CliQ-RRG: Clinical-Knowledge Guided Disease-aware Visual–Textual Alignment for QA-Style Radiology Report Generation

Chandravardhan Singh Raghaw[1], **Shahid Shafi Dar**[1], **Arnav Nishith Sharma**[2], **Shrey Modi**[3], **Nagendra Kumar**[1,*]
[1]Indian Institute of Technology Indore, Indore, India; [2]University of Texas at Dallas, USA;
[3]University of Wisconsin - Madison, USA
{phd2201101016, phd2201201004, nagendra}@iiti.ac.in,
axs230011@utdallas.edu, modi7@wisc.edu
*Corresponding author

Abstract

Radiology reports are essential for diagnostic reasoning and patient care, yet their manual preparation is time-consuming and cognitively demanding. Automatic radiology report generation (RRG) offers a scalable alternative, but existing models often produce lengthy, unstructured narratives that overlook diagnostic cues and multi-view information. We present **Cli**nical-Knowledge Guided Disease-aware Visual–Textual Alignment for **Q**A-Style **R**adiology **R**eport **G**eneration (CliQ-RRG), a unified two-stage framework for interpretable and clinically grounded reporting. In Stage 1, CliQ-RRG employs a Disease-aware Visual–Textual Alignment module that aligns image and text representations using predicted disease embeddings, reinforced by a Prior-Guided Attention Module (PrAM) to capture multi-view dependencies across current and prior scans. In Stage 2, GraphRAG-guided retrieval augments intermediate textual representations with graph-grounded clinical evidence, and a large language model restructures them into concise, interpretable question–answer (QA) pairs with diagnostic summaries. Experiments on two public chest X-ray benchmarks demonstrate that CliQ-RRG consistently outperforms prior methods across both lexical and clinical metrics, generating accurate and clinically coherent QA-style radiology reports. Code is available at https://github.com/chandravardhan-singh-raghaw/CliQ-RRG.

1 Introduction

Radiology report generation (RRG) is vital for chest X-ray (CXR) interpretation, requiring substantial clinical expertise and reasoning (Jin et al., 2024). Manual interpretation, however, remains time-consuming and cognitively demanding, even for experts (Liu et al., 2025a; Park et al., 2025). As imaging volumes rise, radiologists face increasing workload, often affecting report quality and diagnostic accuracy. Consequently, automated RRG has emerged as a promising solution (Hou et al., 2023; Liang et al., 2024; Luo et al., 2024), yet most methods emphasize disease-specific regions, overlooking broader contextual information.

Recent studies have proposed advanced techniques to improve automatic RRG. Li et al. (2024b) integrates graph-enhanced and regional features to describe normal and abnormal findings, yet outputs remain lengthy and less interpretable. To embed clinical knowledge, Hou et al. (2025) and Sun et al. (2025) retrieve domain-specific information, but the generated texts lack diagnostic coherence. Huang et al. (2025) aligns visual and textual features via a cross-modal adapter; however, the approach is inefficient and overlooks supervised labels that strengthen the alignment. Dual-view RRG methods (Chen et al., 2021; Yang et al., 2023) differentiate radiographic views but miss fine anatomical and contextual relations across scans, limiting multi-view consistency in generated reports.

Despite progress in RRG (Zhang et al., 2020b; Yan and Pei, 2022), existing methods face four key limitations (Fig. 1). (**G1**) Prior works (Jin et al.,

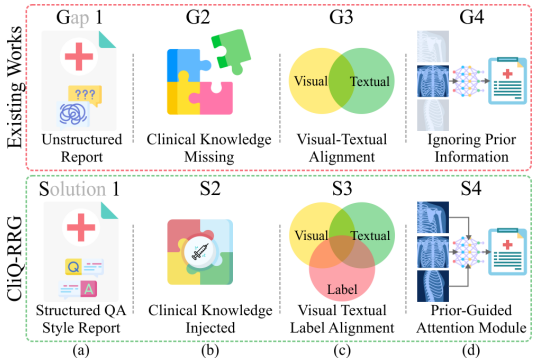


Figure 1: Comparison with existing methods. Existing approaches generate unstructured reports, overlook prior scans, lack clinical knowledge, and exhibit limited alignment. In contrast, CliQ-RRG generates QA-style reports with clinical knowledge enrichment, disease-aware alignment, and prior-guided attention.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 50153–50171
October 24-29, 2026 ©2026 Association for Computational Linguistics

2024; Luo et al., 2024; Wang et al., 2025) generate long free-text narratives, in which key findings are embedded within paragraphs rather than presented in a structured form. Existing studies have limitations of lengthy and inconsistent structured clinical text (Kell et al., 2024; Pattnayak et al., 2025; Schwartz et al., 2011). We therefore explore a QA-style reformulation to make findings explicit and easier to retrieve, aiming to improve structural clarity rather than replace narrative reports. (**G2**) Many studies (Yan et al., 2023; Bu et al., 2024a; Gu et al., 2024) rely on image–report pairs, which may yield fluent but underspecified descriptions, especially for subtle findings. We incorporate external clinical knowledge at the intermediate stage to promote clinically grounded expressions and strengthen alignment between visual evidence and generated text. (**G3**) Recent studies (Li et al., 2023; Shen et al., 2024; Xiao et al., 2025) align visual and textual modalities but neglect diagnostic labels as supervisory signals for joint embedding learning, unlike prompt-based approaches (Jin et al., 2024). We integrate predicted diagnostic labels directly into a contrastive alignment objective, structuring the joint visual–textual space based on diagnostic similarity. (**G4**) While several methods (Hou et al., 2023; Gu et al., 2025; Liu et al., 2025b) leverage historical information, most standard approaches do not explicitly fuse prior scans at the visual representation level. We introduce prior-image fusion before alignment to capture temporal context within the learned feature space.

We present Clinical-Knowledge Guided Disease-aware Visual–Textual Alignment for QA-Style Radiology Report Generation (CliQ-RRG), a unified framework centered on two core contributions: disease-grounded GraphRAG retrieval and fine-tuned LLM-based QA restructuring. As shown in Fig. 1, CliQ-RRG addresses limitations of prior RRG methods. (**S1**) It reformulates free-text reports into clinically interpretable QA-style outputs. (**S2**) It injects external clinical knowledge to enable reasoning beyond data-driven correlations. (**S3**) It aligns visual and textual representations via predicted disease embeddings to achieve semantically grounded features. (**S4**) It incorporates multi-view information from prior examinations through a prior-guided attention mechanism to enhance contextual understanding. Extensive experiments on **MIMIC-CXR** and **IU X-Ray** show that CliQ-RRG consistently outperforms state-of-the-art methods in radiology report generation.

Our main contributions are threefold:

- To the best of our knowledge, CliQ-RRG is the first framework to leverage contrastive alignment for QA-style radiology report generation, restructuring unstructured narratives into concise, structured question-answer pairs.
- We introduce a Disease-aware Visual-Textual Alignment module that aligns image and text representations via predicted disease embeddings, guided by a Prior-Guided Attention Module for multi-view contextual integration.
- We design a GraphRAG-guided generation pipeline that enriches intermediate reports with graph-grounded clinical evidence and employs a large language model to synthesize the final QA-style report.

2 Proposed Approach

2.1 Problem Formulation and Overview

Let $\mathcal{T} = \{(I_k, R_k, L_k)\}_{k=1}^{n}$ be the training set of n studies, each with a CXR $I_k \in \mathbb{R}^{h\times w\times(v+1)}$, associated report R_k, and predicted disease labels L_k, where v is the number of prior scans ($v = 0$ if unavailable), and h, w are image dimensions. Our objective is to generate a QA-style report $R_{qa}(q_i, a_i)$, where each question q_i targets a clinical finding, and the answer $a_i \in \{\text{Yes, No}\}$ is grounded in visual evidence from I. The value m is the total number of generated question–answer pairs. We hypothesize that aligning visual and textual representations with predicted disease labels allows our framework to generate QA-style radiology reports. We formalize the generation process as:

$$R_{qa}(q_k, a_k) = \text{CliQ-RRG}(I, R, L) \quad (1)$$

Our proposed framework, CliQ-RRG, operates in two stages, illustrated in Fig. 2. STAGE 1 employs a disease-aware visual–textual alignment module to align visual features from current and prior chest X-rays with textual representations and predicted disease labels in a unified embedding space. A prior-guided attention module further captures multi-view information from prior scans, enhancing temporal and anatomical consistency. In STAGE 2, a text decoder generates an intermediate report, which we enhance through GraphRAG-guided retrieval of graph-grounded clinical evidence. Finally, an LLM restructures the knowledge-injected report into a concise QA-style format with question–answer pairs and a clinical summary.

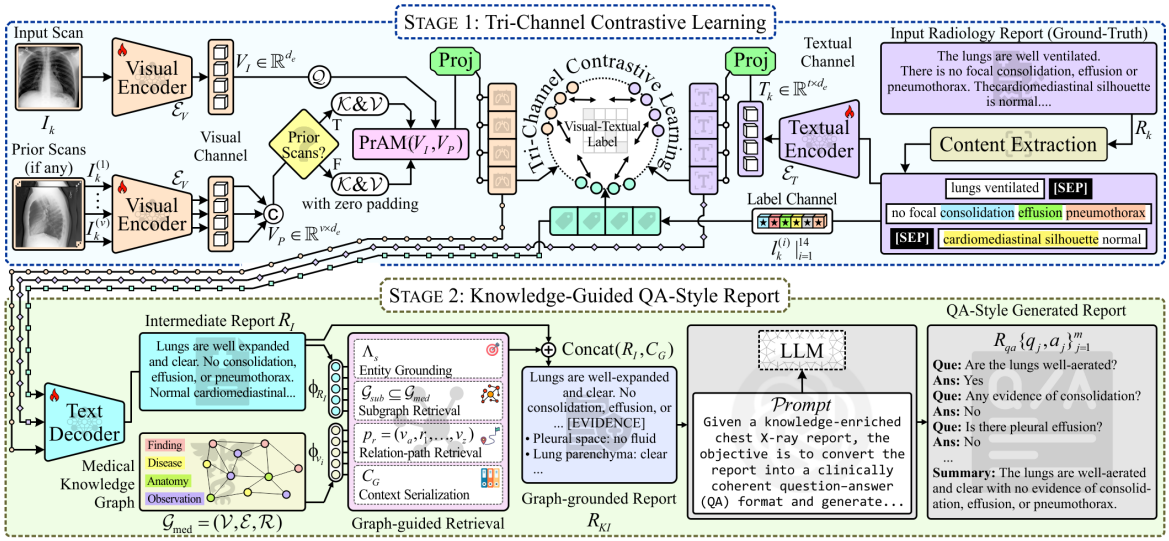


Figure 2: Overview of the proposed CliQ-RRG framework. STAGE 1: Disease-aware Visual–Textual Alignment aligns image and text representations using predicted disease embeddings, while the Prior-Guided Attention Module (PrAM) integrates multi-view information from current and prior scans. STAGE 2: Knowledge-Guided QA-Style Report Generation enriches the intermediate report with graph-connected clinical evidence retrieved from a medical knowledge graph and employs a large language model to produce concise QA pairs and a diagnostic summary.

