

GRAIL: Grounded Visual Alignment of Medical Vision-Language Models for Radiology Report Generation

Sarosij Bose^{1,2†} Ravi K. Rajendran² Biplob Debnath²
Konstantinos Karydis¹ Amit K. Roy-Chowdhury¹ Srimat T. Chakradhar²
¹University of California, Riverside ²NEC Laboratories America

Abstract

Radiology Report Generation (RRG) aims to reduce the burden of clinical documentation, yet generating reports that remain visually grounded in the input image remains challenging. Current medical vision-language models (Med-VLMs) often rely on curated preference data, retrieval augmentation, or large-scale labeled corpora, but remain prone to hallucinated findings due to weak cross-modal alignment. To address this challenge, we propose **GRAIL**: Grounded Visual Alignment of Medical Vision-Language Models for Radiology Report Generation, a two-stage multimodal alignment framework for grounded RRG. First, *Clinically Informed Textual Reasoning* improves report structure and medical terminology using verifiable language rewards with periodic clinical guidance. Second, *Expert-guided Visual Reasoning* leverages a domain expert to rank generated reports using image-report alignment scores, encouraging visually grounded report generation without curated preference pairs or retrieval data. Experiments on multiple benchmarks show that **GRAIL** consistently improves both report generation quality and clinical accuracy, achieving relative gains over strong post-training baselines by up to **15%** in clinical accuracy and **13%** in generation quality. **GRAIL** also outperforms proprietary general-purpose and medical VLMs, highlighting the effectiveness of self-supervised multimodal post-alignment for reducing hallucinations in visually grounded radiology report generation.

1 Introduction

Automated chest X-ray report generation has emerged as a promising approach for reducing the reporting burden on radiologists. Recent advances in Medical Vision-Language Models (Med-VLMs) (Li et al., 2023a; Zhou et al., 2023b,a; Selbergren et al., 2025; Nath et al., 2025) have estab-

[†]Work done during internship at NEC Laboratories America

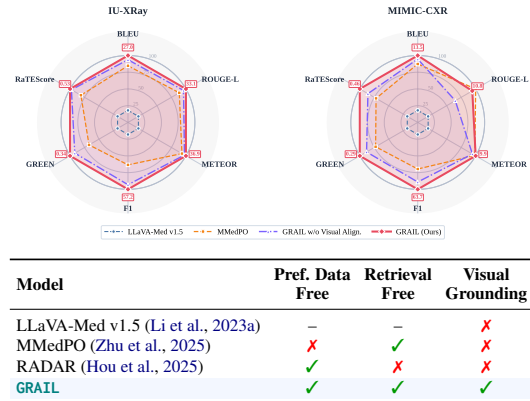


Figure 1: **Comparison with SOTA methods.** **GRAIL** improves generation quality (Papineni et al., 2002; Lin, 2004; Denkowski and Lavie, 2011) and clinical accuracy (Irvin et al., 2019; Ostmeier et al., 2024; Zhao et al., 2024) without curated preference data or retrieval-based augmentation. Unlike prior methods, **GRAIL** explicitly performs disease-specific visual grounding.

lished a dominant paradigm that leverages large-scale multimodal pretraining and clinical knowledge to generate reports directly from medical images. Despite their success, these models frequently generate findings that are clinically plausible but inconsistent with the input image, a phenomenon commonly referred to as *visual hallucination* (Fig. 2). This limitation arises because Med-VLMs often rely on statistical patterns learned from text-heavy pretraining corpora rather than performing genuine cross-modal reasoning (Leng et al., 2024; Huang et al., 2024; Liu et al., 2025; Amirloo et al., 2024; Cao et al., 2024; Zhou et al., 2024). The problem is further exacerbated in retrieval-augmented pipelines, where retrieved clinical statements may conflict with the visual evidence in the input X-ray (Xia et al., 2024b; Hou et al., 2025; Chen et al., 2024a). These limitations motivate a central question: *Can we design a framework that generates clinically accurate reports that remain visually grounded in the input image?*

Existing efforts to mitigate visual hallucinations

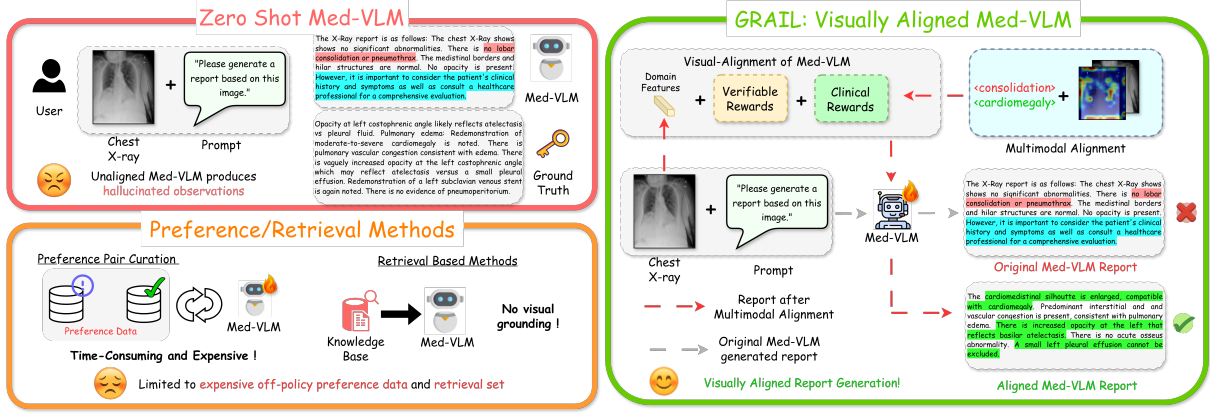


Figure 2: **Overview:** We introduce **GRAIL**, a multimodal alignment framework for reducing visual hallucinations in Med-VLMs (Li et al., 2023a). **Left:** Vanilla Med-VLM generates image-misaligned findings in red and generic non-medical text in cyan. Existing methods rely on costly task-specific preference data or retrieved reports, without directly addressing multimodal alignment. **Right:** **GRAIL** uses visual reasoning with multimodal rewards to generate clinically informed, image-grounded reports, marked in green.

suffer from two key limitations. First, **Data Reliance:** methods based on preference optimization (Chen et al., 2024c; Cui et al., 2024; Ouyang et al., 2022; Zhu et al., 2025) depend heavily on curated preference pairs, which are often constructed using heuristic or weakly informed criteria with limited clinical grounding. Moreover, preference optimization is typically performed entirely in the textual space, without verifying whether generated reports are actually supported by the provided image. As a result, these approaches fail to address the inherent gap between textual reports and visual evidence. Second, **Visual Misalignment:** retrieval-augmented approaches (Jin et al., 2021; Xia et al., 2024b; Sun et al., 2025; Hou et al., 2025) inject external clinical context, but retrieved reports may contain findings absent from the current X-ray (Fig. 2), introducing an additional source of hallucination. These methods also require domain-specific retrieval systems and additional computational overhead. More fundamentally, the semantic gap between retrieved reports and the input image cannot be resolved solely through retrieval quality. Consequently, existing paradigms fail to directly align generated reports with image evidence without relying on curated preference data, retrieval systems, or additional annotations (Fig. 1). In contrast, our goal is to directly optimize report generation using image-report consistency rewards, encouraging alignment with disease-relevant visual information without external guidance.

To this end, we propose **GRAIL: Grounded Visual Alignment of Medical Vision-Language Models for Radiology Report Generation**. Specifically, **GRAIL**

employs a two-stage multimodal alignment framework for radiology report generation. In the first stage, **Clinically Informed Textual Reasoning**, the model is optimized using verifiable language rewards with periodic clinical guidance to improve report structure and medical terminology. In the second stage, **Expert-Guided Visual Reasoning**, **GRAIL** leverages a frozen multi-label expert model to rank image-report consistency between the X-ray and generated candidate reports. These scores are converted into **rank-normalized visual advantages** that provide relative supervision across generated candidates, directly steering policy optimization toward visually grounded report generation without curated preference datasets, retrieval systems, or additional annotations.

In summary, our main contributions are:

1. We introduce a expert-guided visual alignment strategy that converts image-report similarity scores from a frozen domain expert into **rank-normalized visual advantages** for policy optimization, enabling effective visual grounding through image-report consistency optimization.
2. We propose **GRAIL**, a multimodal post-alignment framework that directly aligns generated radiology reports with disease-relevant image evidence through successive textual and image-conditioned reasoning stages, reducing hallucinations while improving image-report consistency.
3. We conduct extensive evaluations on IU-XRay (Demner-Fushman et al., 2015) and

MIMIC-CXR (Johnson et al., 2020), jointly assessing report generation quality, clinical accuracy, and visual grounding. GRAIL consistently shows strong relative gains over post-training baselines by up to **13%** in report generation quality and **15%** in clinical accuracy, while reducing hallucinated findings and improving image-report grounding consistency.

2 Related Works

Radiology Report Generation (RRG). Radiology report generation aims to describe medical images through detailed “Findings” and “Impression” sections, with prior works exploring this task (Huang et al., 2023; Jin et al., 2024; Li et al., 2023b). Existing RRG methods generally fall into two categories. The first improves encoder-decoder architectures for more consistent and clinically accurate reports, including LSTM-based hierarchical models (Hochreiter and Schmidhuber, 1997; Tanida et al., 2023), dynamic graph-based contrastive learning (Li et al., 2023b), and supervised multimodal pre-training (Wang et al., 2023; Park et al., 2025). The second uses retrieval-augmented generation, where report retrievers are trained using clinical similarity, domain knowledge, or disease overlap (Li et al., 2024b; Dong et al., 2025; Xia et al., 2024a; Chu et al., 2025; Hou et al., 2025). However, these methods often rely on paired multi-modal data, retrieved data during training or inference, or may generate clinical terminology misaligned with the X-ray, limiting clinical effectiveness.

Reinforcement Learning in LLMs/VLMs. Beyond Supervised Fine-Tuning (SFT), reinforcement learning (RL)-based alignment has become increasingly popular for improving robustness, aligning model behavior with human preferences, and reducing hallucinations (Yu et al., 2025; Zhou et al., 2024; Wang et al., 2025). In the medical domain, MMedPO (Zhu et al., 2025) adopts Direct Preference Optimization (DPO) with curated preference data to mitigate hallucinations in Med-VLMs. However, such preference data is expensive to collect and may not generalize beyond curated examples. Moreover, preference optimization is typically performed in the textual space without explicitly enforcing alignment with image evidence. Recent work has also explored reinforcement learning for improved radiology report generation (Gunder-

sen et al., 2026). Unlike these approaches, GRAIL performs multimodal post-alignment using image-conditioned rewards derived from image-report consistency, encouraging visually grounded report generation.

Medical Reasoning in Vision Language Models. Following CLIP (Radford et al., 2021), several medical VLMs have emerged, including Med-Flamingo (Moor et al., 2023), BioMedGPT (Zhang et al., 2024a), and HuatuoGPT-Vision (Chen et al., 2024c). Related models such as CheXzero (Tiu et al., 2022), Med-CLIP (Wang et al., 2022), and BioMedCLIP (Zhang et al., 2023) improve image-report alignment, while BioVIL (Boecking et al., 2022) and BioVIL-T (Bannur et al., 2023) strengthen medical text representations. More recent methods incorporate spatial relations (Wu et al., 2023; Jain et al., 2021), expert annotations, or external tools (Li et al., 2024a; Nath et al., 2025), while alignment-based approaches such as RadAlign (Gu et al., 2025) and MCA-RG (Xing et al., 2025) improve RRG through concept-level reasoning and multimodal consistency. Despite this progress, many methods rely on large labeled corpora, retrieval resources, preference annotations, or tool supervision, with limited emphasis on directly aligning model reasoning to the X-ray. In contrast, GRAIL improves visual grounding through image-conditioned post-alignment using image-report consistency rewards, without curated preference data or retrieval-based supervision.

