Radiology Report Generation via Multi-View Contrastive Alignment and Asymmetric Clinical Reinforcement Learning

Anonymous ACL submission

Abstract

Radiology report generation demands not only linguistic fluency but clinical safety: a missed tension pneumothorax is qualitatively different from a missed cardiomegaly. Existing approaches fail on two fronts. First, they process only a single view of chest X-rays, losing the complementary spatial evidence that PA and lateral projections jointly provide. Second, they apply symmetric reward signals during alignment, treating false negatives for life-threatening pathologies identically to those for benign findings. We present a three-component framework that directly addresses these gaps. Specifically, a multi-view soft alignment module establishes patch-level anatomical correspondences between PA and lateral features via a cross-view spatial affinity mechanism, while a clinical soft-label contrastive loss replaces hard negatives with CheXpert-derived similarity weights to prevent false-negative penalisation of clinically similar patients. Furthermore, a confidence-aware adaptive RAG mechanism gates retrieval by a similarity threshold, withholding retrieved context entirely when below it to reduce retrieval-induced vulnerability by 65% under adversarial injection. Finally, an asymmetric clinical GRPO strategy assigns $1.5\times$ the gradient weight to false negatives for critical pathologies compared to false positives, producing the largest per-class CE-F1 gain specifically for pneumothorax ($+28.1\%$ from the retrieval stage to the full framework). On MIMIC-CXR, the complete framework achieves CE-F1 of 0.505 and ROUGE-L of 0.382, surpassing the strongest prior baseline by $+6.8\%$ CE-F1 and $+31.3\%$ ROUGE-L.

1 Introduction

Radiology report generation (RRG) sits at the intersection of clinical safety and natural language generation: a single missed finding can delay life-saving treatment, while a hallucinated one can trigger unnecessary invasive procedures (Zhang et al., 2024). Demand for diagnostic imaging continues to outpace radiologist capacity (Rimmer, 2017; Sloan et al., 2024), making automated report generation a practical necessity. Large-scale datasets such as MIMIC-CXR (Johnson et al., 2019) have catalysed deep learning approaches, yet two fundamental challenges remain largely unresolved.

Challenge 1: View incompleteness. Chest X-ray diagnosis routinely integrates PA and lateral projections, whose orthogonal geometry provides complementary spatial evidence: the PA view exposes cardiac silhouette and lung fields, while the lateral view resolves anterior-posterior lesion depth and posterior costophrenic angles. Most existing models process only a single view; the few that combine views (merc et al., 2024; Sîrbu et al., 2025) use simple concatenation or averaging, without modelling anatomical correspondence between projections. Retrieval-augmented approaches (Hou et al., 2025) further depend on retrieval precision, which in turn depends on encoder quality—yet CLIP-based medical encoders (Wang et al., 2022) are trained with hard-negative objectives that penalise semantically similar cross-patient pairs as negatives, distorting the retrieval space.

Challenge 2: Symmetric alignment objectives. Reinforcement learning approaches have improved factual fidelity (Miura et al., 2021; Delbrouck et al., 2022; zhou et al., 2024), but all existing methods apply *symmetric* reward signals: a missed pneumothorax and a missed cardiomegaly incur equal gradient penalties, despite their vastly different clinical consequences.

Motivating example. Consider a patient with right-sided pleural effusion: the PA view shows only subtle costophrenic blunting, while the lateral view reveals unambiguous posterior angle obliteration. A single-view model retrieves irrelevant cases and generates "clear lungs bilaterally", miss-

ing the effusion entirely. Our system identifies the lateral-view evidence, retrieves a matched historical case, and generates "small to moderate right pleural effusion is seen, likely causing some bibasilar atelectasis"—consistent with the ground truth.

Contributions. We propose a tightly coupled three-component framework:

1. **Multi-view soft alignment** (§3.1). A cross-view spatial affinity module establishes patch-level anatomical correspondences between PA and lateral features without requiring explicit anatomical annotation. Combined with a clinical soft-label contrastive loss, it improves CE-F1 from 0.432 to 0.467 ($+8.1\%$), with the largest gains on spatially dependent pathologies ($+0.057$ posterior effusion CE-F1).

2. **Confidence-aware adaptive RAG** (§3.2). A similarity threshold γ gates whether retrieved reports are injected into generation. Under adversarial retrieval conditions (pathology-inverted injections into 200 test cases), our full framework degrades by only $\Delta{=}{-}0.018$ CE-F1 versus $\Delta{=}{-}0.052$ for a model without this mechanism—a 65% reduction in retrieval-induced vulnerability.

3. **Asymmetric clinical GRPO** (§3.3). We adapt GRPO (Shao et al., 2024) with a composite reward in which false negatives for critical pathologies carry $1.5\times$ the false-positive penalty weight. This produces the largest per-class gain for pneumothorax ($+28.1\%$ CE-F1 from the retrieval stage to the full framework), five times the corresponding cardiomegaly gain.

On MIMIC-CXR, the complete framework achieves CE-F1 of 0.505, outperforming MiniGPT-4+I3+C2FD (Liu et al., 2024a) by $+6.8\%$ CE-F1 and GIT-CXR (Sîrbu et al., 2025) by $+10.3\%$ CE-F1. Cross-domain generalization to IU-Xray is presented in §5.1.