2.2 STAGE 1: Disease-aware Visual–Textual Alignment

Visual Channel: Let I_k denote the current CXR and $I_k^{(j)}{}_{j=1}^{v}$ the set of v prior scans for the k^{th} patient ($v \geq 0$). We use the BioMedCLIP (Zhang et al., 2025a) visual encoder $\mathcal{E}_V(\cdot)$ to extract image features. The current view is encoded as $V_I = \mathcal{E}_V(I_k)$, and each prior view as $V_P^{(j)} = \mathcal{E}_V(I_k^{(j)})$. Prior features are concatenated to form a contextual representation $V_P = \text{Concat}(V_P^{(1)}, \ldots, V_P^{(v)})$, zero-padded when $v = 0$. The visual channel representation is defined as $V_{ch} = \text{PrAM}(V_I, V_P)$, where PrAM fuses the current image features V_I and prior image features V_P, capturing both current and multi-view information.

Prior-guided Attention Module: To handle variable image counts across studies, which complicate concatenation and hinder cross-modal alignment, we introduce the Prior-Guided Attention Module (PrAM). PrAM integrates multi-view information from prior scans through a cross-attention mechanism (Vaswani et al., 2017). The current image feature $V_I \in \mathbb{R}^{d_e}$ serves as the query $\mathcal{Q}$, and the prior features $V_P \in \mathbb{R}^{v \times d_e}$ act as both keys $\mathcal{K}$ and values $\mathcal{V}$, where d_e is the shared embedding dimension. The output is computed as:

$$\text{PrAM}(V_I, V_P) = \text{softmax}\left(\frac{V_I V_P^\top}{\sqrt{d_e}}\right) \cdot V_P \quad (2)$$

producing a unified representation that preserves multi-view context while maintaining consistent dimensionality for contrastive alignment. PrAM does not explicitly encode the temporal order or time intervals between prior scans; hence, it captures historical context but does not explicitly model longitudinal progression patterns.

Textual Channel: To process radiology reports R_k, we construct a textual channel T_{ch} that encodes clinical information. Following (Yan et al., 2023; Liu et al., 2025a, 2026), we extract a sequence of t_c salient clinical phrases from R_k, separated by `[SEP]` tokens, as shown in Fig. 2. This sequence is passed through a textual encoder $\mathcal{E}_T(\cdot)$, instantiated as CXR-BERT (Boecking et al., 2022), yielding textual features $T_k = \mathcal{E}_T(R_k) \in \mathbb{R}^{t_c \times d_e}$. These representations enable alignment with visual and label modalities during contrastive learning. During inference, ground-truth reports are unavailable; therefore, T_{ch} is not used. The decoder relies solely on the visual channel V_{ch} and predicted diagnostic labels L_{ch}. Since V_{ch} has been aligned with T_{ch} during training, it encodes the necessary semantic information to guide generation.

Disease Classifier: To capture disease-relevant supervision, we design a disease classifier that predicts diagnostic labels (L_{ch}) from the visual channel features $V_{ch} \in \mathbb{R}^{d_e}$. Rather than a single softmax jointly over all 13 disease categories, the classifier computes, for each disease category i, a probability distribution over CheXbert's four states (uncertain, negative, pos-

itive, not mentioned) independently, using a cross-attention mechanism (Vaswani et al., 2017):

$$\hat{Y}^{(i)} = \text{Softmax}\left(\frac{V_{ch}\Phi^{(i)\top}}{\sqrt{d_e}}\right), \hat{Y}^{(i)} \in \mathbb{R}^4 \quad (3)$$

$$L_{ch}^{(i)} = \arg\max_{s\in\{1,2,3,4\}} \hat{Y}^{(i,s)} \quad (4)$$

$$\mathcal{L}_{CE} = \sum_{i=1}^{13} \text{CrossEntropy}(\hat{Y}^{(i)}, L_{gt}^{(i)}) \quad (5)$$

where $\Phi \in \mathbb{R}^{(d\times 4)\times d_e}$ is a learnable disease embedding matrix and L_{gt} represents the ground-truth disease labels extracted using CheXbert (Smit et al., 2020). Each label $l_k^{(i)} \in \{-1, 0, 1, 2\}$ denotes uncertainty, negative, positive, or not mentioned, respectively. This yields genuinely multi-label, multi-state prediction, with each of the 13 disease categories classified independently over its own 4-state distribution rather than one distribution shared across all diseases. We retain the first 13 disease categories to identify reports with positive or uncertain findings. Appendix I reports per-state classification performance validating this formulation.

Disease-aware Visual–Textual Contrast: Building on prior contrastive and alignment-based RRG methods that align visual and textual representations through learned knowledge bases or region-level correspondence (Yang et al., 2023; Chen et al., 2024), we propose Disease-aware Visual–Textual Contrast (Di-VTC), which adapts contrastive learning to the generative setting of QA-style radiology report generation. While prior work applies contrastive learning primarily in discriminative settings, Di-VTC learns the parameters θ_V and θ_T of the visual ($\mathcal{E}_V$) and textual ($\mathcal{E}_T$) encoders to align visual (V_{ch}) and textual (T_{ch}) representations in a shared embedding space. For samples g and h in a batch B, the channel features are projected and L_2-normalized to yield $\hat{v}_g$=Norm(Proj($V_{ch}^{(g)}$)) and $\hat{t}_h$=Norm(Proj($T_{ch}^{(h)}$)). A pair is treated as positive if their predicted diagnostic labels are similar, denoted as $l_g \approx l_h$, where two samples are considered similar when their disease label vectors are identical or satisfy a predefined similarity threshold; otherwise, the pair is treated as negative. As shown in Fig. 3, Di-VTC employs a push–pull dynamic that maximizes $\hat{v}_g^\top \hat{t}_h$ for positive pairs and minimizes it for negative pairs. The model is optimized using a bidirectional contrastive loss $\mathcal{L}_{\text{Di-VTC}}$ over θ_V and θ_T.

$$\min_{\{\theta_V,\theta_T\}} \mathcal{L}_{\text{Di-VTC}} = \mathcal{L}_{v\mapsto t} + \mathcal{L}_{t\mapsto v} \quad (6)$$

The $\mathcal{L}_{\text{Di-VTC}}$ loss combines a visual-to-textual loss ($\mathcal{L}_{v\mapsto t}$) and a textual-to-visual loss ($\mathcal{L}_{t\mapsto v}$):

$$\mathcal{L}_{v\mapsto t} = -\sum_{g\in B}\frac{1}{|\text{p}(g)|}\sum_{i\in\text{p}(g)}\log\frac{\exp(\hat{v}_g^\top \hat{t}_i/\tau)}{\sum_{h\in B}\exp(\hat{v}_g^\top \hat{t}_h/\tau)} \quad (7)$$

$$\mathcal{L}_{t\mapsto v} = -\sum_{h\in B}\frac{1}{|\text{p}(h)|}\sum_{i\in\text{p}(h)}\log\frac{\exp(\hat{v}_i^\top \hat{t}_h/\tau)}{\sum_{g\in B}\exp(\hat{v}_g^\top \hat{t}_h/\tau)} \quad (8)$$

where $\text{p}(\cdot)$ denotes the set of positive pair indices for a given anchor sample, and τ is a temperature hyperparameter. The Di-VTC framework aligns visual and textual representations using similarity in predicted disease embeddings as supervision, enabling clinically grounded and coherent QA-style report generation.

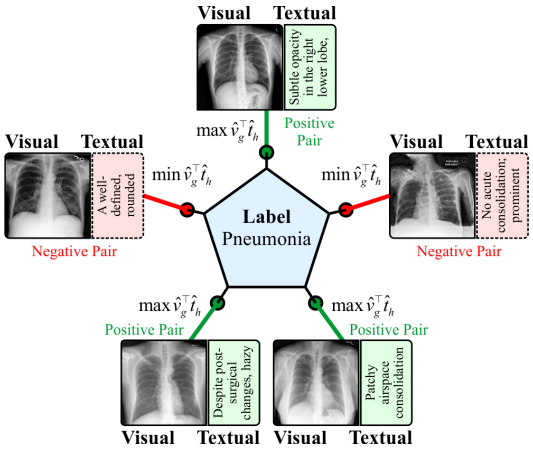


Figure 3: Disease-aware Visual–Textual Contrast (Di-VTC). Positive visual–textual pairs with shared disease labels (green) are pulled closer, while unrelated pairs (red) are pushed apart, enforcing disease-consistent alignment in the shared embedding space.

2.3 Stage 2: Knowledge-Guided QA-Style Report

Intermediate Report Generation: We generate the intermediate report R_I using a text decoder conditioned on visual channel V_{ch}, textual channel T_{ch}, and predicted diagnostic labels L_{ch} from Stage 1. For each k^{th} study, the decoder predicts token R_k^t autoregressively using preceding tokens $R_k^{<t}$ and multi-channel representations. The decoder is optimized using negative log-likelihood loss:

$$\mathcal{L}_G = -\sum_{t=1}^{l}\log(R_k^t | R_k^{<t}, V_{ch}, T_{ch}, L_{ch}) \quad (9)$$

where l denotes report length and $p(R_k^t|\cdot)$ denotes the probability of generating token R_k^t conditioned on preceding tokens and decoder inputs. The loss

$\mathcal{L}_G$ encourages clinically coherent report generation aligned with (V_{ch}, T_{ch}, L_{ch}).

Graph-guided Retrieval: Unlike prior knowledge-injected RRG frameworks that incorporate external clinical text during decoding (Huang et al., 2023; Bu et al., 2024a), CliQ-RRG performs Graph Retrieval-Augmented Generation (GraphRAG) before QA synthesis. Instead of retrieving isolated knowledge fragments, our framework retrieves clinically connected evidence through entity grounding, subgraph retrieval, and relation-path reasoning.

We organize the external clinical repository (Appendix C) as a medical knowledge graph:

$$\mathcal{G}_{med} = (\mathcal{V}, \mathcal{E}, \mathcal{R}) \tag{10}$$

where $\mathcal{V}$ denotes clinical entities including anatomy, findings, and diseases. Edges $e = (v_a, r, v_b) \in \mathcal{E}$ encode medical relations $r \in \mathcal{R}$.