3 Methodology

We describe the GRAIL framework below in two parts: (1) **Instruction-Following Warmup**, which stabilizes the Med-VLM for structured report generation and prevents degenerate outputs during subsequent optimization, and (2) **Multi-Stage Rewards for Clinical and Visual Alignment**, which progressively guides the policy toward clinically precise and visually grounded reports.

3.1 Instruction-Following Warmup

We perform a supervised fine-tuning (SFT) phase to enhance the instruction following capabilities of the Med-VLM $\mathcal{M}(\cdot)$ for radiology report generation. SFT is carried out on a subset of the training data to improve formatted generation and stabilize optimization (Sun et al., 2024; Zhu et al., 2025). Given the SFT dataset $\mathcal{D}_{\text{sft}} = \{(\mathcal{I}, \mathcal{R}_{\text{sft}})\}$ of chest X-rays $\mathcal{I}$, prompts $\mathcal{P}$, and reports $\mathcal{R}_{\text{sft}}$, we update

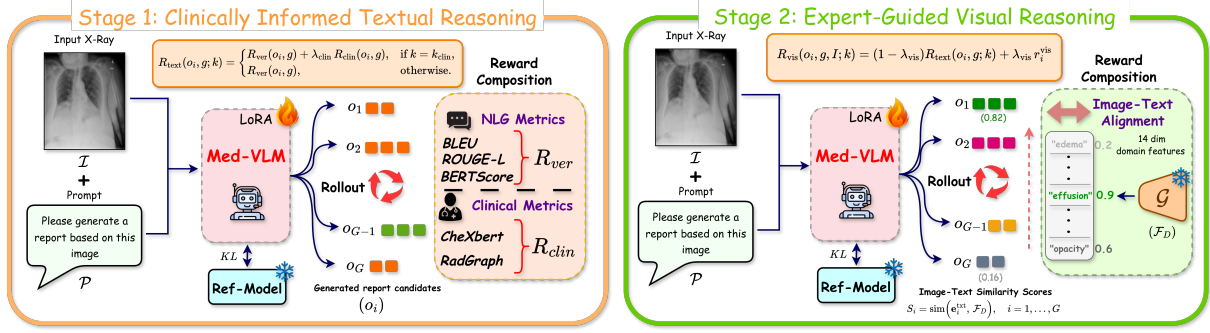


Figure 3: **Workflow of GRAIL.** **Left:** In Stage-1, GRAIL is trained with verifiable textual rewards R_{ver} with periodic clinical guidance R_{clin} every k_{clin} policy steps to familiarize the base Med-VLM $\mathcal{M}(\cdot)$ with radiology reasoning patterns. **Right:** In Stage-2, GRAIL is further optimized with rewards aligned using *image-text similarity scores* computed from embeddings obtained utilizing the multi-label domain expert model $\mathcal{G}$, shifting the model’s reasoning focus from textual knowledge, to grounded, region-wise disease reasoning capabilities in a self-supervised fashion. We also utilize format rewards to maintain structure of generated reports using `<report>` tags.

the parameters of $\mathcal{M}(\cdot)$ by minimizing the negative log-likelihood w.r.t. ground-truth tokens,

$$\mathcal{L}_{SFT}(\theta) = \mathbb{E}_{(\mathcal{I}, \mathcal{R}_{sft}) \sim \mathcal{D}} \left[- \sum_{t=1}^T \log \pi_{\theta}(r_t^{sft} | \mathcal{I}, r_{<t}^{sft}) \right] \quad (1)$$

where $\mathcal{R}_{sft} = (r_1^{sft}, \dots, r_T^{sft})$ denotes the tokenized report, and π_{θ} is the conditional distribution induced by $\mathcal{M}(\cdot)$. This SFT warmup teaches the model to respect the report structure and produce structurally well-formed reports conditioned on the input image $\mathcal{I}$ and prompt $\mathcal{P}$. In practice, we find that starting from an SFT checkpoint leads to faster convergence and avoids degenerate “cold-start” tokens. This SFT fine-tuned policy π_{θ} is then post-trained using our multistage rewards below.

3.2 Multi-Stage Rewards for Clinical and Visual Alignment

To guide the policy toward radiology knowledge and visually grounded outputs, we design a multi-tier reward system. The first stage focuses on semantic alignment, format adherence, and clinical accuracy using textual rewards, whereas the second phase emphasizes image-text alignment between the generated report and input image, injecting disease-relevant visual reasoning.

Clinically Informed Textual Reasoning. Here, we adapt the Med-VLM using *verifiable* text rewards that can be computed directly from the ground truth report. We combine NLG Metrics (Papineni et al., 2002; Lin, 2004), which emphasize report token similarity, with semantics (Zhang et al., 2019), which measures semantic coherence while being less sensitive to generic template text. At each alignment step, $\mathcal{M}(\cdot)$ rolls out a set of

candidate reports $\{o_i\}_{i=1}^G$ from the current policy π_{θ} . For each candidate report o_i and ground truth report g , the verifiable text reward is given as,

$$R_{ver}(o_i, g) = \lambda_{lex} [\text{BLEU}_4(o_i, g) + \text{ROUGE}_L(o_i, g)] + (1 - \lambda_{lex}) \text{BERTScore}(o_i, g). \quad (2)$$

where weight hyperparameter $\lambda_{lex} \geq 0$ controls the tradeoff between lexical fidelity and semantics.

To encourage clinically correct disease terms, we use the micro-averaged F1 scores from CheXbert and RadGraph entities as *periodic* rewards. Let $s_{cx}, s_{rg} \in [0, 1]$ denote normalized F1 scores between labels (or triplets) and entities extracted from o_i and g . We define the clinical reward as,

$$R_{clin}(o_i, g) = s_{cx}(o_i, g) + s_{rg}(o_i, g), \quad (3)$$

Rather than applying clinical rewards at every policy step, which destabilizes the optimization, we inject them only every k_{clin} updates. Let k denote each rollout from the global policy update step. Then, the Stage-1 scalar reward used to compute group relative advantages is combined as,

$$R_{text}(o_i, g; k) = \begin{cases} R_{ver}(o_i, g) + \lambda_{clin} R_{clin}(o_i, g), & \text{if } k = k_{clin}, \\ R_{ver}(o_i, g), & \text{otherwise,} \end{cases} \quad (4)$$

where $\lambda_{clin} \geq 0$ controls the strength of the periodic clinical contribution. This design provides dense, verifiable supervision at every step while periodically nudging the policy toward clinically faithful language without allowing natural language metrics to entirely dominate the training process.

Expert-Guided Visual Reasoning. In Stage-2, we introduce image-text similarity scores that explicitly tie each candidate report with the provided

Table 1: We compare **GRAIL** with the base Med-VLM (Li et al., 2023a), SFT, DPO (Rafailov et al., 2023), GRPO (Shao et al., 2024), and the post-trained RRG baseline MMedPO (Zhu et al., 2025) on generation quality (Pappinen et al., 2002; Lin, 2004; Denkowski and Lavie, 2011) and clinical accuracy (Smit et al., 2020; Ostmeier et al., 2024; Zhao et al., 2024). The last two rows show the effect of visual alignment, with **GRAIL** achieving the best overall performance on both benchmarks. Best results are in **bold** and second best in *italics*.

		Generation Quality			Clinical Accuracy			
		Model	BLEU $\uparrow$	ROUGE-L $\uparrow$	METEOR $\uparrow$	F1-Score $\uparrow$	GREEN $\uparrow$	RaTEScore $\uparrow$
IU-Xray	Baselines	LLaVA-Med v1.5	8.55	7.58	4.56	40.6	0.14	0.18
		+ SFT	14.56	10.31	10.95	38.8	0.11	0.16
		+ DPO (Rafailov et al., 2023)	16.09	12.95	17.13	40.5	0.12	0.16
		+ GRPO (Shao et al., 2024)	17.90	14.22	19.60	41.4	0.14	0.19
	Advancing	MMedPO (Zhu et al., 2025)	23.57	29.52	34.16	49.8	0.26	0.45
		GRAIL (w/o Visual Alignment)	25.61	31.80	35.90	55.8	0.32	0.52
		GRAIL	27.03	33.10	36.88	57.2	0.34	0.53
MIMIC-CXR	Baselines	LLaVA-Med v1.5	3.85	9.42	5.62	50.2	0.08	0.11
		+ SFT	11.11	9.38	7.71	49.8	0.10	0.14
		+ DPO (Rafailov et al., 2023)	11.20	9.45	7.81	50.4	0.11	0.16
		+ GRPO (Shao et al., 2024)	12.26	10.58	9.22	51.2	0.13	0.19
	Advancing	MMedPO (Zhu et al., 2025)	12.02	10.96	9.88	58.7	0.22	0.34
		GRAIL (w/o Visual Alignment)	12.90	10.30	9.65	61.9	0.26	0.40
		GRAIL	13.52	10.85	9.95	63.7	0.29	0.46

X-Ray. For an image I , we obtain a continuous domain feature vector $\mathcal{F}_D = \mathcal{G}(\mathcal{I})$ from the frozen expert model, and for each sampled candidate o_i , we compute a text embedding $\mathbf{e}_i^{\text{txt}} = \mathcal{G}_{\text{txt}}(o_i)$ in the same space as above. We then define the image-text similarity score for each candidate i ,

$$S_i = \text{sim}(\mathbf{e}_i^{\text{txt}}, \mathcal{F}_D), \quad i = 1, \dots, G \quad (5)$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity, and a larger S_i indicates a stronger agreement between the report and disease-relevant visual features.

To obtain a scale-free visual signal, we convert these scores into rank-normalized visual rewards. Let $\text{rank}(S_i)$ return 1 for the largest similarity and z for the smallest; we define,

$$r_i^{\text{vis}} = 1 - \frac{\text{rank}(S_i) - 1}{z - 1}, \quad (6)$$

so that $r_i^{\text{vis}} \in [0, 1]$ assigns the highest score to the most image-consistent candidates in each group.

We retain the scalar textual reward $R_{\text{text}}(o_i, g; k)$ from Eq. (4), and then define the composite visual-alignment reward,

$$R_{\text{vis}}(o_i, g, I; k) = (1 - \lambda_{\text{vis}}) R_{\text{text}}(o_i, g; k) + \lambda_{\text{vis}} r_i^{\text{vis}} \quad (7)$$

where $\lambda_{\text{vis}} \in [0, 1]$ acts as the mixing coefficient that controls the strength of visual grounding. The

group-relative advantages are then obtained by rank-normalization over this composite reward,

$$A_i^{\text{vis}} = \text{RankNorm}(\{R_{\text{vis}}(o_j, g, I; k)\}_{j=1}^G). \quad (8)$$

Thus, we simply replace Stage-1 advantages with A_i^{vis} , which biases the policy toward candidates that are both textually faithful and visually aligned. Therefore, the optimization objective is given as,

$$\mathcal{L}_{\text{GRAIL}}(\theta) = -\frac{1}{G} \sum_{i=1}^G \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \min(r_{i,t}^{\text{vis}}(\theta) A_{i,t}^{\text{vis}}, \text{clip}(r_{i,t}^{\text{vis}}(\theta), 1 - \epsilon, 1 + \epsilon) A_{i,t}^{\text{vis}}) - \beta D_{\text{KL}}[\pi_{\theta} \parallel \pi_{\text{ref}}], \quad (9)$$

where, $D_{\text{KL}}(\pi_{\theta} \parallel \pi_{\text{ref}})$ serves as a regularization term penalizing divergence of the current policy π_{θ} from the reference policy π_{ref} , with β controlling the regularization strength w.r.t. the policy.