2 Related Work

Radiology Report Generation. Early RRG systems adopted CNN-RNN architectures (Jing et al., 2018; Xue et al., 2018); Transformer-era models such as R2Gen (Chen et al., 2020) and ME-Transformer (Wang et al., 2023) substantially improved quality through memory-driven normalisation and multi-expert attention. More recently, large vision-language models have been adapted for RRG: MiniGPT-4+I3+C2FD (Liu et al., 2024a) applies coarse-to-fine decoding to a general-purpose MLLM, RaDialog (Pellegrini et al., 2023) couples BioViL-T with a LLaMA decoder, and RADAR (Hou et al., 2025) augments generation with supplementary knowledge injection. Our work differs by jointly addressing view incompleteness, retrieval noise, and asymmetric clinical risk within a unified framework.

Multi-View Chest X-Ray Models. MAIRA (merc et al., 2024) and GIT-CXR (Sîrbu et al., 2025) both incorporate PA and lateral views, demonstrating that multi-view input improves NLG scores. However, both fuse features by concatenation or simple aggregation, without modelling geometric correspondence between orthogonal projections. Our cross-view spatial affinity module explicitly learns patch-level mappings across views, and our soft-label contrastive alignment corrects the false-negative problem that corrupts the retrieval space.

Contrastive Learning for Medical Imaging. CLIP (Radford et al., 2021) established image-text alignment via contrastive pre-training. MedCLIP (Wang et al., 2022) addressed frequent label co-occurrence in CXR data with a semantic matching loss. CXR-RePaiR (Endo et al., 2021) and X-REM (Jeong et al., 2023) built retrieval-based RRG systems on CLIP-style encoders. Our soft-label alignment extends MedCLIP's insight to the multi-view setting, adding the geometrically constrained cross-view affinity module.

Reinforcement Learning for Report Generation. Miura et al. (2021) optimised entity-level factual completeness; Delbrouck et al. (2022) used RadGraph-based semantic rewards; zhou et al. (2024) incorporated clinical quality metrics directly; Lin et al. (2025) grounded diagnostic reasoning in localised image regions. All existing methods apply *symmetric* penalties across pathology classes. To our knowledge, our asymmetric GRPO objective is the first RRG work to encode risk-stratified false-negative penalties within a GRPO framework, translating *primum non nocere* into a gradient-level bias.

3 Method

Figure 1 illustrates our framework across four structural pipeline stages spanning three tightly coupled core modules. Given paired PA and lateral chest X-rays, a shared ViT encoder extracts view-specific features, which are fused via a cross-view spatial affinity module and aligned with clinical soft labels (§3.1). The fused visual query retrieves top-K reports from a FAISS index, subject to confidence-aware filtering that prevents low-quality retrievals from contaminating generation (§3.2). The retrieved context and projected visual tokens are concatenated as input to Qwen3-VL-Thinking, fine-tuned with an asymmetric clinical GRPO objective (§3.3).

3.1 Multi-View Soft Alignment

Cross-View Spatial Affinity. Given PA features $F_{PA} \in \mathbb{R}^{L\times D}$ and lateral features $F_{Lat} \in \mathbb{R}^{L\times D}$ extracted by shared ViT encoder E_ϕ, we compute a cross-view attention matrix via learnable projections $W_Q, W_K \in \mathbb{R}^{D\times D}$:

$$\mathbf{A}_{cross} = \mathrm{Softmax}\left(\frac{(F_{PA}W_Q)(F_{Lat}W_K)^\top}{\sqrt{D}}\right) \tag{1}$$

The (i, j) entry of $\mathbf{A}_{cross}$ represents the anatomical correspondence strength between the i-th PA patch and the j-th lateral patch, implicitly mapping 2D projection points across views without explicit anatomical annotation. The affinity-enhanced PA representation is:

$$F^+_{PA} = F_{PA} + \mathbf{A}_{cross} \cdot (F_{Lat}W_V) \tag{2}$$

where $W_V \in \mathbb{R}^{D\times D}$ is a learnable value projection. We apply the symmetric operation to obtain F^+_{Lat}.

A lightweight MLP then computes per-view scalar weights:

$$[\alpha_{pa},\ \alpha_{lat}] = \mathrm{Softmax}\big([\mathrm{MLP}(F^+_{PA}),\ \mathrm{MLP}(F^+_{Lat})]\big) \tag{3}$$

$$f_{fused} = \alpha_{pa} \cdot F^+_{PA} + \alpha_{lat} \cdot F^+_{Lat} \tag{4}$$

This dynamic weighting reflects clinical intuition: when a consolidation dominates the PA view, the model assigns high α_{pa}; when only the lateral view reveals posterior costophrenic blunting, α_{lat} dominates (§5).

Clinical Soft-Label Contrastive Learning. Standard InfoNCE treats all off-diagonal pairs as hard negatives. In CXR datasets, two patients both presenting with cardiomegaly share genuine semantic similarity but are penalised as negatives, degrading intra-class clustering and—critically—distorting the retrieval space used by our RAG module.

We construct a **clinical semantic similarity matrix** $\mathbf{Y}^s \in \mathbb{R}^{N\times N}$ from 14-class pathology label vectors $c_i \in \{0, 1\}^{14}$ extracted by the CheXpert labeler:

$$y^s_{i,j} = \frac{\mathrm{sim}(c_i,\ c_j) + 1}{2} \tag{5}$$

This maps identical pathology profiles to y^s=1.0 and orthogonal profiles to y^s=0.5. The soft-label image-to-text contrastive loss is:

$$\mathcal{L}^{soft}_{v2t} = -\frac{1}{N}\sum_{i,j} y^s_{i,j} \cdot \log \frac{\exp\big(\mathrm{sim}(v_i, t_j)\,/\,\tau\big)}{\sum_k \exp\big(\mathrm{sim}(v_i, t_k)\,/\,\tau\big)} \tag{6}$$

with a symmetric $\mathcal{L}^{soft}_{t2v}$ term; total alignment loss $\mathcal{L}_{align} = (\mathcal{L}^{soft}_{v2t} + \mathcal{L}^{soft}_{t2v})\,/\,2$. The fused feature f_{fused} is projected into the LLM token embedding space via a linear projector P: $X_v = P(f_{fused})$.