Given intermediate report R_I, which serves as the retrieval query, and predicted disease labels L_{ch} from Stage 1, we perform entity grounding by projecting report and graph entities into a shared semantic space. Intermediate reports are encoded using CXR-BERT to obtain query representation ϕ_{R_I}, while graph entities are initialized using BioWordVec embeddings ϕ_{v_i}. Graph retrieval relevance score $S(v_i)$ for graph entity v_i is computed as:

$$\begin{aligned} S(v_i) = \alpha\, \mathrm{sim}(\phi_{R_I}, \phi_{v_i}) \\ + (1-\alpha)\, \mathbb{I}\big(L_{ch}[\mathcal{M}(v_i)] \geq \tau\big) \end{aligned} \tag{11}$$

where $\mathrm{sim}(\cdot)$ denotes cosine similarity, $\mathbb{I}(\cdot)$ is an indicator function returning 1 if the mapped label exceeds τ and 0 otherwise, α balances semantic and diagnostic signals, and $\mathcal{M}(\cdot)$ maps graph entities to disease categories.

Top-ranked entities initialize retrieval seeds Λ_s. Starting from Λ_s, bounded breadth-first traversal retrieves evidence subgraph $\mathcal{G}_{sub} \subseteq \mathcal{G}_{med}$ to preserve clinically relevant dependencies while suppressing irrelevant evidence propagation. GraphRAG further retrieves multi-hop clinical relation paths

$$p_r = (v_a, r_1, v_b, \ldots, r_k, v_z) \tag{12}$$

where $(v_m, r_k, v_n) \in \mathcal{E}$ denotes valid graph connectivity. Retrieved paths are ranked and serialized into graph evidence context C_G. Graph-enhanced context is then obtained as:

$$R_{KI} = \mathrm{Concat}(R_I, C_G) \tag{13}$$

where R_{KI} denotes graph-enhanced context for QA-style generation. Implementation details are provided in Appendix D.

QA-Style Report Generation: We employ fine-tuned Llama-3-8B-Instruct, denoted by $\mathcal{F}_{qa}(\cdot)$, to restructure graph-enhanced report $R_{KI}^{(k)}$ into clinically meaningful QA pairs $R_{qa}^{(k)}$ and a concise summary, where k indexes the study. A predefined $\mathcal{P}rompt$ guides $\mathcal{F}_{qa}(\cdot)$ to generate target format:

PROMPT

```
Given a knowledge-enriched chest X-ray report, the
objective is to convert the report into a clinically
coherent question-answer (QA) format and generate a concise
diagnostic summary. This involves decomposing the report
into short, interpretable QA pairs, followed by
synthesizing a summary of the key findings.
```

We formalize QA generation as:

$$R_{qa}^{(k)} = \mathcal{F}_{qa}\left(R_{KI}^{(k)}, \mathcal{P}rompt\right) \tag{14}$$

The output $R_{qa}^{(k)} = \{(q_j, a_j)\}_{j=1}^{m}$ contains clinically interpretable QA pairs, where m denotes the number of QA pairs, q_j denotes a clinical question, and $a_j \in \{\mathrm{Yes}, \mathrm{No}\}$ denotes its answer, followed by a concise diagnostic summary.

2.4 Learning Objective

We train the framework by jointly optimizing the disease-aware visual-textual contrastive, disease classification, and report generation objectives. The total loss is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{Di-VTC}} + \lambda_1 \cdot \mathcal{L}_{\text{G}} + \lambda_2 \cdot \mathcal{L}_{\text{CE}} \tag{15}$$

where $\mathcal{L}_{\text{Di-VTC}}$ denotes the Disease-aware Visual–Textual Contrast loss, $\mathcal{L}_{\text{CE}}$ is the disease classification loss computed using cross-entropy between predicted and CheXbert-derived labels, and $\mathcal{L}_{\text{G}}$ represents the intermediate report generation loss. The coefficients λ_1 and λ_2 balance the contributions of the generative and classification objectives, and are empirically set to 1.0 for equal weighting.

3 Experimental Setting

Datasets: We evaluate CliQ-RRG on public datasets: (1) MIMIC-CXR (Johnson et al., 2019), the largest RRG dataset with 377,110 scans and 227,827 reports, using official train/validation/test split (Chen et al., 2020; Park et al., 2025); and (2) IU X-Ray (Demner-Fushman et al., 2015), a smaller set with 7,470 frontal and lateral scans and 3,955 reports, divided into 7:2:1 splits following prior work (Chen et al., 2020; Park et al., 2025).

Metrics: Following prior work (Jin et al., 2024; Park et al., 2025), narrative reports are evaluated using NLG metrics: BLEU (B) (Papineni et al., 2001), METEOR (MTR) (Denkowski and Lavie, 2011), ROUGE-L (RG-L) (Lin, 2004), CIDEr (C) (Vedantam et al., 2015), and BERTScore (BS) (Zhang et al., 2020a). Clinical Efficacy (CE) is assessed

Table 1: Quantitative comparison of CliQ-RRG with SOTA methods on MIMIC-CXR and IU X-Ray. †: Results from published papers; ⋆: Results reproduced from released code. **Bold**: best and underlined: second-best scores.

Type	Method	Published	MIMIC-CXR						IU X-Ray					
			B1	B4	MTR	RG-L	C	BS#	B1	B4	MTR	RG-L	C	BS#
I	R2GenRL⋆ (Qin and Song, 2022)	ACL'22	0.392	0.113	0.149	0.275	0.298	0.852	0.487	0.177	0.205	0.377	0.452	0.835
	DCL⋆ (Li et al., 2023)	CVPR'23	0.384	0.114	0.147	0.278	0.301	0.844	0.483	0.171	0.201	0.391	0.461	0.838
	MAN† (Shen et al., 2024)	AAAI'24	0.396	0.115	0.151	0.274	–	–	0.501	0.170	0.213	0.386	–	–
	CoFE⋆ (Li et al., 2024a)	ECCV'24	0.393	0.133	0.172	0.316	0.385	0.860	0.406	0.177	0.199	0.423	0.508	0.849
	SEI⋆ (Liu et al., 2024b)	MICCAI'24	0.381	0.119	0.148	0.306	0.328	0.841	0.486	0.173	0.209	0.395	0.472	0.837
	DART† (Park et al., 2025)	CVPR'25	0.437	0.137	0.175	0.310	–	–	0.486	0.208	0.205	0.411	–	–
II	KiUT† (Huang et al., 2023)	CVPR'23	0.393	0.113	0.160	0.285	–	–	0.525	0.185	**0.242**	0.409	–	–
	EKAGen† (Bu et al., 2024a)	CVPR'24	0.419	0.119	0.157	0.287	–	–	0.526	0.203	0.214	0.404	–	–
	PromptMRG† (Jin et al., 2024)	AAAI'24	0.398	0.112	0.157	0.268	–	–	0.401	0.098	0.160	0.281	–	–
	RADAR⋆ (Hou et al., 2025)	ACL'25	0.493	0.277	0.170	0.304	–	0.897	0.411	0.119	0.171	0.293	–	0.840
	REVTAF† (Zhou et al., 2025)	ICCV'25	0.465	0.182	**0.199**	**0.336**	–	–	0.420	0.107	0.176	0.309	–	–
III	R2-LLM⋆ (Liu et al., 2024a)	AAAI'24	0.411	0.132	0.179	0.288	0.354	0.873	0.487	0.178	0.215	0.401	0.485	0.846
	AdaMatch-Cyclic† (Chen et al., 2024)	ACL'24	0.379	0.101	0.163	0.286	–	–	0.416	0.145	0.162	0.366	–	–
	KARGEN† (Li et al., 2024b)	MICCAI'24	0.417	0.140	0.165	0.305	0.289	–	0.490	0.180	0.218	0.385	0.491	–
	LLM-RG4⋆ (Wang et al., 2025)	AAAI'25	0.384	0.136	0.155	0.322	0.372	0.849	0.442	0.179	0.192	0.403	0.479	0.848
	DAD-RG† (Wu et al., 2026)	AAAI'26	0.428	0.131	–	0.291	0.232	–	0.495	0.187	–	0.384	0.634	–
IV	MPO† (Xiao et al., 2025)	AAAI'25	0.416	0.139	0.162	0.309	–	–	0.548	0.209	0.224	0.415	–	–
	HC-LLM⋆ (Liu et al., 2025b)	AAAI'25	0.478	0.215	0.166	0.298	0.362	0.864	0.532	0.199	0.221	0.423	0.503	0.839
	ORID† (Gu et al., 2025)	WACV'25	0.386	0.117	0.150	0.284	–	–	0.501	0.198	0.211	0.400	–	–
V	CoD† (Jin et al., 2025)	TMI'25	0.412	0.129	–	0.286	–	–	0.403	0.091	–	0.288	–	–
VI	RA-RRG† (Park et al., 2026)	ACL'26	0.379	0.080	–	0.249	–	–	0.363	0.067	–	0.272	–	–
VII	MAIRA-2⋆ (Bannur et al., 2024)	arXiv'24	0.401	0.128	0.168	0.296	–	0.855	–	–	–	–	–	–
	MedGemma⋆ (Sellergren et al., 2025)	arXiv'25	0.385	0.118	0.159	0.281	–	0.847	0.412	0.145	0.187	0.362	–	0.831
Ours	CliQ-RRG (Graph-guided)		**0.502**	**0.283**	0.193	**0.336**	**0.396**	**0.898**	**0.565**	**0.221**	0.225	**0.438**	0.589	**0.885**

I: Contrastive-based; II: Knowledge Injected-based; III: LLM-based, IV: Multi View-based; V: QA Style-based; VI: RAG-based; VII: MLLM-based; #: BERT scores are reported for ⋆ methods using reproduced outputs; All reproduced methods use the same test set for fair comparison

Table 2: Comparison of CE metrics of CliQ-RRG on the MIMIC-CXR dataset. †: Results from published papers. **Bold**: best and underlined: second-best scores.