Format Reward. To ensure that generated candidate reports follow the proper format, we impose a format reward (denoted as R_{format}) in both stages to contain the required tag. Specifically, the report must be contained within the <report> and </report> tags, respectively.

4 Experiments

4.1 Experimental Settings

Datasets and Evaluation Metrics. We benchmark **GRAIL** on IU-XRay (Demner-Fushman et al., 2015)

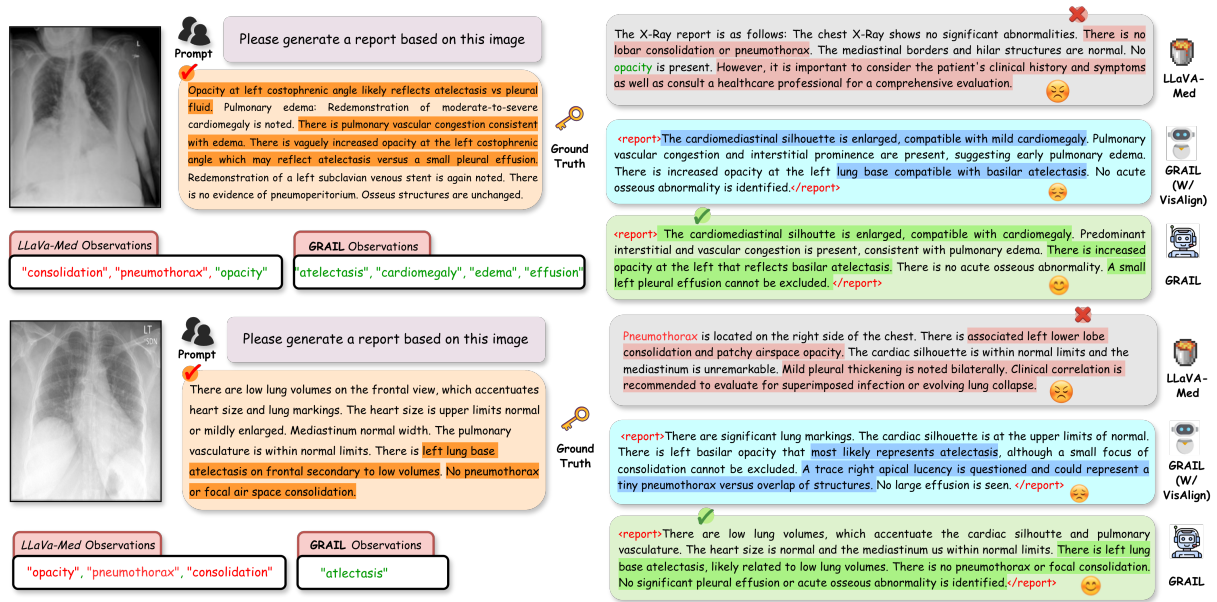


Figure 4: **Generated reports with GRAIL.** **First Case:** Med-VLM fails to detect subtle reasoning patterns needed for accurate diagnosis, resulting in hallucination of disease observations such as “consolidation” (top left). In contrast, GRAIL with verifiable guidance leads to significantly better reasoning and diagnosis capabilities (top right). With visual alignment, GRAIL’s enhanced multimodal reasoning leads to correct diagnosis of all observations, showcasing the need for awareness of disease-relevant portions and focused attention. The Med-VLM again hallucinates diseases due to insufficient visual focus (bottom left), in comparison GRAIL reduces hallucinations and generates reports faithful to the X-Ray exhibiting better anatomical and semantic awareness. (bottom right).

and MIMIC-CXR (Johnson et al., 2020), evaluating both generation quality and clinical accuracy. Following prior RRG work (Chen et al., 2020, 2022; Wang et al., 2023), we report CheXbert-F1 over 14 disease labels (Smit et al., 2020), along with clinically oriented metrics GREEN (Ostmeier et al., 2024) and RaTEScore (Zhao et al., 2024). For report quality, we use standard NLG metrics: BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Denkowski and Lavie, 2011). For ROI overlap, we examine anatomical regions associated with target disease tokens and evaluate attention overlap. We provide additional attention map and grounding evaluation details in Appendix B.

Baselines and Implementation Details. We compare GRAIL against direct baselines built on the base Med-VLM, including SFT, DPO (Rafailov et al., 2023), GRPO (Shao et al., 2024), and GRPO without the proposed clinical and visual alignment. We also compare with MMedPO (Zhu et al., 2025), a recent DPO-based medical post-alignment method which only evaluates generation quality and not clinical accuracy. Following (Zhu et al., 2025), we use LLaVA-Med v1.5 (Li et al., 2023a) as the base Med-VLM and provide additional QwenVL-2.5-7B (Bai et al., 2025) results in the supplementary. Clinical accuracy is evaluated using 14

disease labels with F1, GREEN (Ostmeier et al., 2024), and RaTEScore (Zhao et al., 2024). We further compare against zero-shot proprietary general VLMs, including GPT-4o (Hurst et al., 2024), GPT-4.1 (Achiam et al., 2023), and Gemini-2.5 Pro (Comanici et al., 2025), as well as medical VLMs such as MedFlamingo (Moor et al., 2023), MAIRA-2 (Bannur et al., 2024), and HuatuoGPT-V-8B (Chen et al., 2024b). For GRPO training, we use LoRA fine-tuning (Hu et al., 2022) with batch size 8, learning rate $1e-6$, and train for 3 epochs. Additional details, prompts, and implementation settings are provided in Appendix A.

4.2 Comparison with SOTA Methods

Enhanced Clinical Reasoning and Generation Quality. We show qualitative report examples in Fig. 4. Compared to the baselines, GRAIL generates reports that better match the image evidence while preserving clinically meaningful terminology. In the top row, GRAIL correctly captures findings such as “atelectasis”, “cardiomegaly”, and “edema”, whereas the base Med-VLM either misses subtle abnormalities or produces unsupported observations. Although GRAIL without visual alignment improves over the base model by incorporating disease findings, it can still fail when disease

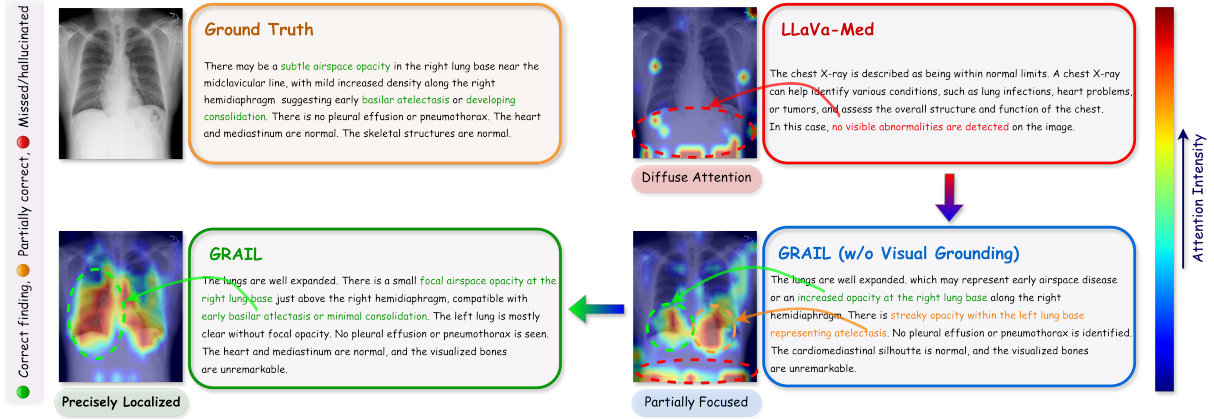


Figure 5: **Attention Map Visualization.** For the input X-Ray $\mathcal{I}$, we visualize the specific areas the Med-VLM (Li et al., 2023a) and **GRAIL** focuses on while generating the report. **Top Right, Both Cases:** is the attention map for the MedVLM (Li et al., 2023a) which misses the lung region due to lack of anatomical reasoning, leading to the generated report containing the inaccurate conclusion that no diseases can be detected.

localization is required. In contrast, the full **GRAIL** model produces disease-relevant observations that are more consistent with the input X-ray, highlighting the role of visual alignment in clinically grounded generation. More details in Appendix E. **Visual Reasoning and Grounding Abilities.** In Fig. 5, we visualize disease-token attention maps to analyze where each model attends while generating clinically relevant findings. The base Med-VLM exhibits diffuse attention over regions unrelated to the reported abnormality, leading to incorrect conclusions such as “no visible abnormalities.” **GRAIL** without visual alignment attends more to thoracic regions but remains partially mislocalized. After visual alignment, **GRAIL** produces more concentrated attention over disease-relevant regions, which aligns with the generated findings “opacity”, “atelectasis”, and “consolidation.” We use these attention maps as post-hoc diagnostic evidence, supported by the quantitative attention analysis in Table 5. Detailed analysis is in Appendix D and more grounding results in Appendix E.

Comparison w.r.t. Post-Trained Models. We evaluate **GRAIL** on two important tasks: generation quality and clinical accuracy on the two radiology report benchmarks IU-XRay (Demner-Fushman et al., 2015) and MIMIC-CXR (Johnson et al., 2020), outperforming state-of-the-art preference tuning method MMedPO (Zhu et al., 2025), DPO (Rafailov et al., 2023) and **GRAIL** without visual grounding in Table 1. We obtain a significant improvement over MMedPO on all three natural language metrics and **GRAIL** with visual

alignment. To assess clinical accuracy, we further report CheXpert-F1 accuracy and more clinically aligned metrics GREEN (Ostmeier et al., 2024) and RaTEScore (Zhao et al., 2024) in Table 1. On IU-Xray, **GRAIL** achieves the best F1 score (57.2%), clearly improving over LLaVA-Med (40.6%) and MMedPO (49.8%) by more than +10%, on GREEN and RaTEScore by more than +15% on average. As it can be seen from both benchmarks, the alignment step of our model is crucial for obtaining a large improvement. Overall, these results demonstrate that **GRAIL** not only produces more fluent and semantically aligned reports but also maintains strong clinical accuracy.

Comparison w.r.t. SOTA Multimodal LLMs and RRG Methods. We also compare our model against proprietary general LLMs such as GPT-4o (Hurst et al., 2024), GPT-4.1 (Achiam et al., 2023) and Gemini Pro (Comanici et al., 2025) and medical LLMs (Moor et al., 2023; Bannur et al., 2024; Chen et al., 2024c) in a zero-shot setting in Table 2. Our method outperforms GPT-4 series (Hurst et al., 2024), Gemini (Comanici et al., 2025) by +10% and SOTA medical MLLMs by an average margin of +12% in generation quality. Therefore, **GRAIL** consistently performs better than both general-purpose and medical multimodal LLMs, showing more anatomy of visual knowledge along with enhanced clinical accuracy despite the significantly larger pre-training knowledge corpus of these models on large amounts of data without the need of curated preference or retrieved data. We provide further results against specialized RRG methods and more baselines in Appendix C.