3.2 Confidence-Aware Adaptive RAG

Offline Index Construction. We encode all MIMIC-CXR training reports with text encoder T_θ trained in §3.1, initialised from BGE-M3 and continued on radiology text. Since T_θ shares the semantic space with the visual encoder, the resulting embeddings $z_j = T_\theta(R_j) \in \mathbb{R}^D$ support cross-modal retrieval. We index all M vectors using FAISS with IVF+PQ compression, enabling sub-linear search at million scale.

Cross-Modal Retrieval and Confidence-Aware Filtering. At inference, the fused visual feature is pooled to a global query $q = \mathrm{Pooling}(f_{fused})$. The top-$K$ reports are recalled by cosine similarity. A critical failure mode in open clinical deployment is **low-confidence retrieval**: for rare or out-of-distribution cases, even the top-ranked retrieved report may be semantically misaligned. Blindly injecting such reports introduces strong misleading priors that the generator struggles to override.

We address this with a confidence-aware filtering

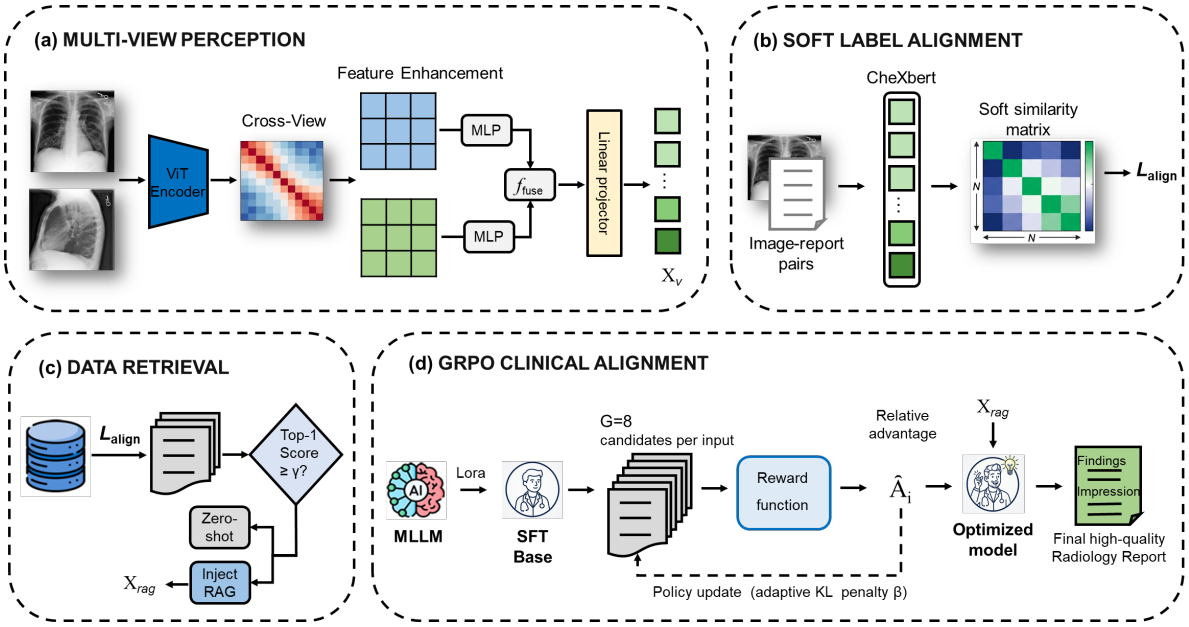


Figure 1: Framework overview. The pipeline flows through four sequential blocks (Multi-View Perception, Soft Label Alignment, Data Retrieval, and GRPO Clinical Alignment) to implement our three tightly coupled core modules addressing view incompleteness, retrieval noise, and asymmetric clinical risk.

mechanism:

$$\mathcal{P}_{\text{input}} = \begin{cases} \text{Prompt}_{\text{RAG}}(X_v,\ \mathcal{R}), \\ \quad \text{if score}(q, z_{\pi_1}) \geq \gamma \\ \text{Prompt}_{\text{ZS}}(X_v), \\ \quad \text{if score}(q, z_{\pi_1}) < \gamma \end{cases} \tag{7}$$

When confidence falls below γ, the system withholds retrieved context and relies on Qwen3-VL-Thinking's zero-shot visual reasoning and chain-of-thought capability—mirroring the clinical practice of expert radiologists who revert to first-principles image analysis for atypical findings. The threshold γ=0.65 is selected by cross-validation over $\gamma \in [0.50, 0.90]$ with step size 0.05, optimizing macro-average CheXbert F1 on the development set.

