracy and consistency.



Structured Finding Comparison

	Atelectasis	Pneumothorax	Support Devices	Pleural Effusion	Cardiomegaly	No Finding
Report Label (SFT)	0	0	1	-	-	0
Answer Label (SFT)	0	1	1	-	-	0
Report Label (RL)	1	1	1	0	-	0
Answer Label (RL)	1	1	1	0	-	0
Ground Truth	1	1	1	0	-	0

Figure 2: Qualitative comparison of report generation and internal coherence. The top row displays the original chest X-ray, the reports generated by SFT and RL models, and the Ground Truth report. The bottom table compares the structured findings derived from the generated report text ("Report findings") versus the model's explicit structured output ("Answer Label").

flicts with the corresponding structured diagnostic labels, such as contradictory indications regarding atelectasis and pneumothorax. In contrast, the RL-based model produces outputs that are more internally coherent and better grounded in the visual evidence. By explicitly enforcing alignment between the reasoning (`<think>`) and the diagnostic summary (`<answer>`), the model reduces factual hallucinations and avoids contradictions between textual findings and labels. As a result, the RL model more accurately captures subtle clinical observations, including "bibasilar atelectasis", reflecting improved fidelity to the ground truth and enhanced clinical reliability.

6 Conclusion

This study introduces a two-stage framework for radiology report generation that jointly optimizes vision–language pairing and consistency-aware learning. Through a systematic evaluation in Phase 1, we identify domain-specific vision encoders, such as XrayCLIP, as the primary determinant of performance, consistently outperforming general-purpose backbones. Notably, we observe that pairing domain-pretrained vision encoders with domain-pretrained LLMs can degrade performance, revealing non-linear interactions that challenge conventional architectural assumptions. Building on these insights, Phase 2 introduces a reinforcement learning framework that substantially improves clinical efficacy while reducing hallucinations and logical inconsistencies between generated reasoning and diagnostic outputs. Overall, our findings highlight that trustworthy radiology report generation requires explicit alignment between reasoning and diagnosis, rather than reliance on surface-level lexical similarity, and point toward future directions in adaptive pairing and consistency-driven training for medical MLLMs.

Limitations

Despite the efficacy of the *Reason-then-Summarize* framework, we acknowledge two primary limitations.

First, our use of a standard MLP projection layer for cross-modal alignment may limit the transmission of fine-grained radiological details. This choice ensures architectural consistency with the LLaVA-based baselines, allowing us to isolate the impact of different vision encoder–LLM pairings. While the robustness of our findings is supported by consistent gains in CE metrics, future work will explore advanced connectors (e.g., Q-Former) to enhance visual granularity.

Second, the reinforcement learning process induced a *length collapse* phenomenon, where the model prioritized succinct diagnostic statements over descriptive verbosity. This brevity adversely affects length-sensitive metrics (e.g., BLEU, ROUGE) and leads to the omission of explicitly normal findings, creating a divergence between linguistic scores and clinical utility. Although length regularization was explored, it was excluded from this version due to training instability. Future research will focus on balancing diagnostic precision with descriptive completeness.

Ethics Statement

This study utilizes the publicly available, de-identified MIMIC-CXR and IU X-Ray benchmarks, strictly adhering to PhysioNet credentialing and HIPAA Safe Harbor protocols. While our "Reason-then-Summarize" framework enhances logical consistency, the resulting models remain research prototypes and are not certified as Software as a Medical Device (SaMD). Given the residual risk of generative confabulation inherent to Large Language Models, these systems are designed exclusively as assistive tools for human augmentation rather than autonomous agents. Consequently, any future clinical deployment necessitates rigorous prospective validation, regulatory clearance, and the implementation of mandatory human-in-the-loop verification protocols to ensure patient safety.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

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

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *IEEvaluation@ACL*.

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, Hany Hassan Awadalla, Julia Gong, Houdong Hu, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Yu Gu, Cliff Wong, Mu-Hsin Wei, and 7 others. 2024. A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings. *Nature Communications*, 16.