Table 2: We compare **GRAIL** against SOTA medical and general VLMs on both RRG benchmarks, reporting generation quality (Papineni et al., 2002; Lin, 2004; Denkowski and Lavie, 2011) and clinical accuracy using F1-Score (Smit et al., 2020). **GRAIL** consistently outperforms both general-purpose and medical VLMs. The best results are highlighted in **bold** and the second best in *italics*.

Type	Models	IU-Xray				MIMIC-CXR			
		BLEU $\uparrow$	ROUGE-L $\uparrow$	METEOR $\uparrow$	F1-Score $\uparrow$	BLEU $\uparrow$	ROUGE-L $\uparrow$	METEOR $\uparrow$	F1-Score $\uparrow$
General VLM	GPT-4o (Hurst et al., 2024)	15.32	18.80	11.50	28.6	12.10	17.20	10.30	30.6
	GPT-4.1 (Achiam et al., 2023)	16.48	19.95	12.34	39.2	12.94	17.88	10.92	35.5
	Gemini-2.5 Pro (Comanici et al., 2025)	17.26	21.10	13.05	–	13.42	18.76	11.48	27.2
Medical VLM	MedFlamingo (Moor et al., 2023)	14.85	24.10	22.44	–	11.76	21.95	19.84	–
	MAIRA-2 (Bannur et al., 2024)	20.22	<i>30.60</i>	<i>32.36</i>	52.9	18.40	<i>28.90</i>	<i>30.10</i>	58.1
	HuatuoGPT-V-8B (Chen et al., 2024c)	18.76	28.40	29.05	<i>54.2</i>	17.12	27.35	28.44	<i>59.6</i>
	GRAIL	27.03	33.10	36.88	57.2	24.30	31.50	34.20	63.7

4.3 Ablation Studies

Need for Multi-Stage Reward Guidance. Fig. 6(a) ablates the Stage-2 mixing weight λ_{vis} in Eq. (7). We observe that BLEU-4, ROUGE-L, and CheXbert F1 consistently peak when $\lambda_{\text{vis}} \in [0.4, 0.6]$, indicating that neither purely textual nor purely visual optimization is sufficient. Lower weightage under-utilizes visual informativeness and reduces Stage-2 to Stage-1 behavior, whereas large values collapse training towards only disease terms. This confirms that gains of **GRAIL** arise from a balanced combination of textual faithfulness and visual grounding, reflected in improvements reported in Table 3.

Clinical/Visual Intervention. Fig. 6(b) studies the periodicity of clinical reward injection through k_{clin} . Very frequent supervision (small k_{clin}) destabilizes optimization and biases the policy toward short, overly conservative clinical phrases, while overly sparse supervision weakens the clinical signal and under-utilizes Stage-1 guidance. A moderate reward period yields the best trade-off, as also shown in Table 4. Fig. 6 justifies this design by showing clear performance jumps both at the onset of periodic clinical guidance and again at the beginning of Stage-2, showing clinical intervention improves report faithfulness, while visual intervention further strengthens grounded generation. More ablations/details in Appendix D.

Grounding Analysis. We quantify disease-token attention localization in Table 5. Compared to the base Med-VLM and **GRAIL** without visual alignment, **GRAIL** assigns more attention mass to clinically relevant regions, improves Hit@1 localization, and produces lower-entropy attention maps, suggesting more image-conditioned reasoning. These localization results suggest that visual

Table 3: **Stage-II visual sensitivity.** Effect of λ_{vis} (Eq. (7)) on BLEU and F1 Score. $\text{--}\times$ denotes non-convergence.

λ_{vis}	IU-Xray		MIMIC-CXR	
	BLEU $\uparrow$	F1 $\uparrow$	BLEU $\uparrow$	F1 $\uparrow$
0.0	25.61	55.8	12.90	61.9
0.5	27.03	57.2	13.52	63.7
1.0	$\text{--}\times$	$\text{--}\times$	$\text{--}\times$	$\text{--}\times$

Table 4: **Periodic clinical reward sensitivity.** Effect of k_{clin} (Eq. (4)) on BLEU and F1 Score. $\text{--}\times$ denotes non-convergence.

k_{clin}	IU-Xray		MIMIC-CXR	
	BLEU $\uparrow$	F1 $\uparrow$	BLEU $\uparrow$	F1 $\uparrow$
dense	$\text{--}\times$	$\text{--}\times$	$\text{--}\times$	$\text{--}\times$
moderate	27.03	57.2	13.52	63.7
rare	26.30	55.1	12.14	62.4

Model	ROI Mass $\uparrow$	Hit@1 $\uparrow$	Entropy $\downarrow$
LLaVA-Med v1.5	0.11	28.4	5.81
GRAIL w/o Visual Align.	0.40	52.7	3.46
GRAIL	0.51	64.9	2.18

Table 5: **Quantitative attention analysis.** ROI Mass and Hit@1 measure whether disease-token attention localizes specifically within clinically plausible regions in the X-Ray, thus lowering entropy.

alignment improves not only where the model attends, but also whether the generated disease findings are supported by the provided X-ray.

Role and Reliability of Expert Ranking. Our Stage-2 reward uses the expert similarity score as a proxy for distinguishing better report candidates. Since the text-side feature extractor is a frozen CheXbert labeler, it is independent of **GRAIL**’s training paradigm. After SFT warmup and Stage-1 alignment, we observe that candidates with higher image-report disease consistency within the same rollout group are generally preferable. This motivates rank-normalizing similarity scores over candidates, making the reward depend only on relatively better reports within each rollout.

To verify whether this ranking is informative, Table 6 compares the quality of expert-ranked candidates across five randomly sampled rollout groups on MIMIC-CXR. Rank-1 candidates outperform Rank-8 candidates by approximately 24% in BLEU

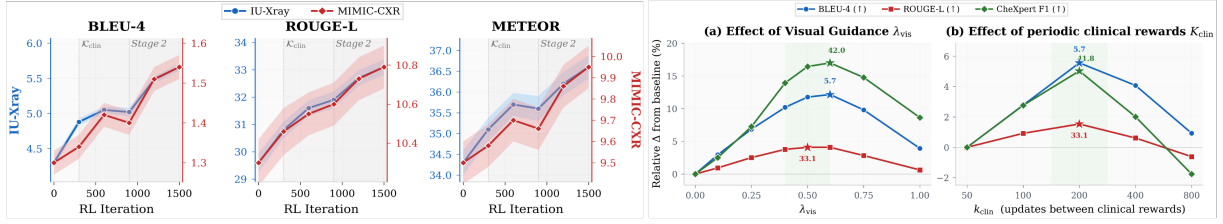


Figure 6: **Evolution of clinical and visual guidance.** *Left:* **GRAIL** performs best with an intermediate visual reward weight, $\lambda_{vis} \sim 0.5$. *Right:* Periodic clinical reward injection with R_{clin} every k_{clin} steps improves performance compared to dense per-step guidance, which destabilizes optimization - dense injection causes non-convergence, moderate intervals give the best optimisation-vs-correctness trade-off.

Expert Rank	BLEU $\uparrow$	RaTEScore $\uparrow$
Rank-1	14.8 ± 0.4	0.48 ± 0.02
Rank-4	13.3 ± 0.5	0.42 ± 0.02
Rank-8	11.9 ± 0.6	0.39 ± 0.03

Table 6: **Reliability of expert-guided candidate ranking on MIMIC-CXR.** We compare reports at different ranks induced by the Stage-2 image-report similarity score across five randomly sampled rollout groups. Higher-ranked candidates consistently exhibit better report quality under both BLEU and RaTEScore.

and 23% in RaTEScore, respectively, indicating that the expert’s relative ranking separates stronger from weaker report candidates and supports its use for computing relative advantages. Our rank-normalization of the expert’s image-report similarity into a group-relative advantage also decouples the reward from the expert’s absolute scale and monotone transforms. Importantly, Stage-2 is not intended to replace textual or clinical guidance; rather, the two stages provide complementary supervision, as reflected by the balanced reward combination in Fig. 6(a).

5 Conclusion

We introduced **GRAIL**, a multimodal alignment framework that improves the clinical and visual reasoning capabilities of existing Medical VLMs. By combining clinically informed textual reasoning with expert-guided visual reasoning, **GRAIL** encourages reports that are both semantically coherent and better grounded in the input X-ray. Our staged design further enables visual consistency to be incorporated through relative expert-guided ranking, complementing textual and clinical supervision without requiring curated preference data or retrieval. Across multiple benchmarks **GRAIL** delivers consistent gains in report generation quality, clinical accuracy, and visual grounding. Our quantitative results and visual analyses further highlight

the importance of visual alignment for reducing unsupported findings and improving disease-relevant reasoning. Overall, **GRAIL** offers a practical step toward more clinically reliable and image-grounded radiology report generation systems.

6 Limitations

While **GRAIL** improves visual grounding and clinical accuracy for radiology report generation, our experiments focus on chest X-ray benchmarks, namely IU-XRay and MIMIC-CXR. Extending the framework to other imaging modalities (CT-Scan for example), temporal imaging of patients, and domain specific reporting styles remains an important direction for future work. In addition, our evaluation is conducted on offline datasets; future work should study how visually aligned report generation can interact with real clinical workflows with experts in-the-loop.

Ethics Statement: VLM based RRG systems may inherit biases from the underlying medical datasets, imaging protocols, patient demographics, and reporting styles used during training. Such biases can lead to uneven performance across clinical subgroups, institutions, or disease presentations, and hallucinated or unsupported findings may potentially affect some groups disproportionately. To mitigate these concerns, systems to take these into account should be developed and evaluated with diverse multi-institutional datasets and subgroup-level fairness analyses.

Acknowledgements

This work was partially supported by NSF grant OAC 2411453 and NEC Laboratories America, Inc. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF or NEC Laboratories America.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Elmira Amirloo, Jean-Philippe Fauconnier, Christoph Roesmann, Christian Kerl, Rinu Boney, Yusu Qian, Zirui Wang, Afshin Dehghan, Yinfei Yang, Zhe Gan, and 1 others. 2024. Understanding alignment in multimodal llms: A comprehensive study. *arXiv preprint arXiv:2407.02477*.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.
- Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, and 1 others. 2024. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*.
- Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, and 1 others. 2023. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15016–15027.
- Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, and 1 others. 2022. Making the most of text semantics to improve biomedical vision-language processing. In *European conference on computer vision*, pages 1–21. Springer.
- Jianjian Cao, Peng Ye, Shengze Li, Chong Yu, Yansong Tang, Jiwen Lu, and Tao Chen. 2024. Madtp: Multimodal alignment-guided dynamic token pruning for accelerating vision-language transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15710–15719.
- 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*.
- Jiawei Chen, Dingkan 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, Zhenyang Cai, Ke Ji, Xiang Wan, and 1 others. 2024b. Towards injecting medical visual knowledge into multimodal llms at scale. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 7346–7370.
- Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, and 1 others. 2024c. Huatuoqpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*.
- Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2022. Cross-modal memory networks for radiology report generation. *arXiv preprint arXiv:2204.13258*.
- Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*.
- 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. 2024d. Chexagent: Towards a foundation model for chest x-ray interpretation. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*.
- Yun-Wei Chu, Kai Zhang, Christopher Malon, and Martin Renqiang Min. 2025. Reducing hallucinations of medical multimodal large language models with visual retrieval-augmented generation. In *Workshop on Large Language Models and Generative AI for Health at AAAI*.
- Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Naveen Sachdeva, Inderjit Dhillon, Marcel Blstein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.
- Hejie Cui, Lingjun Mao, Xin Liang, Jieyu Zhang, Hui Ren, Quanzheng Li, Xiang Li, and Carl Yang. 2024. Biomedical visual instruction tuning with clinician preference alignment. *Advances in neural information processing systems*, 37:96449–96467.
- 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.
- Michael Denkowski and Alon Lavie. 2011. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the sixth workshop on statistical machine translation*, pages 85–91.
- Guanting Dong, Yutao Zhu, Chenghao Zhang, Zechen Wang, Ji-Rong Wen, and Zhicheng Dou. 2025. Understand what llm needs: Dual preference alignment for retrieval-augmented generation. In *Proceedings of the ACM on Web Conference 2025*, pages 4206–4225.
- Difei Gu, Yunhe Gao, Yang Zhou, Mu Zhou, and Dimitris Metaxas. 2025. RadAlign: Advancing Radiology Report Generation with Vision-Language Concept Alignment. In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume LNCS 15966. Springer Nature Switzerland.
- Benjamin Gundersen, Nicolas Deperrois, Samuel Ruiperez-Campillo, Thomas M. Sutter, Julia E Vogt, Michael Moor, Farhad Nooralahzadeh, and Michael Krauthammer. 2026. RadVLM-GRPO: Enhancing chest x-ray report generation and visual grounding via reinforcement learning. In *Medical Imaging with Deep Learning*.

- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Wenjun Hou, Yi Cheng, Kaishuai Xu, Heng Li, Yan Hu, Wenjie Li, and Jiang Liu. 2025. RADAR: Enhancing radiology report generation with supplementary knowledge injection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.
- Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13418–13427.
- Zhongzhen Huang, Xiaofan Zhang, and Shaoting Zhang. 2023. Kiut: Knowledge-injected u-transformer for radiology report generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19809–19818.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Illcus, 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.
- Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, and 1 others. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.
- Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. 2024. Promptmrg: Diagnosis-driven prompts for medical report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2607–2615.
- Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2020. MIMIC-IV. *PhysioNet*. Available online at: <https://physionet.org/content/mimiciv/1.0/> (accessed August 23, 2021), pages 49–55.
- Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882.
- Binxu Li, Tiankai Yan, Yuanting Pan, Jie Luo, Ruiyang Ji, Jiayuan Ding, Zhe Xu, Shilong Liu, Haoyu Dong, Zihao Lin, and 1 others. 2024a. Mmedagent: Learning to use medical tools with multi-modal agent. *arXiv preprint arXiv:2407.02483*.
- Chunyu Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.
- Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. 2023b. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3334–3343.
- Xinze Li, Sen Mei, Zhenghao Liu, Yukun Yan, Shuo Wang, Shi Yu, Zheni Zeng, Hao Chen, Ge Yu, Zhiyuan Liu, and 1 others. 2024b. Rag-ddr: Optimizing retrieval-augmented generation using differentiable data rewards. *arXiv preprint arXiv:2410.13509*.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Bo Liu, Ke Zou, Li-Ming Zhan, Zexin Lu, Xiaoyu Dong, Yidi Chen, Chengqiang Xie, Jiannong Cao, Xiao-Ming Wu, and Huazhu Fu. 2025. Gemex: A large-scale, groundable, and explainable medical vqa benchmark for chest x-ray diagnosis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21310–21320.
- Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. Med-flamingo: a multi-modal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR.
- Vishwesh Nath, Wenqi Li, Dong Yang, Andriy Myronenko, Mingxin Zheng, Yao Lu, Zhijian Liu, Hongxu Yin, Yee Man Law, Yucheng Tang, and 1 others. 2025. Vila-m3: Enhancing vision-language models with medical expert knowledge. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14788–14798.
- Aaron Nicolson, Jason Dowling, and Bevan Koopman. 2023. Improving chest x-ray report generation by leveraging warm starting. *Artificial intelligence in medicine*, 144:102633.
- Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson Md, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, and 1 others. 2024. Green: Generative radiology report evaluation and error notation. In *Findings of the association for computational linguistics: EMNLP 2024*, pages 374–390.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Sang-Jun Park, Keun-Soo Heo, Dong-Hee Shin, Young-Han Son, Ji-Hye Oh, and Tae-Eui Kam. 2025. Dart: Disease-aware image-text alignment and self-correcting re-alignment for trustworthy radiology report generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15580–15589.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmlR.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- 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*.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.
- 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*.
- Liwen Sun, James Jialun Zhao, Wenjing Han, and Chenyan Xiong. 2025. Fact-aware multimodal retrieval augmentation for accurate medical radiology report generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 643–655.
- Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, and 1 others. 2024. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13088–13110.
- 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.
- Ekin Tiu, Ellie Talus, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. 2022. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature biomedical engineering*, 6(12):1399–1406.
- 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.
- Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2022, page 3876.
- Zitian Wang, Yue Liao, Kang Rong, Fengyun Rao, Yibo Yang, and Si Liu. 2025. Instruction-oriented preference alignment for enhancing multi-modal comprehension capability of mllms. *arXiv preprint arXiv:2503.20309*.
- 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.
- Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Medklip: Medical knowledge enhanced language-image pre-training in radiology. *arXiv preprint arXiv:2301.02228*.
- Joy T Wu, Nkechinyere N Agu, Ismini Lourentzou, Arjun Sharma, Joseph A Paguio, Jasper S Yao, Edward C Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, and 1 others. 2021. Chest imagenome dataset for clinical reasoning. *arXiv preprint arXiv:2108.00316*.
- Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. 2024a. Mmed-rag: Versatile multimodal rag system for medical vision language models. *arXiv preprint arXiv:2410.13085*.
- Peng Xia, Kangyu Zhu, Haoran Li, Hongtu Zhu, Yun Li, Gang Li, Linjun Zhang, and Huaxiu Yao. 2024b. Rule: Reliable multimodal rag for factuality in medical vision language models. *arXiv preprint arXiv:2407.05131*.
- Qilong Xing, Zikai Song, Youjia Zhang, Na Feng, Junqing Yu, and Wei Yang. 2025. MCA-RG: Enhancing LLMs with Medical Concept Alignment for Radiology Report Generation. In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume LNCS 15964. Springer Nature Switzerland.
- Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwan He, and 1 others. 2025. Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19985–19995.
- Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D Davison, Hui Ren, and 1 others. 2024a. A generalist vision-language foundation model for diverse biomedical tasks. *Nature Medicine*, 30(11):3129–3141.
- Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, and 1 others. 2023. Biomedclip:

a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*.

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

Xi Zhang, Zaiqiao Meng, Jake Lever, and Edmond SL Ho. 2024b. Gla-ai4biomed at rrg24: Visual instruction-tuned adaptation for radiology report generation. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 624–634.

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

Juexiao Zhou, Xiuying Chen, and Xin Gao. 2023a. Path to medical agi: Unify domain-specific medical llms with the lowest cost. *arXiv preprint arXiv:2306.10765*.

Juexiao Zhou, Xiaonan He, Liyuan Sun, Jiannan Xu, Xiuying Chen, Yuetan Chu, Longxi Zhou, Xingyu Liao, Bin Zhang, and Xin Gao. 2023b. Skingpt-4: an interactive dermatology diagnostic system with visual large language model. *arXiv preprint arXiv:2304.10691*.

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. 2024. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*.

Kangyu Zhu, Peng Xia, Yun Li, Hongtu Zhu, Sheng Wang, and Huaxiu Yao. 2025. [MMedPO: Aligning medical vision-language models with clinical-aware multimodal preference optimization](#). In *Forty-second International Conference on Machine Learning*.

Appendix

In this appendix, we present the following:

- Preliminaries, information on expert guidance, dataset, evaluation and implementation details for [GRAIL](#) (Sec. A, Sec. A.2, Sec. A.3).
- Grounding evaluation and discussion, including the evaluation protocol, grounding metrics, and visual grounding analysis (Sec. B).
- Additional quantitative results, including comparisons with radiology report generation baselines and extended result figures (Sec. C).
- Additional ablations on reward hyperparameters, GRPO candidate size, and model backbones (Sec. D).
- Additional qualitative examples of generated reports and visual grounding behavior (Sec. E).

A Preliminaries and Experimental Setup

A.1 Preliminaries

Problem Formulation. Given a 2D chest X-ray image $\mathcal{I} \in \mathbb{R}^{H \times W \times 3}$, a text prompt $\mathcal{P}$, and a dataset $\mathcal{D} = \{(\mathcal{I}, \mathcal{R}_{\text{gt}})\}$ of images and ground truth reports, the goal of radiology report generation is to produce a report $\mathcal{R}$ that is clinically accurate and aligned with the input image $\mathcal{I}$. We model this with a base Med-VLM $\mathcal{M}_\theta$, which induces a conditional policy over its parameters $\pi_\theta(\mathcal{R} \mid \mathcal{I})$ to generate $\mathcal{R}$ in an autoregressive manner. In this work, we adapt $\mathcal{M}_\theta$ in two stages: first, we refine the text generation policy using verifiable and clinical textual rewards; second, we further align the model with image-conditioned visual rewards derived from a frozen domain expert $\mathcal{G}$, encouraging reports whose disease findings are grounded in the underlying X-ray $\mathcal{I}$.

RRG Objective. We follow existing works and process reports in two parts: the FINDINGS and IMPRESSION sections, where the former contains the broad rationale behind the generated report and the latter contains the potential diseases present in the patient’s medical image. Formally, this process can be represented as follows: given the medical image $\mathcal{I}$, the Med-VLM $\mathcal{M}(\cdot)$ is tasked with interpreting the image, reasoning about it, and generating a descriptive report $\mathcal{R} = \{r_1, r_2, \dots, r_T\}$ where T represents the total length of the report

and $r_t \in \mathbb{V}$ represents the tokens in the vocabulary. The report $\mathcal{R}$ is auto-regressively generated as,

$$p(\mathcal{R} \mid \mathcal{I}) = \prod_{t=1}^T p(r_t \mid r_1, \dots, r_{t-1}, \mathcal{I}). \quad (10)$$

On-Policy Learning Framework. We utilize Group Relative Proximal Optimization (GRPO) to improve the reasoning abilities of medical-VLMs, we use it to improve the reasoning of the base Med-VLM for the task of radiology report generation. GRPO (Shao et al., 2024) is an algorithm similar to Proximal Policy Optimization (PPO) (Schulman et al., 2017) but with two major differences: (1) Unlike PPO, it does not require a trained reward model for calculating the advantages; instead, it relies on group-relative rewards using generalized advantage estimation (GAE), and (2) it can utilize verifiable rule-based outcomes as rewards, which can be as simple as the metrics used for benchmarking and evaluation. At each alignment step, the Med-VLM $\mathcal{M}(\cdot)$ rolls out a set of candidate responses $\{o_i\}_{i=1}^G$ from the current policy $\pi_{\theta_{\text{old}}}$ and each candidate output is assigned a scalar reward r_i at the two stages of the optimization process. The advantages A_i are calculated using the group-relative mean and standard-deviations of the batch G ,

$$A_i = \frac{r_i - \text{mean}(\{r_j\}_{j=1}^G)}{\text{std}(\{r_j\}_{j=1}^G)}. \quad (11)$$

Therefore, the GRPO objective is given as,

$$\mathcal{L}_{\text{GRPO}}(\theta) = -\frac{1}{G} \sum_{i=1}^G \frac{1}{|o_i|} \sum_{t=1}^{|o_i|} \min \left(r_{i,t}(\theta) \hat{A}_{i,t}, \right. \\ \left. \text{clip}(r_{i,t}(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_{i,t} \right) - \beta D_{\text{KL}}[\pi_\theta \parallel \pi_{\text{ref}}]. \quad (12)$$

with

$$r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q, o_{i,<t})}, \quad (13)$$

and where $D_{\text{KL}}[\pi_\theta \parallel \pi_{\text{ref}}]$ serves as a regularization term penalizing divergence from the reference policy π_{ref} , with β controlling its regularization strength of the policy.