3.3 Asymmetric Clinical GRPO

Training Procedure. We adopt GRPO (Shao et al., 2024), which eliminates the critic network required by PPO by estimating advantages from within-group reward comparisons. For each training input $\mathbf{H}$, the policy π_θ generates G=8 candidate reports $\{o_1, \ldots, o_G\}$. Advantages are normalised within the group:

$$\hat{A}_i = \frac{r_i - \text{mean}(r_1, \ldots, r_G)}{\text{std}(r_1, \ldots, r_G)} \tag{8}$$

The policy is updated by maximizing:

$$J_{GRPO}(\theta) = \frac{1}{G}\sum_{i=1}^{G}\Big[\min\big(\rho_i \hat{A}_i,\ \text{clip}(\rho_i,\ 1-\epsilon,\ 1+\epsilon)\hat{A}_i\big) - \beta\, \mathbb{D}_{KL}(\pi_\theta \| \pi_{ref})\Big] \tag{9}$$

where $\rho_i = \pi_\theta(o_i|\mathbf{H})/\pi_{\theta_{old}}(o_i|\mathbf{H})$ and π_{ref} is the SFT-initialized reference policy. The KL penalty prevents reward hacking by anchoring the policy to the natural language prior learned during SFT. We use an adaptive β schedule: $\beta\times1.1$ when $\mathbb{D}_{KL}>0.1$, $\beta\times0.9$ when $\mathbb{D}_{KL}<0.05$.

Composite Reward Function. The reward combines three components:

$$R_{\text{total}} = \lambda_1 R_{\text{clin}} + \lambda_2 R_{\text{ling}} + \lambda_3 R_{\text{fmt}}, \quad (\lambda_1, \lambda_2, \lambda_3) = (0.6,\ 0.3,\ 0.1) \tag{10}$$

Clinical reward R_{clin}. CheXbert (Smit et al., 2020) extracts 14-class pathology labels from both the generated and reference reports; R_{clin} is the weighted F1 over these labels. To reflect the asymmetric clinical cost of missed diagnoses versus false alarms:

$$\omega_{FN}^{(c)} = \begin{cases} 1.5 \cdot \omega_{FP}^{(c)}, & c \in \mathcal{C}_{crit} \\ \omega_{FP}^{(c)}, & c \notin \mathcal{C}_{crit} \end{cases} \tag{11}$$

where $\mathcal{C}_{crit}$ = {Pneumothorax, Pulmonary Embolism, Tension Pneumothorax}. Missing a pneumothorax can precipitate tension pneumothorax within minutes; falsely reporting one leads to a confirmatory CT. The $1.5\times$ asymmetry instantiates *primum non nocere* directly in the optimization objective.

Linguistic reward R_{ling}: average of ROUGE-L F1 and METEOR, discouraging degenerate outputs that achieve high label accuracy through keyword stuffing.

Format reward R_{fmt}: binary check for "Findings" and "Impression" section headers, enforcing radiology reporting standards without post-processing.

4 Experiments

4.1 Experimental Setup

Dataset and Preprocessing. We evaluate on MIMIC-CXR (Johnson et al., 2019), comprising 377,110 images from 65,379 patients with paired free-text reports, following the official train/validation/test split and retaining only studies with both PA and lateral views. Reports are preprocessed with an LLM-assisted cleaning pipeline (Qwen3-235B-A22B-Instruct) to remove temporal comparison statements (e.g., *"in comparison with the preceding study..."*) and technical acquisition notes that introduce temporal hallucination bias. This step alone improves CE-F1 by $+0.027$ (0.388→0.415), confirming that label noise confounds prior comparisons on raw MIMIC-CXR.

Evaluation Metrics. We report BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) for NLG quality. For clinical validity we report **CE-F1**, the macro-averaged F1 over 14 pathology labels extracted by CheXbert (Smit et al., 2020) from both generated and reference reports; CE-F1 also serves as the primary reward signal in GRPO.

Implementation Details. We build on the MS-SWIFT fine-tuning framework with **Qwen3-VL-8B-Thinking** as the generation backbone, trained under DeepSpeed ZeRO-3 on NVIDIA A100 GPUs. SFT uses AdamW with learning rate 1×10^{-5}. GRPO uses learning rate 5×10^{-7}, group size G=8, clip ratio ϵ=0.2, adaptive β initialized at 0.01, 1 epoch, gradient clip 1.0. The FAISS index (IVF+PQ) is built over all training reports encoded by BGE-M3. Retrieval uses Top-K=3 with γ=0.65; end-to-end latency is 4.1 s/case on a single A100.

4.2 Main Results

Each component delivers consistent CE-F1 gains at every stage (Table 1). Our complete framework surpasses MiniGPT-4+I3+C2FD by $+0.032$ CE-F1 (relative $+6.8\%$) and $+0.091$ ROUGE-L ($+31.3\%$), and exceeds MedRG (Delbrouck et al., 2025) by $+0.037$ CE-F1 ($+7.9\%$). While MAIRA-2 (Bannur et al., 2024) reports a clinical metric of 0.281, this uses RadGraph-F1 (a different clinical metric, not directly comparable to CE-F1); nevertheless, our framework's absolute CE-F1 of 0.505 confirms stronger clinical label fidelity. GIT-CXR (Sîrbu et al., 2025), the strongest multi-view baseline in terms of NLG metrics, is exceeded by $+0.047$ CE-F1 despite its higher METEOR score, reflecting a systematic difference between surface-level n-gram overlap and pathology-label accuracy. Crucially, the data cleaning step contributes an independent $+0.027$ CE-F1 that prior comparisons on raw MIMIC-CXR inadvertently confound; our architectural gains are measured on top of this cleaner baseline.

4.3 Ablation Study

We conduct four targeted ablations on Qwen3-VL-8B-Thinking with cleaned MIMIC-CXR data.

Ablation I: Multi-View Alignment Design. We decompose the $+0.035$ CE-F1 gain of the multi-view module (0.432→0.467) by comparing four design choices.