Jiawei Chen, Dingkang Yang, Tong Wu, Yue Jiang, Xiaolu Hou, Mingcheng Li, Shunli Wang, Dongling Xiao, Ke Li, and Lihua Zhang. 2024a. Detecting and evaluating medical hallucinations in large vision language models. *arXiv preprint arXiv:2406.10185*.

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

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*.

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.

Pengfei Gu, Haoteng Tang, Islam A Ebeid, Jose A Nunez, Fabian Vazquez, Diego Adame, Marcus Zhan, Huimin Li, Bin Fu, and Danny Z Chen. 2025. Adapting a segmentation foundation model for medical image classification. In *2025 IEEE 38th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 167–172. IEEE.

Dennis Hein, Zhihong Chen, Sophie Ostmeier, Justin Xu, Maya Varma, Eduardo Pontes Reis, Arne Edward Michalson Md, Christian Bluethgen, Hyun Joo Shin, Curtis Langlotz, and 1 others. 2025. Chexalign: Preference fine-tuning in chest x-ray interpretation models without human feedback. In *Proceedings*

of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 27679–27702.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, D. Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curt P. Langlotz, and Pranav Rajpurkar. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. *ArXiv*, abs/2106.14463.

Alistair Johnson, Tom Pollard, Roger Mark, Seth Berkowitz, and Steven Horng. 2024. Mimic-cxr database. *PhysioNet10*, 13026:C2JT1Q.

Jiayi Kuang, Ying Shen, Jingyou Xie, Haohao Luo, Zhe Xu, Ronghao Li, Yinghui Li, Xianfeng Cheng, Xika Lin, and Yu Han. 2025. Natural language understanding and inference with mllm in visual question answering: A survey. *ACM Computing Surveys*, 57(8):1–36.

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*.

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.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*.

Qika Lin, Yifan Zhu, Bin Pu, Ling Huang, Haoran Luo, Jingying Ma, Zhen Peng, Tianzhe Zhao, Fangzhi Xu, Jian Zhang, and 1 others. 2025. A foundation model for chest x-ray interpretation with grounded reasoning via online reinforcement learning. *arXiv preprint arXiv:2509.03906*.

Fenglin Liu, Hongjian Zhou, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S Chen, Yining Hua, Peilin Zhou, and 1 others. 2025a. Application of large language models in medicine. *Nature Reviews Bioengineering*, pages 1–20.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Hong Liu, Dong Wei, Zhe Xu, Xian Wu, Yefeng Zheng, and Liansheng Wang. 2025b. Rrg-dpo: Direct preference optimization for clinically accurate radiology report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 552–562. Springer.

Tengfei Liu, Jiapu Wang, Yongli Hu, Mingjie Li, Junfei Yi, Xiaojun Chang, Junbin Gao, and Baocai Yin. 2025c. Hc-llm: Historical-constrained large language models for radiology report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5595–5603.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson, Michael E. Moseley, Curtis P. Langlotz, Akshay S. Chaudhari, and Jean-Benoit Delbrouck. 2024. Green: Generative radiology report evaluation and error notation. *ArXiv*, abs/2405.03595.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Noam Rotstein, David Bensaid, Shaked Brody, Roy Ganz, and Ron Kimmel. 2024. Fusecap: Leveraging large language models for enriched fused image captions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 5689–5700.

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

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla P. Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, and 62 others. 2025b. Medgemma technical report. *ArXiv*, abs/2507.05201.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, and 1 others. 2025. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*.

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*.

Haoteng Tang, Xiyao Fu, Lei Guo, Yalin Wang, Scott Mackin, Olusola Ajilore, Alex Leow, Paul Thompson, Heng Huang, and Liang Zhan. 2022. Functional2structural: Cross-modality brain networks representation learning. *arXiv preprint arXiv:2205.07854*.

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, and 1 others. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356.

Chenyu Wang, Weichao Zhou, Shantanu Ghosh, Kayhan Batmanghelich, and Wenchao Li. 2025a. Semantic consistency-based uncertainty quantification for factuality in radiology report generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1739–1754, Albuquerque, New Mexico. Association for Computational Linguistics.