A.2 Dataset Details

Datasets. We utilize the following two standard benchmarks for the task of radiology report generation in our paper.

IU-Xray (Demner-Fushman et al., 2015). The Indiana University Chest X-ray collection contains 7,470 images paired with 3,955 free-text reports;

each report is associated with either a frontal projection (AP/PA) or a lateral view. We follow the standard 7:1:2 train/val/test split used in prior work. **MIMIC-CXR (Johnson et al., 2020).** MIMIC-CXR is a large-scale clinical corpus with 227,835 radiology reports collected at Beth Israel Deaconess Medical Center. We use the official train/test partitions to ensure comparability with baselines (Zhu et al., 2025). Apart from the experimental settings described in Sec. 4, we also present the overall statistics of each dataset in Table 7 which was used for our work.

Table 7: **Dataset Statistics.** Summary for our pre-processing setup. We list the modality, no. of image-report pairs, the views available, splits and the evaluation labels used respectively. Note that these are before pre-processing of the respective datasets.

Dataset	Modality	#Images	#Reports	Proj.	Split	Observations
IU-Xray	CXR	7,470	3,955	AP/PA, Lateral	70/10/20	14 Labels
MIMIC-CXR	CXR	377,110	227,835	AP/PA, Lateral	Official	14 Labels

Benchmarks and split curation. We adopt the experimental settings and benchmarks as in MMedPO, evaluating the accuracy and generation quality on the test splits of the IU-XRay (Demner-Fushman et al., 2015) and MIMIC-CXR (Johnson et al., 2020). For SFT, we use the same image-ground truth report pairs as the baseline (Zhu et al., 2025) and in line with other visual instruction tuning baselines (Zhang et al., 2024b). Note that this is not a specifically curated set, just a subset of the original training data. For comparison against proprietary models, we utilize the MIMIC-CXR test split since it is available for MAIRA-2 (Bannur et al., 2024) for a fair comparison instead of the MMedPO test split data. For IU-XRay, we utilize the original protocol.

A.3 Implementation Details

We warm-up the base Med-VLM $\mathcal{M}(\cdot)$ with SFT and optimize with GRPO (Shao et al., 2024) using verifiable rewards R_{ver} (BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004) and BERTScore (Zhang et al., 2019)) and periodic clinical rewards R_{clin} (CheXbert (Smit et al., 2020) and RadGraph (Jain et al., 2021)). In the second stage, we devise R_{vis} to further align $\mathcal{M}(\cdot)$ with the input X-ray. In this section, we further list out more details on the candidate generation process, hyperparameters used and the training regimen for both the stages.

Rollout/Decoding Strategy: For generating multiple report candidates for each GRPO policy

step, we change the following parameters to alter the decoding process for each report candidate: temperature, top_p, repetition penalty, no-repeat-ngram and do_sample. By varying these parameters we induce enough diversity within the generated samples o_i . We use a batch-update size ($G = 8$) for all our experiments.

Stage-1: Clinically-guided textual alignment reward weights. We set $\lambda_{\text{lex}} = 0.10$ and $\lambda_{\text{sem}} = 0.20$ in Eq. 2, where the verifiable reward term splits as BLEU:ROUGE-L = 0.05:0.05. Further we enforce format constraints with a format reward R_{format} given by `<report>` and `</report>` tags. We keep a target token range in-between [140, 260] and a hard cap of 320 tokens. For the clinical rewards outlined in Eq. 3, we use $\lambda_{\text{cx}} = 0.30$ (CheXbert) and $\lambda_{\text{rg}} = 0.40$ (RadGraph), which are then mixed into Eq. 4 with $\lambda_{\text{clin}} = 0.8$ but periodically every k_{clin} policy updates: $k_{\text{clin}} = 300$ (IU-Xray) and $k_{\text{clin}} = 500$ (MIMIC-CXR). We normalize/clip these per-batch rewards to ensure smooth reward updates.

Stage-2: Visual Alignment reward weights. We form R_{vis} per Eq. 7 using rank-normalized similarity advantages with mixing $\lambda_{\text{vis}} = 0.5$ (tuned within the [0.4, 0.6] as illustrated in the peak of Fig. 6 in the main paper). The expert embeddings are frozen during this entire process; only the Med-VLM adapter layers and the projection blocks are trained in this stage.

Optimization, hyperparameters and training details. We use group size $G = 8$ candidates per policy step, with GRPO clip $\epsilon = 0.1$, KL regularization coefficient $\beta = 0.02$ with adaptive target 0.04 to a frozen reference, entropy coefficient 0.005, and rank-norm standardization, gradient clip 1.0, AdamW optimizer with $\text{lr} = 1 \times 10^{-6}$, $\beta = (0.9, 0.95)$, and no weight decay. As discussed in the decoding/rollout strategy section, we set fixed sampling settings for each rollout (for example): temperature 0.6, top-p=0.90, no-repeat-ngram = 3, $N = 8$ for best-of-N sampling, max new tokens = 96 (min = 16). Batch size = 8 (accumulation = 1); for a total of 3 epochs over the train split. All experiments were carried out in an NVIDIA 80GB A100 GPU.

LoRA Fine-Tuning. We adopt LoRA (Hu et al., 2022) with $r=16$, $\alpha=32$, dropout = 0.05, bias=none on a small set of modules while keeping the rest of backbone frozen. In Stage-1, adapters are applied to the text decoder’s self-attention and MLP blocks, yielding an adapted decoder $\mathcal{D}_{\theta}$ of

Model	Covered F1 $\uparrow$	Uncovered F1 $\uparrow$	Uncovered Drop $\downarrow$
GRAIL w/o Visual Align.	55.8	42.1	–
GRAIL	57.2	42.4	-0.3

Table 8: Coverage evaluation of expert-label dependence. “Covered” denotes disease labels within the frozen expert’s CheXpert top-5 label space, while “Uncovered” denotes labels outside this space. **GRAIL** improves covered-label performance without any drop in uncovered-label performance, indicating that the expert serves as a grounding prior.

the Med-VLM. In Stage-2, we additionally enable LoRA on the vision-language projection layer $\mathcal{F}_{\text{proj}}$, resulting in the stage-2 trainable parameters in $\{\mathcal{D}_\theta, \mathcal{F}_{\text{proj}}\}$.

Expert Model Architecture. We instantiate the frozen domain expert $\mathcal{G}$ as a CheXpert-style multi-label CXR disease classifier trained to predict the standard 14 CheXpert observations (Irvin et al., 2019). Given an input X-ray $\mathcal{I}$, we obtain a 14-dimensional disease logit/probability vector $F_D = \mathcal{G}_{\text{img}}(\mathcal{I})$ from the frozen CheXpert-style classifier. For each generated candidate o_i , we apply a standard CheXbert labeler to obtain a corresponding 14-dimensional observation vector $e_i^{\text{txt}} = \mathcal{G}_{\text{txt}}(o_i)$ in the same disease-label space. We then compute cosine similarity between these image- and text-side disease vectors and rank-normalize the resulting scores for Stage-2 rewards. Importantly, $\mathcal{G}$ is kept strictly frozen throughout training and is used only to score candidates within each rollout, ensuring that visual alignment is driven by a fixed, domain-features rather than additional supervision or task-specific reward model training.

Expert Model Discussion. We use the disease-relevant visual features of the frozen expert model $\mathcal{G}$ as an auxiliary signal to align the Med-VLM with the provided chest X-ray. In our setting, $\mathcal{G}$ operates over the standard CheXpert 14-label space, which captures a clinically meaningful set of chest X-ray observations. Importantly, we do not use $\mathcal{G}$ as an report-generation prior; instead, it provides anatomical priors that helps rank candidate reports according to their consistency with the input image. The label space of $\mathcal{G}$ covers the standard label set of CheXpert chest X-ray observation labels, and guides the output space of the Med-VLM. Since this reward is combined with textual and clinical rewards, the model can still generate findings beyond the expert’s commonly covered labels. In Table 8, **GRAIL** improves Covered F1 from 55.8 to 57.2, while Uncovered F1 is not degraded and

slightly improves from 42.1 to 42.4. The negative “uncovered” drop indicates that visual alignment does not affect findings outside the expert’s top-5 label space.

Evaluation Criteria/Protocol. As discussed in Sec. 4 onwards, BLEU-n, ROUGE-L and METEOR are reported for NLG quality, CheXbert-F1 scores, GREEN and RateScore are reported for assessing the clinical accuracy (Table 1) and we also perform visual analysis of enhanced attention focusing capabilities of **GRAIL** and zero-shot comparisons with SOTA MLLMs in Table 2 on both benchmarks (main paper) and include ablations on visual guidance and clinical intervention in Table 3 and Table 4 respectively.

Prompt Design and Diversity. We present the prompts used in our paper as follows: the Global Instruction Prompt applied to the Med-VLM (Li et al., 2023a) is shown in Fig. 7, and the Query Prompt for querying each X-ray is shown in the adjacent box. Different base prompts can lead to diverse candidates but in this work we keep the same prompt for evaluation fairness w.r.t. SOTA baselines. We change the decoding hyperparameters of the Med-VLM for diverse candidate generation instead.

B Grounding Evaluation and Discussion

B.1 Attention Map Visualization.

To analyze whether the generated reports rely on visually meaningful evidence, we visualize image-token attention maps from the Med-VLM during report generation. For each input X-ray, we use the same prompt, image pre-processing, decoding configuration, and generated report format across all compared models. After generation, we perform a forward pass with the generated report and extract the attention weights from the penultimate decoder layer, which provides a stable representation before the final token-prediction layer. For a generated report, let $\mathcal{T}$ denote the set of clinically relevant disease-bearing tokens, such as tokens corresponding to findings like opacity, atelectasis, consolidation, cardiomegaly, or pleural effusion. For multi-token findings, we average the attention maps across all subword tokens belonging to the same clinical phrase. We then select the attention weights from these textual tokens to the visual patch tokens produced by the image encoder and passed through the multimodal projection layer. The atten-

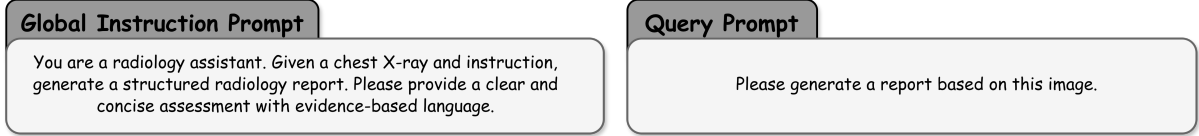


Figure 7: **Left:** The global instruction prompt given to the Med-VLM (Li et al., 2023a; Bai et al., 2025). **Right:** The query prompt used for each X-ray.