Naive concatenation yields only $+0.009$ CE-F1, confirming that feature stacking introduces redundancy without modelling inter-view correspondence. The cross-view affinity module contributes a further $+0.014$ by establishing patch-level anatomical correspondences. Replacing hard InfoNCE with soft-label alignment provides the remaining $+0.012$, resolving the false-negative problem in which clinically similar cross-patient pairs are incorrectly penalised as negatives during contrastive training.

Ablation II: GRPO Reward Design. Starting from the base+MV+RAG configuration (CE-F1=0.488), we compare three reward designs.

GRPO with a symmetric reward improves CE-F1 by $+0.006$ over SFT via direct clinical label feedback. The asymmetric $1.5\times$ false-negative weight

Table 1: Performance comparison on the MIMIC-CXR test set. † = LLM-cleaned data; **Bold** = best overall. CE-P and CE-R are reported only for our models (baseline papers use different evaluation or averaging schemes, precluding direct precision/recall comparison).

Model	NLG Metrics						Clinical Metrics		
	B-1	B-2	B-3	B-4	R-L	MET	CE-P	CE-R	CE-F1
Prior Baselines									
CKNet (2023)	0.356	0.239	0.157	0.113	0.289	0.146	–	–	0.358
HERGen (2024)	0.395	0.248	0.169	0.122	0.285	0.156	–	–	0.317
MiniGPT-4+I3+C2FD (2024a)	0.402	0.262	0.180	0.128	0.291	0.175	–	–	0.473
GIT-CXR (2025)	0.403	0.286	0.215	0.168	0.312	**0.369**	–	–	0.458
MedRG (2025)	0.385	0.264	0.191	0.143	0.315	0.172	–	–	0.468
Ours (cleaned data†, incremental build-up)									
Qwen3-VL base SFT	0.410	0.280	0.198	0.165	0.360	0.205	0.440	0.425	0.432
+ Multi-View Soft Align.	–	–	–	0.169	0.368	0.210	–	–	0.467
+ Conf.-Aware RAG	–	–	–	0.177	0.375	0.221	–	–	0.488
+ Asymmetric GRPO **(Full)**	**0.435**	**0.305**	**0.225**	**0.185**	**0.382**	0.235	**0.512**	**0.498**	**0.505**

Table 2: Multi-view alignment design ablation.

Configuration	CE-F1
Single-view (PA only)	0.432
Multi-view concat	0.441
Cross-affinity + Hard InfoNCE	0.455
Cross-affinity + Soft-label (Ours)	**0.467**

Table 3: GRPO reward design ablation.

Configuration	CE-F1
SFT only (no RL)	0.488
GRPO + Symmetric penalty	0.494
GRPO + Asymmetric (Ours)	**0.505**

Table 4: Robustness to adversarial retrieval injection (200 cases); full test-set CE-F1 for this configuration is 0.488, see Table 1)

Configuration	Normal	Adv.	Δ
Base+MV+RAG, SFT only	0.478	0.426	−0.052
Full framework (Ours)	**0.505**	**0.484**	**−0.018**

Table 5: Chain-of-thought ablation (full framework).

Configuration	CE-F1
Qwen2.5-VL-Instruct (no CoT arch.)	0.474
Qwen3-VL-Thinking, CoT masked	0.481
Qwen3-VL-Thinking, full CoT (Ours)	**0.505**

for $\mathcal{C}_{crit}$ yields a further $+0.011$ CE-F1, with impact concentrating specifically on critical pathology classes as shown in Table 8 (§5).

Ablation III: RAG Robustness. To test confidence-aware filtering, we inject adversarially mismatched reports (pathology-inverted, randomly sampled) for 200 positive test cases and measure CE-F1 degradation.

Without GRPO, adversarial injection degrades CE-F1 by -0.052: SFT models lack a mechanism to override misleading text context. Our full framework's degradation is only -0.018, a **65% reduction in vulnerability**. GRPO's clinical reward builds a "visual evidence first" inductive bias: whenever retrieved text conflicts with X_v during training, the policy is penalised, teaching the model to treat retrieval as auxiliary evidence rather than a binding constraint—a critical safety property for clinical deployment.

Ablation IV: Chain-of-Thought Reasoning. We isolate the contribution of explicit reasoning by comparing three configurations under our full framework.

Upgrading the backbone while masking CoT yields only $+0.007$ CE-F1, attributable to architectural improvements (DeepStack multi-layer visual injection, 256K context window). Enabling full CoT provides the dominant gain of $+0.024$ CE-F1. For high-risk pathologies with subtle imaging features, the `<think>` block acts as a self-verification layer that explicitly revisits ambiguous regions before committing to a diagnosis—a behaviour absent in single-pass generation, consistent with the dual-process model of expert clinical reasoning.

5 Analysis

5.1 Cross-Domain Generalization

To assess whether RAG learns transferable pathological reasoning or merely memorizes source-domain writing templates, we evaluate the com-

Table 6: Cross-domain generalization (knowledge base: MIMIC-CXR training set; test domain: IU-Xray).

Configuration	B-4	R-L	MET	CE-F1
In-domain (MIMIC-CXR, Full)	0.185	0.382	0.235	0.505
IU-Xray, no RAG baseline	0.142	0.340	0.198	0.458
IU-Xray, full framework	0.158	0.355	0.212	**0.491**

Table 7: Sensitivity to retrieval count K (full framework).