Dandan Wang and Shiqing Zhang. 2024. Large language models in medical and healthcare fields: applications, advances, and challenges. *Artificial intelligence review*, 57(11):299.

Pan Wang, Siwei Song, Hui Ji, Siqi Cao, Heng Yu, Zhijian Liu, Huanrui Yang, Yingyan (Celine) Lin, Beidi Chen, Mohit Bansal, Xiaoming Liu, Pengfei Zhou, Ming-Hsuan Yang, Tianlong Chen, and Jingtong Hu. 2025b. From models to systems: A comprehensive survey of efficient multimodal learning. *Authorea Preprints*.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. 2025. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 16(1):7866.

Dexuan Xu, Yanyuan Chen, Jieyi Wang, Yue Huang, Hanpin Wang, Zhi Jin, Hongxing Wang, Weihua Yue, Jing He, Hang Li, and 1 others. 2024. Mlevlm: Improve multi-level progressive capabilities based on multimodal large language model for medical visual question answering. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4977–4997.

Lingrui Yang, Yuxing Zhou, Jun Qi, Xiantong Zhen, Li Sun, Shan Shi, Qinghua Su, and Xuedong Yang. 2025a. Aligning large language models with radiologists by reinforcement learning from ai feedback for chest ct reports. *European Journal of Radiology*, 184:111984.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, and 1 others. 2025b. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*.

Ziruo Yi, Ting Xiao, and Mark V. Albert. 2025. A survey on multimodal large language models in radiology for report generation and visual question answering. *Information*.

Feng Yu, Masahiro Endo, Rayan Krishnan, Ian Pan, Andrew Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, H. H. Lee, Zohreh Hossein Abad, Andrew Y. Ng, C. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. 2022. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4.

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, and 1 others. 2025. A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings. *Nature Communications*, 16(1):3108.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.

Kun Zhao, Siyuan Dai, Yingying Zhang, Guodong Liu, Pengfei Gu, Chenghua Lin, Paul M Thompson, Alex Leow, Heng Huang, Lifang He, and 1 others. 2025. R-genima: Integrating neuroimaging and genetics with interpretable multimodal ai for alzheimer's disease progression. *arXiv preprint arXiv:2512.18986*.

Kun Zhao, Chenghao Xiao, Sixing Yan, Haoteng Tang, William K Cheung, Noura Al Moubayed, Liang Zhan, and Chenghua Lin. 2024. X-ray made simple: Lay radiology report generation and robust evaluation. *arXiv preprint arXiv:2406.17911*.

Yaowei Zheng, Junting Lu, Shenzhi Wang, Zhangchi Feng, Dongdong Kuang, and Yuwen Xiong. 2025. Easyr1: An efficient, scalable, multi-modality rl training framework.

Zijian Zhou, Miaojing Shi, Meng Wei, Oluwatosin Alabi, Zijie Yue, and Tom Vercauteren. 2024. Large model driven radiology report generation with clinical quality reinforcement learning. *arXiv preprint arXiv:2403.06728*.

A Appendix

A.1 Instruction

Table A.1 is an example of the prompt.

A.2 More Cases

We have provided more cases in Figure 3.

A.3 Systematic Failure Mode Analysis

In addition to qualitative assessment, we perform a systematic per-class F1 analysis (Table 4) to characterize failure patterns across clinical findings. The results reveal three primary error regimes:

Rare-class Instability: Findings such as *Enlarged Cardiomediastinum*, *Consolidation*, and *Fracture* exhibit consistently low F1 scores under both RL and SFT paradigms. This suggests that performance degradation is primarily a function of class imbalance and data sparsity rather than a byproduct of the reinforcement learning process.

Report-level Compression: We observe performance drops when transitioning from label-level to report-level F1 for specific findings (e.g., *Lung Opacity*). This indicates that errors often arise during the sentence-level aggregation and narrative synthesis phase, correlating with the "length collapse" phenomenon observed in our generation traces.

Diagnostic Bias: Both models prioritize salient pathological findings while frequently omitting explicitly "normal" statements. This prioritization explains the discrepancy between high CE metrics and lower SCS scores, reflecting a bias toward clinically significant content over surface-level completeness.