Method	Pre	Rec	F1
R2GenRL† (Qin and Song, 2022)	0.342	0.294	0.292
DCL† (Li et al., 2023)	0.471	0.352	0.373
MAN† (Shen et al., 2024)	0.411	0.398	0.389
CoFE† (Li et al., 2024a)	0.489	0.370	0.405
DART† (Park et al., 2025)	0.533	0.520	0.546
KiUT† (Huang et al., 2023)	0.371	0.318	0.321
EKAGen† (Bu et al., 2024a)	0.517	0.483	0.499
PromptMRG† (Jin et al., 2024)	0.501	0.509	0.476
REVTAF† (Zhou et al., 2025)	**0.628**	0.613	0.592
R2-LLM† (Liu et al., 2024a)	0.465	0.482	0.473
LLM-RG4† (Wang et al., 2025)	0.583	0.593	0.588
MPO† (Xiao et al., 2025)	0.436	0.376	0.353
ORID† (Gu et al., 2025)	0.435	0.295	0.352
CoD† (Jin et al., 2025)	0.487	0.521	0.479
CliQ-RRG (Graph-guided)	0.601	**0.621**	**0.610**
Disease Classifier	*0.496*	*0.515*	*0.505*

using CheXbert (Smit et al., 2020) F1, Precision (Pre), and Recall (Rec). For QA pairs, we follow recent studies (Manes et al., 2024; Kim et al., 2024) and evaluate semantic (BS) and clinical metrics (CheXbert-F1, RadGraph-F1 (Jain et al., 2021)). Inter-annotator agreement is measured using pairwise agreement and Fleiss' kappa (κ_F).

Implementation Details The framework is implemented in TensorFlow and trained on an NVIDIA 64 GB GPU using the AdamW optimizer (Loshchilov and Hutter, 2019). We train for 30 epochs with batch size 16, learning rate $3e{-}5$, and weight decay 0.01. Additional implementation details are provided in Appendix E.

4 Result and Discussion

Quantitative Analysis: We compare CliQ-RRG against seven method categories (Appendix B) on MIMIC-CXR and IU X-Ray. Reproduced baselines (⋆) follow their original settings and evaluation protocols, without adding prior scans or external knowledge. For fair evaluation, we consider two settings: (i) CliQ-RRG (Graph-guided), where narrative reports are evaluated using standard NLG and CE metrics; and (ii) QA-Pair (R_{qa}). Since QA outputs differ from narrative reports, conventional NLG metrics are not directly applicable. We therefore evaluate QA outputs using semantic (BERTScore) and clinical (CheXbert-F1, RadGraph-F1) metrics, following prior QA-style evaluation practice (Manes et al., 2024; Kim et al., 2024). However, because QA-style text explicitly states disease names and their assertions, these metrics may not be directly comparable to scores on narrative reports. We therefore use Expert Evaluation as the primary evidence of clinical validity for the generated QA pairs.

Table 3: Ablation analysis on MIMIC-CXR showing incremental performance gains from integrating PrAM, Di-VTC (across V_{ch}–T_{ch} and L_{ch}), and $\mathcal{G}_{med}$ into the base model. Results are reported as mean ± std over five runs; * indicates statistically significant improvement over the base model ($p \leq 0.05$).

Model	Stage 1			Stage 2	NLG Metrics					CE Metrics			
	PrAM	$V_{ch}-T_{ch}$	L_{ch}	$\mathcal{G}_{\text{med}}$	B1	B4	MTR	RG-L	∇_{N}	Pre	Rec	F1	∇_{C}
Base[x]	×	×	×	×	0.381±0.004	0.185±0.005	0.157±0.003	0.295±0.004	–	0.455	0.471	0.463	–
(a)	×	✓	×	×	0.399±0.003	0.197±0.004	0.160±0.002	0.301±0.005	3.8%	0.472	0.493	0.482	4.2%
(b)	✓	×	×	×	0.404±0.005	0.208±0.006	0.163±0.003	0.302±0.004	5.8%	0.485	0.501	0.493	6.5%
(c)	✓	✓	×	×	0.422±0.004	0.233±0.005	0.169±0.004	0.311±0.006	11.5%	0.498	0.519	0.508	9.8%
(d)	×	✓	✓	×	0.442±0.006	0.224±0.004	0.169±0.003	0.309±0.005	12.4%	0.511	0.541	0.526	13.6%
(e)	✓	✓	✓	×	0.465±0.005	0.247±0.006	0.175±0.004	0.314±0.004	18.0%	0.542	0.565	0.553	19.5%
(f)	×	✓	✓	✓	0.481±0.006	0.262±0.001	0.183±0.005	0.325±0.004	22.8%	0.563	0.592	0.577	24.6%
(g)[y]	✓	✓	✓	✓	0.502±0.003*	0.283±0.002*	0.193±0.003*	0.336±0.006*	29.0%	0.601	0.621	0.610	31.9%

PrAM: Prior-Guided Attention Module; V_{ch}: Visual Channel; T_{ch}: Textual Channel; L_{ch}: Predicted diagnostic label; $\mathcal{G}_{\text{med}}$: Graph-guided retrieval. ∇_{N} and ∇_{C}: Average improvement over the base configuration for NLG and CE metrics; x: CliQ-RRG (Base Model); y: CliQ-RRG (Graph-guided)

Comparison with SOTA: As shown in Table 1, CliQ-RRG achieves SOTA performance across most metrics on MIMIC-CXR and IU X-Ray. Compared with contrastive methods, our framework shows stronger lexical and clinical alignment with GT reports. Among knowledge-based methods, RADAR (Hou et al., 2025) remains competitive, while our graph-guided retrieval achieves superior overall performance. Among knowledge-based methods, REVTAF (Zhou et al., 2025) achieves the best MTR and second-best RG-L on MIMIC-CXR, while DAD-RG (Wu et al., 2026) performs strongly on IU X-Ray. CliQ-RRG outperforms multi-view methods, including MPO (Xiao et al., 2025) and QA-style baseline CoD (Jin et al., 2025), demonstrating the effectiveness of PrAM, channel alignment, and graph-guided retrieval. Notably, CliQ-RRG also outperforms the recent RA-RRG (Park et al., 2026) and MLLM-based MAIRA-2 (Bannur et al., 2024) and MedGemma (Sellergren et al., 2025) across the reported metrics.

CE Evaluation: Table 2 reports CE metrics on MIMIC-CXR, where CliQ-RRG achieves superior performance and outperforms REVTAF (Zhou et al., 2025) in Rec and F1. We further evaluate the disease classifier, confirming its effectiveness in guiding multimodal alignment.

QA-pair Evaluation: Table 4 reports QA-pair evaluation on MIMIC-CXR and IU X-Ray. CliQ-RRG achieves strong BERTScore, CheXbert-F1, and RadGraph-F1 performance, indicating semantic fidelity and clinical consistency in QA pairs.

Ablation Study: We analyze each component of CliQ-RRG on MIMIC-CXR (Table 3) and IU X-Ray (Appendix F). The **base** model uses a visual encoder and Transformer decoder without alignment or graph-guided retrieval. Integrating visual (V_{ch}) and textual (T_{ch}) channels **(model a)** yields marginal gains (B4: 0.197, F1: 0.482), showing the limitation of CLIP-style alignment. Adding PrAM **(model b)** improves clinical performance (F1: 0.493). Combining both **(model c)** further improves F1 to 0.508. Replacing CLIP-style alignment with Di-VTC **(model d)**, which aligns V_{ch}–T_{ch}–L_{ch}, yields stronger gains (B4: 0.224), showing the benefit of label-aware alignment. Integrating PrAM with Di-VTC **(model e)** further improves NLG and CE metrics. Adding $\mathcal{G}_{med}$ **(model f)** improves semantic completeness (F1: 0.577). The full graph-guided framework **(model g)** achieves the best performance (NLG: +29.0%, CE: +31.9%), validating that $\mathcal{G}_{med}$ complements $V_{ch}-T_{ch}-L_{ch}$ alignment. Results are reported as mean ± std over five runs, and model (g) shows statistically significant gains ($p \leq 0.05$). We further analyze LLM-based QA restructuring and $\mathcal{G}_{med}$ in Appendix G and H.

Table 4: QA-pair evaluation using semantic and clinical metrics.

Dataset	BERTScore	CheXbert-F1	RadGraph-F1
MIMIC-CXR	0.921	0.639	0.325
IU X-Ray	0.903	0.445	0.267

Qualitative Analysis; Fig. 4 illustrates how each component of CliQ-RRG progressively enhances report quality on a MIMIC-CXR sample. The base model produces generic statements such as *"no acute abnormality"*, lacking clinical detail. Adding the *PrAM* introduces precise findings (*"no evidence of consolidation"*), while the *Disease-aware Visual–Textual Contrast* identifies key observations like *"focal consolidation, pleural effusion, or pneumothorax"*. Incorporating predicted disease embeddings improves diagnostic grounding, captur-

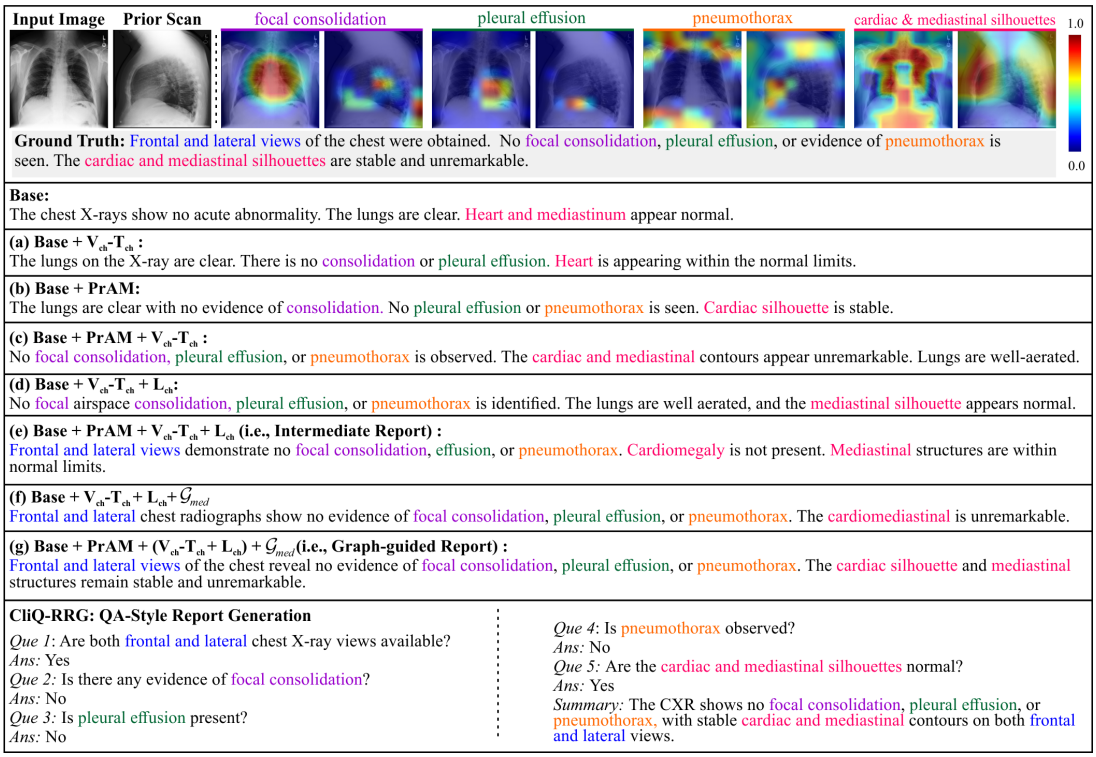


Figure 4: Qualitative analysis of CliQ-RRG on MIMIC-CXR. Report comparison across models given input and prior scans, with ground truth (GT). Shared medical terms use consistent color coding, and Grad-CAM highlights anatomically and pathologically relevant regions.

ing terms such as *"cardiomegaly"*. Graph-guided retrieval further refines report quality through clinically grounded evidence, producing findings such as *"stable and unremarkable"*, aligning more closely with the ground truth and enabling structured QA-style outputs. GradCAM (Selvaraju et al., 2017) visualizations confirm that CliQ-RRG focuses on relevant anatomical regions, validating anatomically grounded reasoning.