K	B-4	R-L	CE-F1	Lat. (s)
1	0.171	0.366	0.480	3.2
3	**0.185**	**0.382**	**0.505**	4.1
5	0.182	0.379	0.496	5.8
10	0.174	0.370	0.482	9.3

plete framework on IU-Xray (Demner-Fushman et al., 2016) using a knowledge base built exclusively from MIMIC-CXR training reports. IU-Xray originates from Indiana University Health and differs substantially from MIMIC-CXR (Massachusetts General Hospital) in report phrasing and stylistic conventions, constituting a controlled domain-shift test.

Under domain shift, NLG metrics decline moderately (BLEU-4: $0.185 \rightarrow 0.158$), as cross-domain retrievals differ in surface phrasing from IU-Xray references. However, **CE-F1 remains robust at 0.491**, versus 0.458 for the no-RAG baseline (Δ=+0.033), and degrades only 0.011 from in-domain performance. This "NLG gap but CE-F1 robustness" dissociation demonstrates that RAG's core contribution is transferable pathological inference logic—the association between blunted costophrenic angles and pleural effusion, for instance—rather than domain-specific sentence templates.

Retrieval Count Sensitivity. All metrics follow an inverted-U. From K=1 to K=3, CE-F1 rises +0.022 as diverse references allow triangulation across phrasings and complementary findings. Beyond K=3, performance degrades monotonically: lower-ranked retrievals decay in semantic relevance, and the expanding text context competes with visual tokens X_v for effective attention—consistent with context dilution in long-context LLMs (Liu et al., 2024b). The K=5 marginal cost is −0.006 CE-F1 at +41% latency, confirming K=3 as optimal.

Table 8: Per-pathology CE-F1 across framework stages.

Pathology	Base	+MV	+RAG	+GRPO
High-frequency				
Cardiomegaly	0.621	0.638	0.651	0.659
Pleural Effusion	0.573	0.601	0.624	0.631
Consolidation	0.512	0.531	0.558	0.566
Spatially dependent				
Posterior Effusion	0.341	0.398	0.412	0.421
Retrocardiac Opacity	0.298	0.362	0.374	0.382
Critical / low-frequency				
Pneumothorax	0.301	0.319	0.334	**0.428**
Pulm. Embolism	0.244	0.258	0.271	**0.341**
Atelectasis	0.443	0.461	0.479	0.491

Per-Pathology Breakdown. Table 8 stratifies CE-F1 by pathology across framework stages, revealing three distinct improvement patterns.

Spatially dependent pathologies benefit most from multi-view alignment: posterior effusion CE-F1 improves +0.057 (Base→+MV), the largest single-module gain, directly validating the cross-view affinity module's ability to compensate for retrocardiac and posterior angle occlusion. **High-frequency pathologies** gain most from RAG (+0.023–+0.027 per stage), as retrieved reports supply standardized terminology. **Critical pathologies** exhibit a distinctive profile: moderate early gains, then a disproportionately large GRPO jump. Pneumothorax CE-F1 improves +0.094 from +RAG to +GRPO (+28.1% relative)—nearly twelve times the corresponding cardiomegaly gain of +0.008—directly attributable to the asymmetric false-negative penalty in Eq. (11). This confirms that the asymmetric reward reshapes the decision boundary specifically for $\mathcal{C}_{crit}$, without homogenizing gains across all classes.

View Attention Weight Visualization. Figure 2 visualizes dynamic view weights $(\alpha_{pa}, \alpha_{lat})$ for two representative cases. For Case A (right upper lobe consolidation), the model assigns α_{pa}=0.85, as the lesion dominates the PA view. For Case B (posterior trace effusion visible only in the lateral view), weights invert to α_{lat}=0.80. The MLP weight predictor learns this diagnostic value asymmetry entirely from data, confirming that the module internalizes orthogonal projection geometry without explicit anatomical supervision.

6 Conclusion

We presented a safety-aware framework for chest X-ray report generation integrating three coupled

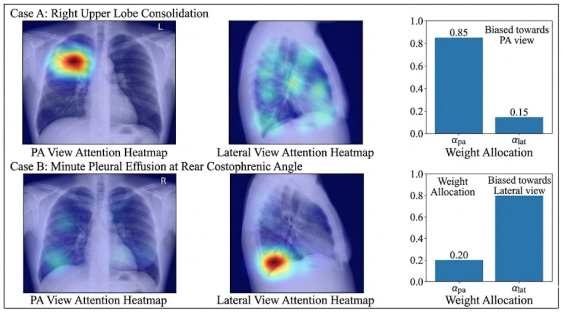


Figure 2: Dynamic view weight visualization. Case A (consolidation, top): α_{pa}=0.85. Case B (trace posterior effusion, bottom): α_{lat}=0.80. Heatmaps show ViT Rollout attention projected back to image space.

components: multi-view soft alignment resolves view incompleteness and retrieval space distortion; confidence-aware adaptive RAG prevents low-quality retrieval contamination; and asymmetric clinical GRPO encodes risk stratification directly into the optimization objective. On MIMIC-CXR, the complete framework achieves CE-F1 of 0.505, surpassing the strongest prior baseline by $+6.1\%$ CE-F1. Component analysis reveals module-specific strengths: multi-view alignment is most impactful for spatially dependent pathologies ($+0.057$ posterior effusion CE-F1); confidence-aware RAG reduces retrieval vulnerability by 65% under adversarial injection; and asymmetric GRPO produces a $+28.1\%$ pneumothorax CE-F1 gain without homogenizing gains across all pathology classes. Future work will explore LLM-as-judge reward models (Zheng et al., 2023) for finer-grained clinical feedback, and prospective radiologist blind-review validation (Tanno et al., 2025) to close the gap between automated metrics and expert assessment.