Importantly, our analysis demonstrates that RL does not introduce novel systematic failure modes relative to SFT. Improvements remain consistent in high-prevalence conditions (e.g., *Cardiomegaly*, *Atelectasis*), while rare findings remain challenging across both training regimes.

A.4 Computational Cost

All experiments in Phase 1 and Phase 2 were conducted on 8 NVIDIA A100 80GB GPUs using DeepSpeed ZeRO-3 with full-parameter fine-tuning under CUDA 12.2 and the TinyLLaVA framework. Phase 1 consists of pretraining and supervised fine-tuning. During the pretraining stage, both the LLM and vision encoder were frozen. We trained for one epoch with a batch size of 32, which required approximately 12 hours per vision encoder–LLM pair. Peak GPU memory consumption ranged between 60GB and 80GB per GPU. In the supervised fine-tuning stage, the vision encoder was frozen while the LLM and projector were updated. We trained for three epochs with a batch size of 16, requiring approximately 40 hours. In total, each encoder–LLM pair required about 52 hours of training time. This cost is comparable to standard LLaVA-style SFT training under the same hardware configuration, as Phase 1 follows an identical optimization paradigm.

Phase 2 includes a cold-start stage and reinforcement learning training. For cold start, we randomly selected 5,000 samples and trained with a batch size of 16, which required approximately 1 hour. For the RL stage, we used TRL version 0.25.0 and vLLM version 0.11.0, with 7 GPUs allocated for training and 1 GPU dedicated to the vLLM server. We set the number of generations per prompt to 16 and trained for 1,500 steps and other hyperparameters are followed the default setting of TRL. Due to on-policy sampling, vLLM returns generated reports approximately every 5 to 7 minutes before gradient updates are applied. The total RL training time was approximately 30 to 35 hours. Although RL introduces additional sampling overhead compared to SFT-only training, peak GPU memory consumption does not exceed the 80GB per GPU limit.

During inference, the model requires a single A100 80GB GPU. Using vLLM for optimized decoding, the average generation latency per medical report is approximately 1.5 to 2.0 seconds, depending on report length. Memory usage during inference is approximately 40GB to 50GB, which is substantially lower than during training. Importantly, although Phase 2 increases total training time relative to SFT baselines, inference cost and deployment requirements remain unchanged, as both approaches share the same architecture and parameter size.

A.5 Sensitivity Analysis of Reward Weights

Regarding the sensitivity of reward weights (λ_1–λ_5), our study includes a component-wise ablation analysis (Table 3) that systematically evaluates the contribution of each reward term. The results demonstrate that removing the consistency reward (R_1) leads to a substantial decline in SCS (from 92.16% to 75.68%), indicating that the framework's behavior is primarily governed by the structural reward design rather than delicate weight tuning. Similarly, excluding R_2 and R_4 reduces semantic fidelity without improving consistency, confirming the complementary functional roles of these components.Under the current configuration ($\lambda_1 = 0.2, \lambda_2 = 0.5, \lambda_3 = 1.0, \lambda_4 =$

Class	RL (L)	SFT (L)	RL (R)	SFT (R)	RL (SCS)	SFT (SCS)
No Finding	0.7368	0.7175	0.7359	0.7150	0.9804	0.9851
Enlarged Cardiomediastinum	0.0059	0.0271	0.0000	0.0111	0.5455	0.6061
Cardiomegaly	0.6024	0.5688	0.6063	0.5676	0.9775	0.9722
Lung Opacity	0.5000	0.4321	0.4508	0.3844	0.9325	0.9148
Lung Lesion	0.1310	0.1717	0.1310	0.1753	1.0000	0.9683
Edema	0.5200	0.5249	0.5133	0.5220	0.9905	0.9723
Consolidation	0.0320	0.0588	0.0331	0.0593	0.8000	0.9778
Pneumonia	0.1419	0.0468	0.1812	0.2132	0.7531	0.3200
Atelectasis	0.5147	0.4282	0.5147	0.4266	0.9964	0.9928
Pneumothorax	0.4978	0.4588	0.5000	0.4549	0.9770	0.9938
Pleural Effusion	0.6850	0.6953	0.6746	0.6963	0.9843	0.9949
Pleural Other	0.0323	0.0000	0.0323	0.0303	1.0000	0.8750
Fracture	0.0382	0.0494	0.0382	0.0497	1.0000	0.9524
Support Devices	0.7926	0.8198	0.7703	0.7907	0.9651	0.9437

Table 4: Per-class F1 analysis at Label (L) and Report (R) levels, including Semantic Consistency Scores (SCS).