Parameter Sensitivity Analysis: We analyze the sensitivity of key hyperparameters λ_1, λ_2, and τ on MIMIC-CXR. As shown in Fig. 5, we vary $\lambda_1, \lambda_2 \in 0.5, 1.0, 1.5$ and $\tau \in 0.03, 0.05, 0.07, 0.10$ while keeping other parameters fixed. The method achieves optimal performance across all metrics when $\lambda_1 = \lambda_2 = 1.0$, indicating that equal weighting balances generation and classification objectives. Performance peaks at $\tau = 0.07$, suggesting an appropriate temperature for contrastive alignment.

Expert Evaluation: We conduct a two-tier clinical validation to assess the reliability of the generated QA pairs. (i) We randomly sample 1,000 QA pairs generated by CliQ-RRG and evaluate using a standard Likert-scale protocol. Three medical professionals (2 doctors and 1 medical student) independently rate each QA pair against the ground-truth

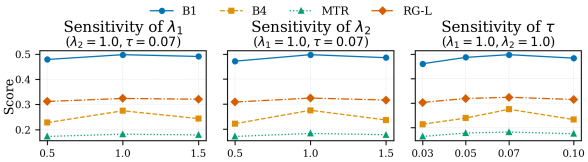


Figure 5: Parameter sensitivity analysis on MIMIC-CXR showing the effect of generation loss weight (λ_1), classification loss weight (λ_2), and temperature (τ).

report on a 5-point scale across five criteria: overall quality, consistency, clinical relevance, specificity, and fluency, following the rubric of Hashemi et al. (2024) (Table 5 (i)). We measure annotation reliability using Fleiss' kappa (κ_F) and pairwise agreement (PA). (ii) To assess clinical accuracy and hallucination, we perform an independent clinical validation with a senior radiologist with over 20 years of experience. We select 500 QA pairs using stratified random sampling across the 13 CheXbert disease categories and difficulty levels, and present them in randomized order without revealing their source. Each pair is compared with the corresponding GT report and categorized as fully acceptable, clinically acceptable, or hallucinated. As shown in Table 5 (ii), 467 pairs are fully acceptable, 31 are clinically acceptable, and 2 are hallucinated,

yielding a 99.6% clinical reliability rate and a 0.4% hallucination rate. The detailed sampling protocol, difficulty distribution, error taxonomy, and inter-rater agreement are provided in Appendix K.

Table 5: Human evaluation of generated QA pairs: (i) 5-point Likert ratings (R_0–R_4) with inter-annotator agreement; (ii) independent clinical assessment of QA integrity and hallucinations against ground truth.

	Dataset	R_0	R_1	R_2	R_3	R_4	PA	κ_F
(i)	MIMIC-CXR	4.07	4.03	4.32	4.17	4.29	0.83	0.77
	IU X-Ray	4.15	4.07	4.51	3.97	4.21	0.86	0.79

	Outcome	Count	%
(ii)	Fully acceptable	467	93.4%
	Clinically acceptable (minor linguistic deviations)	31	6.2%
	Hallucinated (contradictory/unsupported findings)	2	0.4%
	Total Reliable (Fully + Clinically) Pairs	**498**	**99.6%**

Generalizability Analysis: To assess generalization under distribution shift, we evaluate CliQ-RRG zero-shot on the held-out ReXGradient-160K (Zhang et al., 2025b), a large-scale multi-institutional CXR dataset spanning 3 health systems and 79 medical sites. Neither training nor fine-tuning is performed on this dataset. Table 6 reports Stage 1 results, while Table 7 reports the final QA-pair performance. Models trained on MIMIC-CXR consistently outperform those trained on IU X-Ray, reflecting the benefit of larger and more diverse training data. Although surface-level metrics show a moderate decrease under distribution shift, semantic and clinical metrics remain relatively stable, indicating that CliQ-RRG preserves clinical information across institutions and devices. This evaluation supports cross-hospital and cross-device generalization, while extension to other anatomical regions and modalities remains future work.

Table 6: Zero-shot generalization of Stage 1 on ReXGradient-160K.

Train	Test	B1	B4	MTR	RG-L	C	BS
MIMIC-CXR	ReXGradient	0.441	0.221	0.161	0.288	0.312	0.861
IU X-Ray	ReXGradient	0.398	0.165	0.148	0.259	0.271	0.831

Table 7: Zero-shot QA-pair evaluation on ReXGradient-160K.

Train	Test	BERTScore	CheXbert-F1	RadGraph-F1
MIMIC-CXR	ReXGradient	0.897	0.581	0.289
IU X-Ray	ReXGradient	0.869	0.502	0.243

5 Related Work

Due to space constraints in the main text, we provide related work in Appendix A.

6 Conclusion

In this paper, we present CliQ-RRG for QA-style radiology report generation. CliQ-RRG introduces a disease-aware visual–textual contrastive scheme to align visual, textual, and predicted diagnostic labels, enhanced by prior-guided attention for integrating multi-view chest X-rays. Additionally, graph-guided retrieval incorporates clinically connected evidence to improve report quality and support structured QA generation. Qualitative analysis shows that CliQ-RRG effectively aligns visual, textual, and label modalities to generate clinically reliable QA-style radiology reports. Experiments on MIMIC-CXR and IU X-Ray benchmarks highlight the superiority of our framework over state-of-the-art methods.

7 Limitations

While CliQ-RRG demonstrates strong performance, several limitations suggest directions for future improvement. First, the current study focuses on 2D CXR modalities, which enable a controlled analysis of disease-aware alignment and QA-style reporting but do not cover modalities with richer spatial structure. Second, the large language model is used as a fixed restructuring component to convert graph-guided reports into structured yes or no QA pairs, rather than being jointly optimized for multimodal reasoning. Third, training is limited to MIMIC-CXR and IU X-Ray, which may constrain the learned representations to their specific acquisition protocols and population characteristics. Fourth, although PrAM incorporates historical scans, it does not explicitly encode temporal order or time intervals between examinations and therefore does not directly model longitudinal progression patterns. Finally, the framework relies on supervised learning with paired images, reports, and diagnostic labels, which may constrain scalability in data-scarce settings.

References

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*.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. 2022. Making the most of text semantics to improve biomedical vision-language processing. In *European conference on computer vision*, volume 13696 LNCS, pages 1–21.

Shenshen Bu, Taiji Li, and Zhiming Dai. 2023. Enhancing medical report generation in multi-slice fusion scenarios. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, page 1030–1037. IEEE.

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

Shenshen Bu, Yujie Song, Taiji Li, and Zhiming Dai. 2024b. Dynamic knowledge prompt for chest x-ray report generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5425–5436.

Wenting Chen, Linlin Shen, Jingyang Lin, Jiebo Luo, Xiang Li, and Yixuan Yuan. 2024. Fine-grained image-text alignment in medical imaging enables explainable cyclic image-report generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 9494–9509. Association for Computational Linguistics.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2021. Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics.

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

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 *WMT 2011 - 6thWorkshop on Statistical Machine Translation, Proceedings of the Workshop*, pages 85–91.

Tiancheng Gu, Dongnan Liu, Zhiyuan Li, and Weidong Cai. 2024. Complex organ mask guided radiology report generation. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, page 7980–7989. IEEE.

Tiancheng Gu, Kaicheng Yang, Xiang An, Ziyong Feng, Dongnan Lin, and Weidong Cai. 2025. Orid: Organ-regional information driven framework for radiology report generation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, page 378–387. IEEE.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. 2024. Llm-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 13806–13834. Association for Computational Linguistics.

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 (Volume 1: Long Papers)*, pages 26366–26381. Association for Computational Linguistics.

Wenjun Hou, Kaishuai Xu, Yi Cheng, Wenjie Li, and Jiang Liu. 2023. Organ: Observation-guided radiology report generation via tree reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Xiyang Huang, Yingjie Han, Yaoxu Li, Runzhi Li, Pengcheng Wu, and Kunli Zhang. 2025. CmEAA: Cross-modal enhancement and alignment adapter for radiology report generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8546–8556. Association for Computational Linguistics.

Zhongzhen Huang, Xiaofan Zhang, and Shaoting Zhang. 2023. Kiut: Knowledge-injected u-transformer for radiology report generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 19809–19818. IEEE.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Matthew Lungren, Andrew Ng, Curtis Langlotz, and Pranav Rajpurkar. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. In *Advances in Neural Information Processing Systems 35*, NeurIPS 2021, page 1–12. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Haibo Jin, Haoxuan Che, Sunan He, and Hao Chen. 2025. A chain of diagnosis framework for accurate and explainable radiology report generation. *IEEE Transactions on Medical Imaging*, 44(12):4986–4997.

Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. 2024. Promptmrg: Diagnosis-driven prompts for medical report generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(3):2607–2615.

Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. 2019. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *Preprint*, arXiv:1901.07042.

Gregory Kell, Angus Roberts, Serge Umansky, Linglong Qian, Davide Ferrari, Frank Soboczenski, Byron C Wallace, Nikhil Patel, and Iain J Marshall. 2024. Question answering systems for health professionals at the point of care—a systematic review. *Journal of the American Medical Informatics Association*, 31(4):1009–1024.

Yunsoo Kim, Jinge Wu, Yusuf Abdulle, and Honghan Wu. 2024. Medexqa: Medical question answering benchmark with multiple explanations. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, page 167–181. Association for Computational Linguistics.

Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. 2023. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 3334–3343. IEEE.