7 Limitations

We identify three primary limitations of the current framework.

CheXbert reward proxy. The 14-class CheXbert vocabulary used as our GRPO reward signal may suppress low-frequency synonymous expressions (e.g., "parenchymal opacification" for "consolidation"), systematically penalising linguistically diverse outputs. The discrete label space also cannot capture spatial lesion grading or diagnostic confidence nuance. This creates a risk of reward hacking toward keyword-dense but clinically shallow reports.

Long-tail rare pathologies. Our adversarial robustness experiments are conducted on 200 cases with common pathologies. For ultra-rare findings (e.g., pneumomediastinum, chylothorax), the FAISS index contains few relevant neighbors, the asymmetric penalty provides minimal gradient signal, and chain-of-thought reasoning has limited training examples to generalize from. Performance on such long-tail cases remains unquantified.

Visual resolution ceiling. The ViT encoder compresses images to fixed-size patches, potentially discarding sub-centimeter lesions (e.g., early nodules <5 mm) that may be clinically significant. This is an architectural constraint shared with all patch-based vision-language models.

Ethics Statement

Data. All experiments use MIMIC-CXR (Johnson et al., 2019), a publicly available, IRB-approved, de-identified dataset. No patient identifiers are stored or processed in our pipeline.

Intended use. This system is designed as a *decision support tool* for qualified radiologists, not as a replacement for expert clinical judgment. Generated reports should be reviewed and verified by a licensed clinician before any clinical action is taken.

Deployment prerequisites. Our evaluation is conducted on retrospective benchmark data. Before clinical deployment, prospective validation through radiologist blind-review trials under appropriate IRB oversight is mandatory. The system's performance on patient populations underrepresented in MIMIC-CXR (e.g., paediatric patients, non-Western demographics) has not been characterized.

Failure mode disclosure. As documented in §7, the system shows reduced reliability for ultra-rare pathologies. Any deployment must include human oversight mechanisms and clear communication of these limitations to end users.

Acknowledgments

This work was supported by [funding sources]. We thank the anonymous reviewers for their constructive feedback.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Ex-*

trinsic Evaluation Measures for Machine Translation and/or Summarization, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, and 1 others. 2024. Maira-2: Grounded radiology report generation. *Computing Research Repository*, arXiv:2406.04449.

Zhihong Chen, Yizhou Song, Tsung-Hui Chang, and Wan Xiang. 2020. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449.

Jean-Benoit Delbrouck, Pierre Chambon, Zhihong Chen, and 1 others. 2022. Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4348–4360.

Jean-Benoit Delbrouck, Junyi Xu, Julian Moll, and 1 others. 2025. Automated structured radiology report generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26813–26829.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, and 1 others. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Mark Endo, Ray Krishnan, Viswanath Krishna, and 1 others. 2021. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In *Proceedings of Machine Learning for Health*, pages 209–219. PMLR.

Wenjun Hou, Yu Cheng, Kaidi Xu, and 1 others. 2025. Radar: Enhancing radiology report generation with supplementary knowledge injection. *Preprint*, arXiv:2505.14318.

Jaehwan Jeong, Katherine Tian, Andrew Li, and 1 others. 2023. Multimodal image-text matching improves retrieval-based chest x-ray report generation. In *Medical Imaging with Deep Learning (MIDL)*, pages 1–13.

Baoyu Jing, Pengtao Xie, and Eric Xing. 2018. On the automatic generation of medical imaging reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2577–2586.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, and 1 others. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317.

Bing Kang, Yating Xiong, Jianbo Jiao, Yue Zhang, Xiaohui Jia, and Jing Li. 2023. Bridging the gap: Cross-modal knowledge driven network for radiology report generation. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1202–1209. IEEE.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Q Lin, Y Zhu, B Pu, and 1 others. 2025. A foundation model for chest x-ray interpretation with grounded reasoning via online reinforcement learning. *Preprint*, arXiv:2509.03906.

Chang Liu, Yuanhe Tian, Weidong Chen, Yan Song, and Yongdong Zhang. 2024a. Bootstrapping large language models for radiology report generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):18635–18643.

Nelson F. Liu, Kevin Lin, Peter Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024b. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Shreyasi Srivastav merc, Mercy Ranjit, Fernando Pérez-García, and 1 others. 2024. Maira at rrg24: A specialised large multimodal model for radiology report generation. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 597–602.

Yasuhide Miura, Yuhao Zhang, Emily B Tsai, and 1 others. 2021. Improving factual completeness and consistency of image-to-text radiology report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 5288–5304.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Chantal Pellegrini, Ege Özsoy, Benjamin Busam, and 1 others. 2023. Radialog: A large vision-language model for radiology report generation and conversational assistance. *Preprint*, arXiv:2311.18681.

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 (ICML)*, pages 8748–8763. PMLR.

Abi Rimmer. 2017. Radiologist shortage leaves patient care at risk, warns royal college. *British Medical Journal*, 359.

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

Iustin Sîrbu, Iulia-Renata Sîrbu, Jasmina Bogojeska, and Traian Rebedea. 2025. Git-cxr: End-to-end transformer for chest x-ray report generation. *Information*, 16(7):524.

Philip Sloan, Philip Clatworthy, Edwin Simpson, and Majid Mirmehdi. 2024. Automated radiology report generation: A review of recent advances. *IEEE Reviews in Biomedical Engineering*, 18:368–387.

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. *Preprint*, arXiv:2004.09167.