$0.3, \lambda_5 = 0.5$), the full RL model consistently outperforms strong SFT baselines across clinical efficacy metrics; for instance, Answer-Level Macro-F1 increases from 35.71% to 37.36%, and SCS Macro-F1 improves from 89.06% to 92.16%. These gains were achieved without exhaustive hyperparameter sweeps, suggesting that performance improvements stem from the structural reward formulation rather than fine-grained coefficient optimization. Furthermore, pilot experiments showed stable behavior within a reasonable neighborhood of these values, indicating that the framework is not overly sensitive to precise weight selection.To address potential biases introduced by CheXpert label uncertainties, our design incorporates three mitigation strategies. First, uncertain labels are assigned a neutral reward ($+0.5$) rather than being treated as definitive positives or negatives, thereby reducing gradient bias toward noisy supervision. Second, we integrate both text-level (R_2) and answer-level (R_3) accuracy rewards, preventing over-reliance on any single noisy labeling signal. Third, the semantic fidelity reward (R_4) ensures alignment with the full free-text report rather than solely structured labels, providing an additional supervision pathway that is less sensitive to label artifacts.Ultimately, the composite reward is designed to remain interpretable, clinically grounded, and computationally efficient—attributes we believe are essential for the development of trustworthy models in medical deployment.

Prompt Example

<|user|>

You are an expert radiologist. Your task is to analyze the given chest X-ray image, write a detailed finding, and then provide a structured summary of your finding.

You must follow this two-step process:

1. First, enclose your detailed, free-text radiology findings in `<think>` tags. This is your reasoning where you describe what you see.
2. Second, based **ONLY** on the text you wrote in `<think>`, provide a structured JSON object in `<answer>` tags. The JSON must summarize your findings according to the 14 official CheXpert labels.

CheXpert labels (14):

- Atelectasis
- Cardiomegaly
- Consolidation
- Edema
- Enlarged Cardiomediastinum
- Fracture
- Lung Lesion
- Lung Opacity
- No Finding
- Pleural Effusion
- Pleural Other
- Pneumonia
- Pneumothorax
- Support Devices

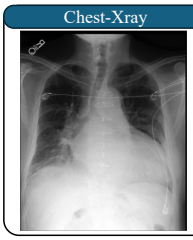
Use 1.0 for positive findings, 0.0 for negative, and -1.0 for uncertain.

<|assistant|>

EXAMPLE:

```
<think>
The cardiomediastinal silhouette is within normal limits. The lungs are clear.
There is no pleural effusion or pneumothorax. A port-a-cath is in place, with its
tip in the expected location.
</think>
<answer>
{
  "Atelectasis": 0.0,
  "Cardiomegaly": 0.0,
  "Consolidation": 0.0,
  "Edema": 0.0,
  "Enlarged Cardiomediastinum": 0.0,
  "Fracture": 0.0,
  "Lung Lesion": 0.0,
  "Lung Opacity": 0.0,
  "No Finding": 1.0,
  "Pleural Effusion": 0.0,
  "Pleural Other": 0.0,
  "Pneumonia": 0.0,
  "Pneumothorax": 0.0,
  "Support Devices": 1.0
}
</answer>
```

Chest-Xray



SFT Generated Report

The chest x-ray shows moderate cardiomegaly with pulmonary vascular congestion and mild interstitial edema. There is a small right pleural effusion and no pneumothorax. The patient has a history of median sternotomy and CABG. The impression is cardiomegaly with pulmonary vascular congestion and mild interstitial edema.