Model	Avg. In-ROI $\uparrow$	Avg. Out-ROI $\downarrow$	Higher In-ROI $\uparrow$
LLaVA-Med v1.5 (Li et al., 2023a)	0.118	0.091	36.2%
GRAIL w/o Visual Align.	0.181	0.104	55.8%
GRAIL	0.247	0.121	68.7%

Table 9: Attention-ROI overlap analysis. Avg. In-ROI and Avg. Out-ROI denote the average disease-token attention score inside and outside clinically relevant regions, respectively. Higher In-ROI reports the percentage of samples where the average in-region attention exceeds the average out-region attention.

tion weights are averaged across heads and normalized to obtain a single attention vector over image patches.

Formally, for a token $t \in \mathcal{T}$, let $A_{t \rightarrow v}^{(\ell, h)}$ denote the attention from token t to visual patch token v at layer ℓ and head h . We use the penultimate layer $\ell = L - 1$ and aggregate over heads as,

$$\alpha_t(v) = \frac{1}{H} \sum_{h=1}^H A_{t \rightarrow v}^{(L-1, h)}, \quad (14)$$

where H is the number of attention heads. For multi-token clinical phrases, we further average α_t over the corresponding tokens. The resulting patch-level attention vector is reshaped into the spatial patch grid of the vision encoder, bilinearly upsampled to the original X-ray resolution, and min-max normalized to $[0, 1]$ for visualization. Importantly, these attention maps are used only as post-hoc diagnostic visualizations and are not used as an additional training signal. We therefore interpret them as qualitative evidence of image-conditioned reasoning rather than as standalone proof of grounding. To make the comparison fair, all attention maps are generated using the same extraction procedure and visualization normalization across the base Med-VLM, GRAIL without visual alignment, and the full GRAIL model.

B.2 Attention-ROI overlap

Table 9 evaluates whether disease-token attention is preferentially inside the clinically relevant ROIs. For a held-out set of samples in the test split for ROI-overlap analysis, we define clinically relevant

regions using standard chest anatomical region definitions following Chest ImaGenome (Wu et al., 2021), mapping each disease-bearing token to its corresponding anatomical regions. We see that the base Med-VLM shows only a small gap between in-ROI and out-ROI attention, while GRAIL substantially increases the average in-ROI attention and the fraction of samples where in-ROI attention dominates. This supports the qualitative observation that visual alignment encourages the model to focus on disease-relevant anatomical regions during report generation. These localization results suggest that visual alignment improves not only where the model attends, but also whether the generated disease findings are supported by the input X-ray.

C Additional Quantitative Results

C.1 Comparison with Additional RRG Baselines

In Table 10 we compare GRAIL against specialized RRG methods which mostly use specialized encoder/decoder architectures trained in a fully-supervised manner on ground truth reports or use medical specific tools which may result in error propagation in the generated reports. Most of these specialized RRG systems are trained end-to-end for report generation on task-specific datasets, whereas GRAIL studies a different question: whether a Med-VLM can be self-supervised post-aligned to generate clinically accurate and visually grounded reports without curated preference data, retrieval databases, or additional visual annotations. GRAIL outperforms all these specialized RRG methods by a margin of **6%** on BERT-Score, **8%** on F1 accuracy of generated reports and by **4%** on GREEN score showing that directly incorporating image-conditioned visual rewards during alignment helps in generating semantically and clinically rich, visually grounded reports.

Comparison against retrieval baselines. We compare the clinical efficacy of GRAIL w.r.t. SOTA retrieval method RADAR (Hou et al., 2025) which

Method	Paradigm	Self-Sup. Visual Align.	BERT-Score $\uparrow$	F1-Score $\uparrow$	GREEN $\uparrow$
R2Gen (Chen et al., 2020)	Specialized RRG	$\times$	27.1	27.6	–
CheXagent (Chen et al., 2024d)	X-Ray Agent (with tools)	–	24.6	55.8	0.25
CheXpert-Plus (Chambon et al., 2024)	X-Ray Agent (with tools)	–	36.1	48.7	0.25
Med-CXRGen-F (Zhang et al., 2024b)	Specialized RRG	$\times$	22.1	41.2	0.25
RGRG (Tanida et al., 2023)	Specialized RRG	$\times$	34.8	24.5	0.27
RadFM (Wu et al., 2025)	Specialized RRG	$\times$	31.3	38.8	0.18
MAIRA-2 (Bannur et al., 2024)	Trained Encoder-Decoder arch.	$\times$	30.8	58.1	0.15
CvT2DistilGPT2 (Nicolson et al., 2023)	VLM (diff. backbone)	$\times$	33.1	54.2	0.16
GRAIL	Med-VLM Post-Alignment	$\checkmark$	41.6	63.7	0.29

Table 10: **Comparison against specialized RRG backbone models.** on MIMIC-CXR (Johnson et al., 2020). We report BERT-Score (Zhang et al., 2019), F1-Score (Irvin et al., 2019), and GREEN (Ostmeier et al., 2024) to compare both generation quality and clinical correctness. We highlight the differences of **GRAIL** relative to the broader RRG literature and it’s different setting: self-supervised visual post-alignment without task-specific retrieval databases or curated preference pairs.

curates specific retrieval data for each dataset and trains a domain specific retriever for that purpose alongwith utilizing the knowledge of an expert model. **GRAIL** still outperforms by an average of 3% across both benchmarks without using any additional data, highlighting the versatility of the rewards used in our method.

C.2 Classwise Clinical Performance

Classwise clinical performance. We investigate whether after performing the two-stage alignment over the base Med-VLM, whether **GRAIL** is still able to maintain high clinical performance across various disease classes or not - it shows consistent gains over the baseline (shown in Fig. 9) on both benchmarks indicating strong clinical knowledge with no forgetting, a common issue encountered with post-trained models.

C.3 Extended Result Figures

D Ablation and Sensitivity Analysis

D.1 Reward Hyperparameters λ_{lex} and λ_{clin}

Table 12(a,b) shows the role of λ_{lex} and λ_{clin} in the lexical and clinical guidance alignment process. For λ_{lex} , the best performance is obtained at 0.10, indicating that a relatively lower lexical component helps preserve report-level phrasing while retaining the semantics of the report provided by BERTScore. Similarly, $\lambda_{clin}=0.8$ gives the best balance between generation quality and clinical correctness; lower values slightly weaken the clinical signal, while larger values provide no consistent benefit.

D.2 Effect of GRPO Candidate Size G

Table 12(c) studies the effect of the GRPO candidate group size G . Increasing G from 4 to 8 provides a consistent but moderate improvement

across both datasets, suggesting that a larger candidate set improves group-relative ranking by exposing the policy to more diverse report candidates. We therefore use $G=8$ in all main experiments, which provides a better optimization signal without changing the model backbone or requiring additional supervision. Additional candidates per roll-out is significantly more expensive and leads to little/no improvements.

D.3 Backbone Ablation

Comparisons with different backbone model.

We also utilize QwenVL-2.5-7B (Bai et al., 2025) as an additional LVLM and align it with **GRAIL**’s two stage rewards (shown in Table 13). We obtain significant more gains over the baselines and SOTA models. An initial SFT phase becomes crucial to a significant performance boost as it becomes familiar with clinical terminology and patterns and **GRAIL** enforces both clinically-informed knowledge and visually aligned generation. **GRAIL** is highly versatile since it performs visual alignment in an entirely self-supervised manner (Stage-2) and uses commonly used NLG/Clinical metrics as rewards (Stage-1) which enhances the medical reasoning capabilities of any backbone VLM leading to enhanced gains across both generation quality and clinical accuracy by an average of 10% and 15% respectively.

E Additional Qualitative Examples

Model	Precision $\uparrow$	Recall $\uparrow$	F1 $\uparrow$
LLaVA-Med v1.5 (Li et al., 2023a)	41.2	38.4	40.6
MMedPO (Zhu et al., 2025)	47.2	48.6	49.8
RADAR (Hou et al., 2025)	51.2	52.6	52.4
GRAIL	56.6	57.8	57.2

Figure 8: CheXbert disease classification metrics (**F1**) on the **IU-Xray** dataset. Scores are computed using the CheXpert label set diseases; higher values indicate better alignment with individual disease labels.

Model	Precision $\uparrow$	Recall $\uparrow$	F1 $\uparrow$
LLaVA-Med v1.5 (Li et al., 2023a)	48.9	51.5	50.2
RADAR (Hou et al., 2025)	64.7	65.0	65.3
MMedPO (Zhu et al., 2025)	57.6	59.3	58.7
GRAIL	64.9	63.5	63.7

Table 11: **CheXbert disease classification metrics on MIMIC-CXR (Johnson et al., 2020)**. Precision/Recall/F1 with higher scores indicating better alignment with disease labels.

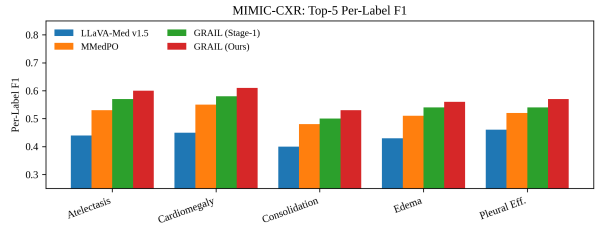
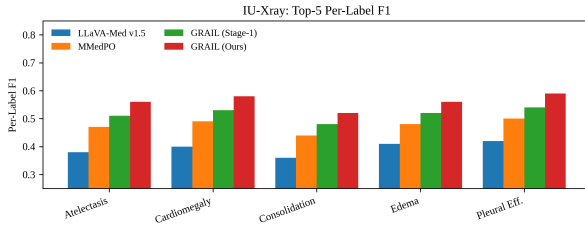


Figure 9: **Performance of GRAIL across top-5 disease labels.** We evaluate the classwise F-1 scores of **GRAIL** across 5 major disease labels: atelectasis, cardiomegaly, consolidation, edema and pleural effusion. We compare against the baseline Med-VLM and **GRAIL** Stage-1 and **GRAIL** respectively. As it can be seen, **GRAIL** exhibits superior clinical accuracy scores.

(a) Effect of λ_{lex}					(b) Effect of λ_{clin}					(c) Effect of candidate size G				
λ_{lex}	IU-Xray		MIMIC-CXR		λ_{clin}	IU-Xray		MIMIC-CXR		G	IU-Xray		MIMIC-CXR	
	BLEU $\uparrow$	F1 $\uparrow$	BLEU $\uparrow$	F1 $\uparrow$		BLEU $\uparrow$	F1 $\uparrow$	BLEU $\uparrow$	F1 $\uparrow$		BLEU $\uparrow$	F1 $\uparrow$	BLEU $\uparrow$	F1 $\uparrow$
0.05	26.78	56.8	13.34	63.2	0.5	26.52	56.2	13.08	62.7	4	26.41	56.6	13.10	62.8
0.10	27.03	57.2	13.52	63.7	0.8	27.03	57.2	13.52	63.7	8	27.03	57.2	13.52	63.7
0.20	26.91	57.0	13.41	63.3	1.0	26.88	57.0	13.39	63.4					

Table 12: Sensitivity analysis for the remaining reward and optimization hyper-parameters. We report BLEU and CheXbert-F1 on IU-Xray and MIMIC-CXR using LLaVA-Med v1.5 as the backbone. The selected setting, $\lambda_{\text{lex}}=0.10$, $\lambda_{\text{clin}}=0.8$, and $G=8$, provides the best overall trade-off while showing stable behavior across nearby configurations.