Ryutaro Tanno, David GT Barrett, Andrew Sellergren, and 1 others. 2025. Collaboration between clinicians and vision-language models in radiology report generation. *Nature Medicine*, 31:599–608.

Fuying Wang, Du Shenghui, and Lequan Yu. 2024. Hergen: Elevating radiology report generation with longitudinal data. *Preprint*, arXiv:2407.15158.

Zifeng Wang, Lei Liu, Lei Wang, and Ligong 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 (CVPR)*, 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 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3876–3887. Association for Computational Linguistics.

Yuan Xue, Tao Xu, L Rodney Long, and 1 others. 2018. Multimodal recurrent model with attention for automated radiology report generation. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 457–466.

Sheng Zhang, Samuel Sambara, Oishi Banerjee, and 1 others. 2024. Radflag: A black-box hallucination detection method for medical vision language models. *Preprint*, arXiv:2411.00299.

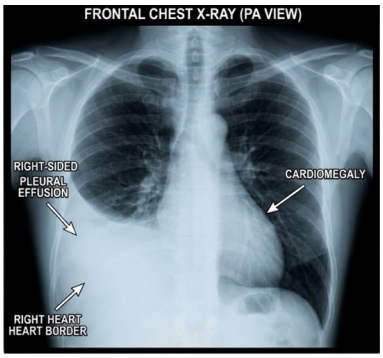
Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Z zhou, M Shi, M Wei, and 1 others. 2024. Large model driven radiology report generation with clinical quality reinforcement learning. *Preprint*, arXiv:2403.06728.

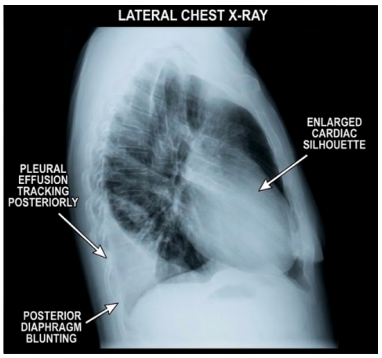
A Qualitative Analysis and Case Study

Figure 3 illustrates how multi-view alignment, confidence-aware RAG, and chain-of-thought reasoning interact on a representative MIMIC-CXR test case. The baseline single-view model hallucinates "clear lungs bilaterally", missing both the pleural effusion and the cascading atelectasis. Our framework cross-inspects the lateral view, gates retrieval above the confidence threshold, and produces a report fully congruent with the ground truth.

As illustrated in Figure 3, the single-view or naive baseline fails to resolve the posterior blind spots and falls prey to clinical false negatives (reporting "clear lungs bilaterally"). In contrast, our framework activates a systemic verification process within the thinking trace, anchoring its clinical inference heavily on the visual anomalies exposed by the cross-view affinity module. This demonstrates that the framework does not merely copy non-parametric writing templates but actively utilizes them as diagnostic hypotheses to be verified against raw multi-view pixel evidence.



(a) PA View — Cardiomegaly is prominent; posterior costophrenic angle is **obscured**.



(b) Lateral View — Posterior costophrenic angle **blunting** is unambiguous.

(c) Confidence-Aware RAG — Top-1 Retrieved Reference $\texttt{score}(q, z_{\pi_1}) = 0.81 \geq \gamma$

"The heart size is moderately enlarged. There is a right-sided pleural effusion with associated lower lobe compressive atelectasis."

(d) Baseline (Qwen2.5-VL-Instruct) — single view, no RAG, no GRPO

"Cardiomegaly is present. **Clear lungs bilaterally.** No pneumothorax."

× Missed findings: right pleural effusion, lower lobe atelectasis.

(e) Our Framework — Internal Chain-of-Thought Trace `<think>`

Step 1 [Observation]: Cardiac silhouette is enlarged — cardiomegaly confirmed.

Step 2 [Multi-View Cross-Inspection]: Lateral view (Fig. 3b) reveals unambiguous posterior costophrenic angle blunting, which is occluded by the cardiac shadow in the PA view (Fig. 3a).

Step 3 [Context Alignment]: Retrieved reference reports right-sided pleural effusion. Cross-verifying with lateral evidence (α_{lat}=0.80): hypothesis confidence is high.

Step 4 [Cascade Verification]: Pleural effusion at the right base typically induces adjacent lower lobe compressive atelectasis; right-base density pattern is consistent.

Step 5 [Conclusion]: Synthesize into structured FINDINGS / IMPRESSION output.

(f) Our Framework — Final Generated Report

Findings: Moderate cardiomegaly is noted. A small to moderate right pleural effusion is seen, which appears to cause some surrounding bibasilar compressive atelectasis. The left lung remains clear. No pneumothorax identified.

Impression: (1) Moderate cardiomegaly. (2) Small to moderate right pleural effusion with associated lower lobe atelectasis.

✓ Fully congruent with ground truth.

(g) Ground Truth Report

"The heart size is moderately enlarged. There is a right-sided pleural effusion with associated lower lobe atelectasis."

Figure 3: Qualitative case study: right-sided pleural effusion with cardiomegaly. **(a)–(b)** Input PA/lateral pair; complementary spatial evidence. **(c)** Top-1 retrieval (similarity 0.81, above threshold γ). **(d)** Baseline hallucinates "clear lungs", missing effusion and atelectasis. **(e)** Our framework's thinking trace cross-references lateral blunting with the retrieved template via the multi-view affinity module (α_{lat}=0.80, dominant weight on lateral view). **(f)** Final report captures all findings; **(g)** ground truth confirms full congruence.