Mingjie Li, Haokun Lin, Liang Qiu, Xiaodan Liang, Ling Chen, Abdulmotaleb Elsaddik, and Xiaojun Chang. 2024a. Contrastive learning with counterfactual explanations for radiology report generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 162–180. Springer Nature Switzerland.

Yingshu Li, Zhanyu Wang, Yunyi Liu, Lei Wang, Lingqiao Liu, and Luping Zhou. 2024b. *KARGEN: Knowledge-Enhanced Automated Radiology Report Generation Using Large Language Models*, page 382–392. Springer Nature Switzerland.

Xiao Liang, Yanlei Zhang, Di Wang, Haodi Zhong, Ronghan Li, and Quan Wang. 2024. Divide and conquer: Isolating normal-abnormal attributes in knowledge graph-enhanced radiology report generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 4967–4975. ACM.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Qika Lin, Kai He, Yifan Zhu, Fangzhi Xu, Erik Cambria, and Mengling Feng. 2025. Cross-modal knowledge diffusion-based generation for difference-aware medical vqa. *IEEE Transactions on Image Processing*, 34:2421–2434.

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.

Kang Liu, Zhuoqi Ma, Xiaolu Kang, Yunan Li, Kun Xie, Zhicheng Jiao, and Qiguang Miao. 2025a. Enhanced contrastive learning with multi-view longitudinal data for chest x-ray report generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 10348–10359. IEEE.

Kang Liu, Zhuoqi Ma, Xiaolu Kang, Zhusi Zhong, Zhicheng Jiao, Grayson Baird, Harrison Bai, and Qiguang Miao. 2024b. *Structural Entities Extraction and Patient Indications Incorporation for Chest X-Ray Report Generation*, page 433–443. Springer Nature Switzerland.

Kang Liu, Zhuoqi Ma, Mengmeng Liu, Zhicheng Jiao, Xiaolu Kang, Qiguang Miao, and Kun Xie. 2026. Factual serialization enhancement: A key innovation for chest x-ray report generation. *Expert Systems with Applications*, 324:132550.

Tengfei Liu, Jiapu Wang, Yongli Hu, Mingjie Li, Junfei Yi, Xiaojun Chang, Junbin Gao, and Baocai Yin. 2025b. Hc-llm: Historical-constrained large language models for radiology report generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(6):5595–5603.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019*.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Yuanjiang Luo, Hongxiang Li, Xuan Wu, Meng Cao, Xiaoshuang Huang, Zhihong Zhu, Peixi Liao, Hu Chen, and Yi Zhang. 2024. *Textual Inversion and Self-supervised Refinement for Radiology Report Generation*, page 681–691. Springer Nature Switzerland.

Itay Manes, Naama Ronn, David Cohen, Ran Ilan Ber, Zehavi Horowitz-Kugler, and Gabriel Stanovsky. 2024. K-qa: A real-world medical q& a benchmark. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, page 277–294. Association for Computational Linguistics.

Luis-Jesus Marhuenda, Miquel Obrador-Reina, Mohamed Aas-Alas, Alberto Albiol, and Roberto Paredes. 2025. Unveiling differences: A vision encoder-decoder model for difference medical visual question

answering. In *Medical Imaging with Deep Learning*, pages 1095–1106.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, ACL '02, pages 311–318. Association for Computational Linguistics.

Jonggwon Park, Byungmu Yoon, Soobum Kim, and Kyoyun Choi. 2026. Ra-rrg: Multimodal retrieval-augmented radiology report generation with key phrase extraction. In *Findings of the Association for Computational Linguistics: ACL 2026*, page 5029–5048. Association for Computational Linguistics.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 15580–15589. IEEE.

Priyaranjan Pattnayak, Hitesh Patel, Amit Agarwal, Srikant Panda, Bhargava Kumar, and Tejaswini Kumar. 2025. Clinical qa 2.0- multi-task learning for answer extraction and categorization. In *2025 IEEE International Conference on Electro Information Technology (eIT)*, page 1–7. IEEE.

Han Qin and Yan Song. 2022. Reinforced cross-modal alignment for radiology report generation. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics.

Lawrence H. Schwartz, David M. Panicek, Alexandra R. Berk, Yuelin Li, and Hedvig Hricak. 2011. Improving communication of diagnostic radiology findings through structured reporting. *Radiology*, 260(1):174–181.

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*.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, page 618–626. IEEE.

Hongyu Shen, Mingtao Pei, Juncai Liu, and Zhaoxing Tian. 2024. Automatic radiology reports generation via memory alignment network. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(5):4776–4783.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. 2020. Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

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)*, page 643–655. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 2017-December.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 4566–4575. IEEE.

Zhuhao Wang, Yihua Sun, Zihan Li, Xuan Yang, Fang Chen, and Hongen Liao. 2025. Llm-rg4: Flexible and factual radiology report generation across diverse input contexts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(8):8250–8258.

Puzhen Wu, Hexin Dong, Yi Lin, Yihao Ding, and Yifan Peng. 2026. A disease-aware dual-stage framework for chest x-ray report generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(40):33953–33961.

Ting Xiao, Lei Shi, Peng Liu, Zhe Wang, and Chenjia Bai. 2025. Radiology report generation via multi-objective preference optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(8):8664–8672.

Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *32nd International Conference on Machine Learning, ICML 2015*, volume 3, pages 2048–2057.

Benjamin Yan, Ruochen Liu, David Kuo, Subathra Adithan, Eduardo Reis, Stephen Kwak, Vasantha Venugopal, Chloe O'Connell, Agustina Saenz, Pranav Rajpurkar, and Michael Moor. 2023. Style-aware radiology report generation with radgraph and few-shot prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, page 14676–14688. Association for Computational Linguistics.

Bin Yan and Mingtao Pei. 2022. Clinical-bert: Vision-language pre-training for radiograph diagnosis and reports generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3):2982–2990.

Shuxin Yang, Xian Wu, Shen Ge, Zhuozhao Zheng, S Kevin Zhou, and Li Xiao. 2023. Radiology report generation with a learned knowledge base and multi-modal alignment. *Medical Image Analysis*, 86:102798.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, and 5 others. 2025a. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1).

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. Bertscore: Evaluating text generation with bert. In *8th International Conference on Learning Representations, ICLR 2020*.

Xiaoman Zhang, Julián N. Acosta, Josh Miller, Ouwen Huang, and Pranav Rajpurkar. 2025b. Rexgradient-160k: A large-scale publicly available dataset of chest radiographs with free-text reports. In *Machine Learning for Health (ML4H) 2025*.

Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. 2020b. When radiology report generation meets knowledge graph. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12910–12917.

Qin Zhou, Guoyan Liang, Xindi Li, Jingyuan Chen, Zhe Wang, Chang Yao, and Sai Wu. 2025. Learnable retrieval enhanced visual-text alignment and fusion for radiology report generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, page 22529–22538. IEEE.

Appendix

Related Work ..... Appendix A
Compared Baselines ..... Appendix B
Knowledge Base Construction ..... Appendix C
Graph-guided Retrieval Details ..... Appendix D
Implementation Details ..... Appendix E
Ablation Results (IU dataset) ..... Appendix F
LLM Choice on QA ..... Appendix G
Impact of Graph-retrival ..... Appendix H
Per-State Disease Classification ..... Appendix I
Noise-injection Sensitivity ..... Appendix J
Expanded Expert Evaluation ..... Appendix K
Computational Efficiency ..... Appendix L
Practical Implication ..... Appendix M
Future Work ..... Appendix N
Reproducibility ..... Appendix O

A Related Work

Encoder–Decoder and Multi-View RRG: Early RRG methods rely on encoder-decoder architectures to generate natural language descriptions from visual inputs (Xu et al., 2015; Lu et al., 2017; Chen et al., 2020). To better capture spatial and anatomical details, recent models incorporate multi-view X-rays and multi-slice features (Chen et al., 2021; Bu et al., 2023). For longitudinal tracking, HC-LLM (Liu et al., 2025b) uses historical constraints to guide large language models in generating progression-aware reports.

Contrastive Alignment: Multimodal contrastive learning is widely used to improve visual-textual alignment, ranging from global image-report matching to fine-grained region-word pairs (Li et al., 2023; Shen et al., 2024; Huang et al., 2025). Models like PromptMRG (Jin et al., 2024) use diagnostic labels as soft prompts to guide generation, while other methods integrate learned knowledge bases (Yang et al., 2023). AdaMatch-Cyclic (Chen et al., 2024) applies fine-grained cyclic alignment between image regions and text.

Knowledge-Injected Generation: Recent works inject external clinical knowledge into the generation process via attention mechanisms and knowledge graphs (Hou et al., 2025; Sun et al., 2025; Liu et al., 2024a). Notably, REVTAF (Zhou et al., 2025) fuses modal alignment and knowledge injection using a learnable retrieval enhancer in hyperbolic space and optimal transport-based cross-attention. Similarly, KiUT (Huang et al., 2023) integrates symptom graphs into a U-Transformer. Other approaches, like EKAGen (Bu et al., 2024a), retrieve instance-level expert knowledge and highlight key pathological regions, while DKP (Bu et al., 2024b) generates dynamic knowledge prompts from anomaly-driven features.

QA-Style Generation: Motivated by limitations of narrative reporting (Jin et al., 2024; Luo et al., 2024; Wang et al., 2025; Kell et al., 2024; Pattnayak et al., 2025; Schwartz et al., 2011), limited clinical grounding (Yan et al., 2023; Bu et al., 2024a; Gu et al., 2024), underuse of diagnostic labels (Li et al., 2023; Shen et al., 2024; Xiao et al., 2025), and limited temporal fusion (Hou et al., 2023; Gu et al., 2025), recent work explores structured and explainable outputs. For example, CoD (Jin et al., 2025) prompts LLMs using findings extracted from diagnostic QA pairs. In medical VQA, MENDER (Lin et al., 2025) leverages cross-modal knowledge

diffusion to achieve accurate answering. VED (Marhuenda et al., 2025) uses a vision encoder-decoder to detect and explain radiological changes across longitudinal X-rays.

Summary: Despite progress in alignment, knowledge injection, and QA formatting, prior works largely overlook longitudinal priors for structured reporting. CliQ-RRG addresses this by unifying disease-aware contrastive alignment with staged knowledge retrieval. Finally, it uses multi-view priors and employs an LLM to restructure standard narrative reports into precise, clinically meaningful QA-style outputs.

B Compared Baselines

We benchmark our proposed CliQ-RRG against state-of-the-art baselines, which we classify into seven categories.