Table 13: We compare the performance of **GRAIL** against the following baselines: Qwen (Bai et al., 2025), SFT, naive DPO (Rafailov et al., 2023), GRPO (Shao et al., 2024) on the generation quality (Papineni et al., 2002; Lin, 2004; Denkowski and Lavie, 2011) of reports as well as clinical relevance and accuracy (Smit et al., 2020; Ostmeier et al., 2024; Zhao et al., 2024). Overall **GRAIL** achieves the best performance on both benchmarks. The best results are in **bold** and the second best in *italics*.

		<i>Generation Quality</i>			<i>Clinical Accuracy</i>		
	<i>Model</i>	BLEU ↑	ROUGE-L ↑	METEOR ↑	F1 ↑	GREEN ↑	RaTEScore ↑
IU-Xray	Qwen2.5-VL	6.20	6.40	3.70	29.5	0.06	0.09
	Baselines						
	+ SFT	11.80	9.10	8.60	31.0	0.07	0.11
	+ DPO (Rafailov et al., 2023)	12.90	9.90	9.40	32.2	0.08	0.12
	+ GRPO (Shao et al., 2024)	15.40	12.80	12.10	34.0	0.10	0.15
	GRAIL (w/o Visual Alignment)	23.10	28.10	31.20	48.0	0.25	0.40
MIMIC-CXR	Qwen2.5-VL	2.90	8.10	4.50	41.8	0.05	0.07
	Baselines						
	+ SFT	8.60	8.60	6.40	43.0	0.06	0.09
	+ DPO (Rafailov et al., 2023)	9.10	8.85	6.70	43.5	0.07	0.10
	+ GRPO (Shao et al., 2024)	10.30	9.70	7.80	44.6	0.08	0.12
	GRAIL (w/o Visual Alignment)	<i>12.10</i>	10.20	8.90	<i>57.0</i>	<i>0.20</i>	<i>0.32</i>
MIMIC-CXR	GRAIL	12.60	10.60	9.20	58.5	0.22	0.35

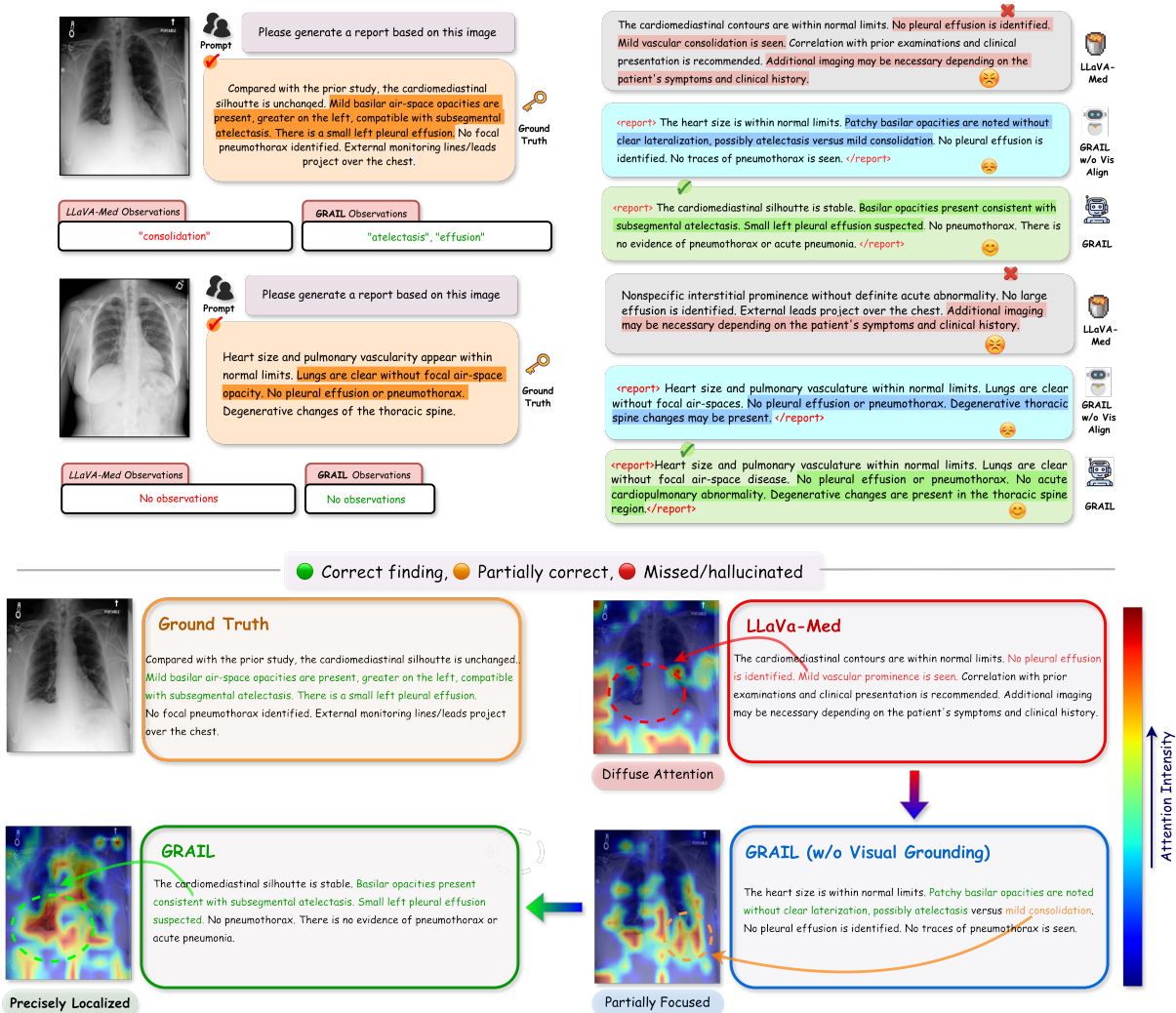


Figure 10: Report examples (failure modes of baselines and improvements with GRAIL). **Case 1 (LLaVA-Med v1.5 backbone):** The base Med-VLM produces a generic “no acute abnormality” style report and misses the subtle left basilar opacity, occasionally introducing unsupported or overly confident statements. Vanilla GRPO improves medical phrasing and overall fluency but can remain imprecise in localization (e.g., “bibasilar” wording) In contrast, **GRAIL** correctly grounds to the image by localizing the finding (left basilar atelectasis). **Case 2 (Qwen2.5-VL backbone):** tends to hallucinate broad pathologies (e.g., edema) and under-identifies abnormalities (left pleural effusion), reflecting weak image grounding. Vanilla GRPO reduces some generic language but often remains non-committal (“atelectasis vs consolidation”) and still underestimates effusion severity. **GRAIL** yields clinically informed and image-consistent conclusions, while maintaining concise, structured reporting.

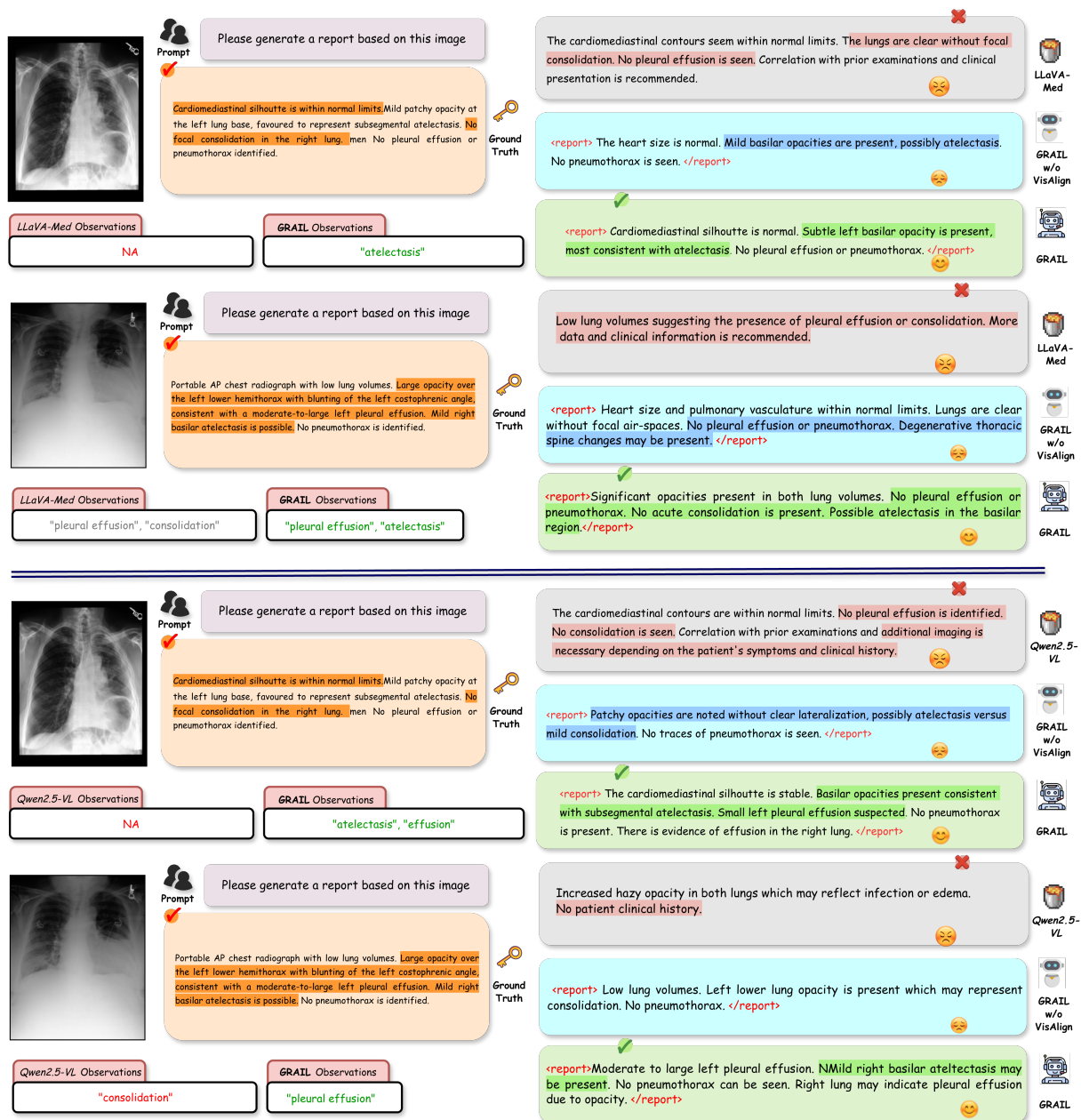


Figure 11: Report examples (failure modes of baselines and improvements with GRAIL). **Case 1, Upper Half (LLaVA-Med v1.5 backbone):** The base Med-VLM produces a generic “no acute abnormality” style report and misses the subtle left basilar opacity, occasionally introducing unsupported or overly confident statements. Vanilla GRPO improves medical phrasing and overall fluency but can remain imprecise in localization (e.g., “bibasilar” wording) In contrast, GRAIL correctly grounds to the image by localizing the finding (left basilar atelectasis). **Case 2, Lower Half (Qwen2.5-VL backbone):** tends to hallucinate broad pathologies (e.g., edema) and under-identifies abnormalities (left pleural effusion), reflecting weak image grounding. Vanilla GRPO reduces some generic language but often remains non-committal (“atelectasis vs consolidation”) and still underestimates effusion severity. GRAIL yields clinically informed and image-consistent conclusions, while maintaining concise, structured reporting.