RL Generated Report

The chest x-ray shows moderate cardiomegaly with mild pulmonary edema. There is no large pleural effusion or pneumothorax.

Ground Truth Report

The chest x-ray reveals midline sternotomy wires and mediastinal clips, with stable mediastinal contour. No pneumothorax or large effusion is present, and there are intact osseous structures. Additionally, there is no free air below the right hemidiaphragm. However, the heart remains markedly enlarged, and there is mild pulmonary edema noted. These findings suggest marked cardiomegaly with mild pulmonary edema.

Structured Finding Comparison

	Cardiomegaly	Edema	Enlarged Cardiomediastinum	Pleural Effusion	Pneumothorax	No Finding
Report Label (SFT)	1	1	-	1	0	0
Answer Label (SFT)	1	1	-	1	-	0
Report Label (RL)	1	1	-	0	-	0
Answer Label (RL)	1	1	-	0	-	0
Ground Truth	1	1	-1	0	-	0

Chest-Xray

SFT Generated Report

The chest x-ray shows an endotracheal tube terminating 5.5 cm above the carina. The right IJ catheter terminates at the cavoatrial junction. There is no pneumothorax. The cardiomediastinal silhouette is enlarged, and there are bilateral parenchymal opacities, which may be due to pulmonary edema or pneumonia. There is no large pleural effusion.

RL Generated Report

The chest x-ray shows mild pulmonary edema, cardiomegaly, and a right-sided central venous catheter in the upper SVC. There is no large pneumothorax or pleural effusion. The mediastinum appears widened.

Ground Truth Report

The chest X-ray shows slightly limited evaluation due to overlying external artifacts. The endotracheal tube is appropriately positioned with the tip terminating 4.7 cm above the carina. There is mild pulmonary interstitial edema. The mediastinum appears widened, which is probably due to the AP technique and supine positioning. The cardiac silhouette is mildly enlarged, likely due to the AP view. There is no pleural effusion or pneumothorax, and there is no definite free air under the diaphragm.

Structured Finding Comparison

	Cardiomegaly	Edema	Enlarged Cardiomediastinum	Lung Opacity	Support Devices	Pneumonia
Report Label (SFT)	-	-1	-	1	1	-1
Answer Label (SFT)	-	-1	1	1	1	-1
Report Label (RL)	1	1	1	-	1	-
Answer Label (RL)	1	1	1	-	1	-
Ground Truth	1	1	1	-	1	0

Chest-Xray

SFT Generated Report

The chest x-ray shows a normal heart size with a stable mediastinal contour. The lungs appear hyperinflated with flattening of the diaphragms, which is consistent with COPD. There is a new opacity in the left lower lobe, which could be due to pneumonia or atelectasis. No pleural effusion or pneumothorax is observed. The impression is that there is a new left lower lobe opacity, which could be due to pneumonia or atelectasis.

RL Generated Report

The chest x-ray shows a small left pleural effusion with adjacent atelectasis. There is no evidence of pneumothorax or focal airspace consolidation. The heart size is normal, and the mediastinal and hilar contours are unremarkable.

Ground Truth Report

Findings show a small to moderate left pleural effusion with mild atelectasis, improvement of atelectasis at the right lung base and no focal consolidation, pneumothorax, or pulmonary edema. The cardiomediastinal silhouette is normal. The impression is that there is an interval re-accumulation of a small to moderate left pleural effusion.

Structured Finding Comparison

	Pleural Effusion	Edema	Enlarged Cardiomediastinum	Lung Opacity	Atelectasis	Pneumonia
Report Label (SFT)	0	-	-1	1	-1	-1
Answer Label (SFT)	0	-	0	1	-1	-1
Report Label (RL)	1	-	0	-	1	-
Answer Label (RL)	1	-	0	-	1	-
Ground Truth	1	-1	0	-	1	-

Figure 3: More cases for qualitative comparison of report generation and internal coherence. The top row displays the original chest X-ray, the reports generated by SFT and RL models, and the Ground Truth report. The bottom table compares the structured findings derived from the generated report text ("Report Label") versus the model's explicit structured output ("Answer Label").