I: Contrastive-based Methods

- *R2GenRL* (Qin and Song, 2022) optimizes the mapping between visual regions and textual words using reinforcement learning based on generation metrics.
- *DCL* (Li et al., 2023) constructs dynamic relation graphs and applies contrastive learning to align visual features with medical entities.
- *MAN* (Shen et al., 2024) uses a shared memory mechanism to capture cross-modal correspondence and guide attention during decoding.
- *CoFE* (Li et al., 2024a) aligns representations via contrastive learning that maximize similarity for factual pairs while repelling negatives.
- *SEI* (Liu et al., 2024b) aligns extracted anatomical entities and patient indications with corresponding visual regions.
- *DART* (Park et al., 2025) leverages diagnostic labels to ground text generation and incorporates a self-correcting mechanism to refine image-text consistency.

II: Knowledge Injected-based Methods

- *KiUT* (Huang et al., 2023) integrates clinical knowledge into a U-Transformer architecture using a symptom graph and an adaptive distiller to guide word prediction.
- *EKAGen* (Bu et al., 2024a) enhances generation by combining expert knowledge with discriminative attention mechanism to focus on pathological regions.
- *PromptMRG* (Jin et al., 2024) converts disease predictions into soft prompts and retrieves similar reports as in-context guidance.
- *RADAR* (Hou et al., 2025) injects external clinical knowledge and aligns it with visual features to enhance report reliability.
- *REVTAF* (Zhou et al., 2025) utilizes semantic hierarchy in hyperbolic space to get reference reports to improve clinical findings.

III: LLM-based Methods

- *R2-LLM* (Liu et al., 2024a) frames report generation as instruction following using a frozen LLM guided by visual prompts.
- *AdaMatch-Cyclic* (Chen et al., 2024) employs a cyclic framework with adaptive patch–word alignment to guide both report generation and image synthesis.
- *KARGEN* (Li et al., 2024b) integrates a medical knowledge graph with a frozen LLM to generate disease-sensitive reports.
- *LLM-RG4* (Wang et al., 2025) supports diverse input scenarios through adaptive token fusion and loss reweighting to reduce hallucinations.
- *DAD-RG* (Wu et al., 2026) utilizes a dual-attention framework that balances global radiological context with region-specific anatomical features to improve narrative consistency.

IV: Multi View-based Methods

- *MPO* (Xiao et al., 2025) adapts report generation to different user preferences by optimizing weighted objectives through reinforcement learning.
- *HC-LLM* (Liu et al., 2025b) employs historical constraints to improve progression-aware report generation with large language models.
- *ORID* (Gu et al., 2025) filters irrelevant noise using an instruction-tuned LLaVA-Med model to generate organ-specific descriptions prioritized by graph-based analysis.

V: QA Style-based Method

- *CoD* (Jin et al., 2025) improves clinical accuracy by generating QA pairs through a diagnostic conversation framework to guide a large language model during RRG.

VI: RAG-based Method

- *RA-RRG* (Park et al., 2026) is a retrieval-augmented framework that extracts clinically essential key phrases and retrieves image-consistent phrases for hallucination-suppressed radiology report generation.

VII: MLLM-based Methods

- *MAIRA-2* (Bannur et al., 2024) is a radiology-specific multimodal model that generates grounded and non-grounded chest X-ray reports using current and prior images, reports, and clinical context.
- *MedGemma* (Sellergren et al., 2025) is a medical vision-language foundation model for interpreting medical images and text, with strong performance across medical VQA and chest X-ray report generation.

C Medical Knowledge Base Construction

We construct an external clinical repository $\mathcal{K}_{ext} = \{K_m\}_{m=1}^{M}$ to support GraphRAG-based retrieval in STAGE 2.

Step 1: Source Collection: We curate approximately 5,000 medical documents from PubMed abstracts related to thoracic imaging and chest radiology. Documents are filtered to retain clinically relevant evidence describing disease manifestations, imaging findings, and anatomical structures.

Step 2: Clinical Fact Extraction: Raw text is processed using BioClinicalBERT-based biomedical NER to identify radiology-relevant entities including diseases, symptoms, anatomical regions, and imaging observations. Sentences containing clinical entities are normalized into atomic knowledge units by removing redundant modifiers and standardizing terminology. Near-duplicate facts are filtered using cosine similarity thresholding.

Step 3: Disease Label Association: Each knowledge entry K_m is associated with disease category $y_m \in \mathcal{Y}$ to maintain consistency with predicted diagnostic labels L_{ch} obtained in STAGE 1.

Step 4: Embedding Initialization: Knowledge units are embedded using BioWordVec to construct a searchable semantic space. The resulting embeddings support downstream graph grounding and retrieval.

D Graph-guided Retrieval Details

(i) Medical Graph Construction: GraphRAG is implemented using a medical knowledge graph $\mathcal{G}_{med}$ constructed from approximately 5,000 domain-specific medical summaries and clinical resources described in Appendix C. Graph entities represent diseases, anatomical structures, imaging findings, and clinical observations. Directed edges encode semantic medical relations $\mathcal{R}$, including `located_in`, `associated_with`, `manifestation_of`, `suggestive_of`, and `coexists_with`. Graph entities are pre-computed using BioWordVec embeddings to support downstream graph retrieval. Graph storage and traversal are implemented using Neo4j.

(ii) Entity Grounding: During inference, intermediate report R_I is encoded into report representation ϕ_{R_I}, while predicted disease labels L_{ch} provide diagnostic constraints. Entity grounding evaluates retrieval relevance using Eq. 11. For diagnostic supervision, graph entities are mapped into the disease label space through a mapping operator $\mathcal{M} : \mathcal{V} \rightarrow \{1, 2, \ldots, |\mathcal{Y}|\}$, where $\mathcal{V}$ denotes graph entities and $\mathcal{Y}$ denotes diagnostic categories. Disease entities are mapped directly, while anatomical findings and imaging observations inherit labels through graph connectivity:

$$\mathcal{M}(v_i) = \arg \max_{m \in \{1,\ldots,|\mathcal{Y}|\}} \text{sim}(\phi_{v_i}, \phi_{y_m}) \quad (16)$$

where ϕ_{v_i} denotes graph entity embedding and ϕ_{y_m} denotes the embedding of disease category $y_m \in \mathcal{Y}$. Threshold $\tau \in [0, 1]$ filters low-confidence disease predictions in L_{ch}, allowing only reliable diagnostic information to contribute to retrieval scoring. Top-k ranked entities initialize retrieval seed set Λ_s, enabling graph retrieval using both report semantics and predicted disease labels.

(iii) Subgraph Retrieval and Relation-path Discovery: Starting from Λ_s, bounded breadth-first traversal retrieves evidence subgraph $\mathcal{G}_{sub}$ while suppressing irrelevant evidence propagation. Traversal depth is restricted to two hops to preserve clinically relevant dependencies. Example relation paths include:

- `Pleural Effusion` → `manifestation_of` → `Costophrenic Angle Blunting`
- `Opacity` → `located_in` → `Left Lower Lobe`
- `Cardiomegaly` → `associated_with` → `Enlarged Cardiac Silhouette`

(iv) Graph Context Serialization: Retrieved relation paths are ranked using retrieval relevance and diagnostic consistency. High-confidence paths are serialized into graph evidence context C_G using structural delimiters while preserving relation

ordering and graph connectivity. Serialized graph evidence is appended to intermediate report representation R_I to obtain graph-enhanced context R_{KI} for downstream QA-style generation.

E Implementation Details

Input CXRs are resized and cropped to 224×224 pixels. In STAGE 1, visual and textual encoder outputs are projected to $d_e = 768$ through a linear layer, while STAGE 2 utilizes a Transformer (Vaswani et al., 2017) decoder with 8 attention heads and a hidden size of 256. The generation length for R_{KI} is capped at 100 tokens. For GraphRAG retrieval, we construct a medical knowledge graph aligned with predicted disease labels (Appendix C). Graph entities are initialized using BioWordVec embeddings and stored within a Neo4j graph database. During inference, grounded entities initialize graph retrieval, where traversal is restricted to two hops for evidence subgraph extraction. We use top-$k_t = 10$ retrieval seeds for graph evidence retrieval.[1] QA restructuring is performed using Llama-3-8B-Instruct adapted through QLoRA fine-tuning with 4-bit NF4 quantization, LoRA rank $r = 16$, scaling factor $\alpha = 32$, and dropout 0.05. Fine-tuning is performed using AdamW with a learning rate of 2×10^{-4} for 3 epochs. All experiments were supported by an Intel(R) Xeon(R) Silver 4316 CPU @ 2.30 GHz with 512 GB RAM, and inference was performed with $m = 5$.

F Ablation Analysis on IU Dataset

We further analyze each component on IU X-Ray (Table 8). The **base** encoder–decoder achieves a B1 score of 0.435. Integrating visual (V_{ch}) and textual (T_{ch}) channels in **model (a)** improves average performance by 5.4%, showing the limitation of standard visual–textual alignment. Adding PrAM in **model (b)** yields an 8.2% gain, highlighting the benefit of multi-view context. Combining both in **model (c)** further improves performance. Replacing standard alignment with Di-VTC in **model (d)**, which aligns V_{ch}–T_{ch}–L_{ch}, improves B1 to 0.498, demonstrating stronger label-aware grounding. Integrating PrAM with Di-VTC in **model (e)** further improves NLG performance. Adding $\mathcal{G}_{med}$ in **model (f)** enhances clinical completeness. The full graph-guided framework **(model g)** achieves the best performance (+27.1%), confirming the effectiveness of graph-guided retrieval. These observations remain consistent with MIMIC-CXR and demonstrate the robustness of CliQ-RRG.

[1] We tested top-$k_t \in [1, 15]$ and selected 10 based on validation performance.

Table 8: Ablation analysis on IU X-Ray showing incremental gains from integrating PrAM, Di-VTC (across V_{ch}–T_{ch} and L_{ch}), and graph-guided retrieval ($\mathcal{G}_{med}$) into the base model.

Model	STAGE 1			STAGE 2	NLG Metrics				
	PrAM	$V_{ch}-T_{ch}$	L_{ch}	$\mathcal{G}_{med}$	B1	B4	MTR	RG-L	∇_N
Base	×	×	×	×	0.435	0.144	0.174	0.387	–
(a)	×	✓	×	×	0.452	0.168	0.184	0.398	5.4%
(b)	✓	×	×	×	0.468	0.165	0.196	0.404	8.2%
(c)	✓	✓	×	×	0.489	0.194	0.205	0.418	14.6%
(d)	×	✓	✓	×	0.498	0.188	0.201	0.411	13.9%
(e)	✓	✓	✓	×	0.531	0.207	0.213	0.425	20.7%
(f)	×	✓	✓	✓	0.551	0.217	0.221	0.432	24.6%
(g)	✓	✓	✓	✓	**0.565**	**0.221**	0.225	0.438	27.1%

PrAM: Prior-Guided Attention Module; V_{ch}: Visual Channel; T_{ch}: Textual Channel; L_{ch}: Predicted diagnostic label; $\mathcal{G}_{med}$: Graph-guided retrieval. ∇_N: Average improvement across NLG metrics over the base model.

G Impact of LLMs on QA Restructuring

To assess QA restructuring, Fig. 6 compares multiple language models across prompting and fine-tuning settings using RadGraph-F1, CheXbert-F1, and BERTScore. Few-shot prompting ($k = 5$) improves over zero-shot settings across different models. Fine-tuned Llama-3-8B-Instruct with QLoRA achieves the best overall performance, outperforming larger closed-source models such as GPT-5 and Gemini-3 Pro. These findings suggest that lightweight task adaptation is more effective than prompt design alone for preserving clinical entities and semantic alignment.

H Impact of Graph-retrieval

We analyze the contribution of graph-guided retrieval (see Table 9) through retrieval seed initialization (Λ_s), relation-path retrieval (p_r), graph serialization (C_G), and traversal depth (Hop). Removing Λ_s causes the largest performance drop, reducing CheXbert-F1 from 0.610 to 0.591, highlighting the importance of grounded retrieval initialization. Removing C_G limits sequential coherence also degrades performance, resulting in a severe drop in RG-L to 0.299 indicating that structured graph evidence improves report generation quality. Traversal analysis shows that 2-hop retrieval achieves the best performance, while 1-hop limits evidence coverage and 3-hop introduces irrelevant anatomical associations that degrade generation quality across all lexical and clinical metrics.

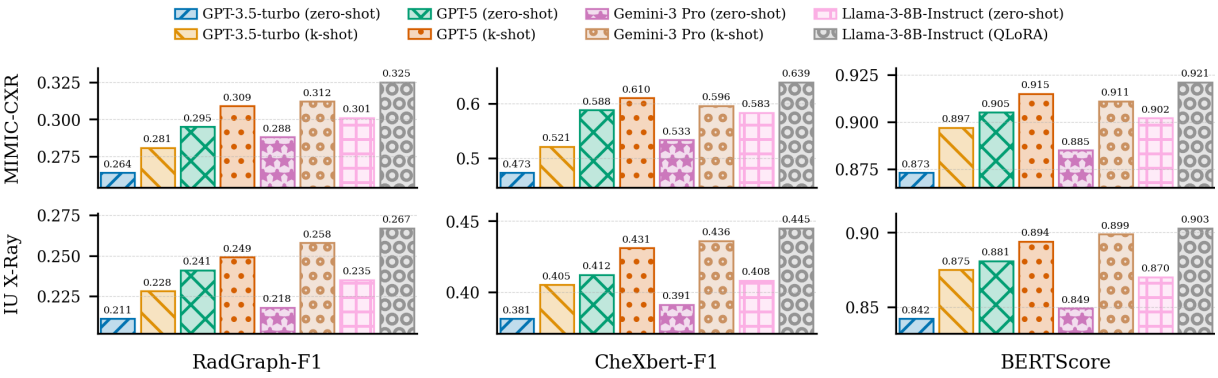


Figure 6: Comparison of LLMs for QA restructuring under zero-shot, few-shot, and QLoRA fine-tuned settings on MIMIC-CXR and IU X-Ray using RadGraph-F1, CheXbert-F1, and BERTScore.

Table 9: Impact of graph-guided retrieval on MIMIC-CXR. Higher is better.

Setting	Λ_s	p_r	C_G	Hop	B4	RG-L	F1
w/o Λ_s	×	✓	✓	2	0.242	0.311	0.548
w/o p_r	✓	×	✓	2	0.259	0.318	0.574
w/o C_G	✓	✓	×	2	0.248	0.299	0.561
1-hop	✓	✓	✓	1	0.270	0.324	0.589
3-hop	✓	✓	✓	3	0.276	0.329	0.595
2-hop (Full)	✓	✓	✓	2	**0.283**	**0.336**	**0.610**

Λ_s: retrieval seed initialization; p_r: relation-path retrieval; C_G: graph context serialization

I Per-State Disease Classification

To verify that the disease classifier captures the four-state labeling scheme used by CheXbert (Smit et al., 2020), we report its per-state performance on MIMIC-CXR in Table 10. For each of the 13 disease categories, the classifier independently predicts one of four states: positive, negative, uncertain, or not mentioned. The reported values are averaged across the 13 disease categories. The classifier achieves a macro-average F1 of 0.505, consistent with the overall disease-classification F1 reported in the main paper. Performance is higher for the positive and negative states, while the uncertain state is comparatively more challenging. This indicates that the classifier learns distinctions across the four CheXbert states rather than reducing the task to binary disease prediction.

Table 10: Per-state disease classification performance on MIMIC-CXR.

State	Precision	Recall	F1	Support (%)
Positive	0.541	0.558	0.549	28.4
Negative	0.612	0.634	0.623	41.2
Uncertain	0.389	0.402	0.395	12.6
Not Mentioned	0.442	0.466	0.454	17.8
Macro-average	**0.496**	**0.515**	**0.505**	**100.0**

J Noise-injection Sensitivity

To examine the sensitivity of Di-VTC to errors in the predicted disease labels L_{ch}, we conduct a controlled noise-injection study on MIMIC-CXR. During Stage 1 training, we synthetically corrupt the labels used to construct positive and negative pairs at rates of $0\%, 10\%, 20\%$, and 30%, while keeping all other settings fixed. We then evaluate the resulting classifier, intermediate report generation, and final QA-pair quality. As shown in Table 11, increasing label noise consistently degrades performance across all stages. At 20% noise, B4 and QA-pair RadGraph-F1 decrease by 10.95% and 10.46%, respectively, relative to the noise-free setting. At 30% noise, the degradation becomes more pronounced, with B4 and RadGraph-F1 decreasing to 0.221 and 0.256, respectively. These results indicate that Di-VTC is moderately tolerant to label noise at low levels but becomes increasingly sensitive as classifier errors accumulate. Since the reported disease classifier achieves an F1 score of 0.505, improving disease prediction remains an important direction for enhancing the reliability of disease-aware alignment and downstream generation.

Table 11: Sensitivity of Di-VTC and downstream generation to injected disease-label noise on MIMIC-CXR.

Noise (%)	F1	B4	RG-L	RadGraph-F1
30 %	0.402	0.221	0.278	0.256
20 %	0.446	0.252	0.309	0.291
10 %	0.481	0.271	0.327	0.314
0 % (Ours)	**0.505**	**0.283**	**0.336**	**0.325**

K Expanded Expert Evaluation

To further assess the clinical reliability of the generated QA pairs, we expanded the evaluation to 500 QA pairs from MIMIC-CXR. The samples were selected using stratified random sampling across the 13 CheXbert disease categories, with balanced representation of positive, negative, and uncertain findings, and were further grouped by difficulty (Table 12). The QA pairs were presented in randomized order without revealing their source. Table 13 summarizes the observed errors, with most involving minor linguistic deviations and only two cases (0.4%) containing hallucinated findings. Inter-rater agreement is reported in Table 5, with a pairwise agreement of 0.83 and Fleiss' κ of 0.77, indicating substantial agreement and supporting the reliability of the clinical assessment.

Table 12: Difficulty distribution of the 500 QA pairs sampled for expert evaluation.

Difficulty	Definition	Count	(%)
Easy	Single, unambiguous finding	233	46.6
Moderate	Single subtle/borderline finding	187	37.4
Hard	Multiple findings or rare pathology	80	16.0
Total		**500**	**100.0**

Table 13: Error types identified among the non-fully-acceptable QA pairs.

Error Type	Count	Errors (%)	Total (%)
Minor linguistic deviation	21	63.6	4.2
Missing finding	6	18.2	1.2
Wrong laterality	1	3.0	0.2
Incorrect severity	1	3.0	0.2
Temporal inconsistency	2	6.1	0.4
Hallucinated finding	2	6.1	0.4
Total	**33**	**100.0**	**6.6**

L Computational Efficiency

We analyze the computational efficiency of CliQ-RRG on a single NVIDIA 64 GB GPU, as summarized in Table 14. The framework consists of a two-stage architecture with $\sim$ 170M trainable parameters, including STAGE 1 multimodal alignment and STAGE 2 graph-guided report generation. The average inference time is 180–300 ms per study, excluding the final QA restructuring step. Graph-guided retrieval introduces minimal overhead through lightweight entity grounding and graph traversal operations. QA restructuring is performed independently using QLoRA fine-tuned Llama-3-8B-Instruct, introducing an additional latency of 1.35–2.75 s per study.

Table 14: Computational efficiency of CliQ-RRG.

Method	Parameters	Inference Time
CliQ-RRG (Graph-guided)	$\sim$ 170M	180–300 ms / study*
CliQ-RRG (QA Pair)		1.35–2.75 s / study*

$*$ Reported for single-scan and prior-scan settings.

M Practical Implication

CliQ-RRG offers several practical benefits for real-world clinical deployment and research use. First, the QA-style report formulation converts free-text narratives into a structured format that is easier to read, review, and query, which supports faster clinical decision making. Second, the disease-aware design ties each reported finding to explicit diagnostic evidence, thereby improving report consistency and enabling clinicians to verify conclusions with greater confidence. Third, the use of prior examinations supports longitudinal assessment and clearer tracking of disease changes across visits. Fourth, the integration of clinical knowledge enriches reports with relevant medical context, improving coverage of main findings without increasing text density. Finally, CliQ-RRG streamlines the radiology workflow by generating structured and reliable reports that clinicians can directly utilize.

N Future Work

Future work will focus on extending the framework along several methodological and practical dimensions. First, extending the framework toward Open-Set VQA generation is a promising direction for future work. Second, we will extend the framework to additional imaging modalities, including 3D data such as CT and MRI, to assess its applicability to volumetric and anatomically complex scenarios. We will also investigate explicit temporal modeling, such as time-interval embeddings, temporally ordered attention, or recurrent architectures, to better capture longitudinal progression across historical examinations. Third, we plan to explore vision–language models that directly perform multimodal question answering in an interactive manner. Fourth, we will investigate training on broader multi-center and cross-institutional datasets, along with prospective evaluation, to further assess robustness across diverse clinical environments. Fifth, we aim to investigate semi-supervised and weakly supervised learning strategies to reduce dependence on fully annotated data.

O Reproducibility

To support reproducibility, we release our code at: https://github.com/chandravardhan-singh-raghaw/CliQ-RRG