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

Anonymous ACL submission

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, domain-specific clinical knowledge is injected into intermediate textual representations, 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://anonymous.4open.science/r/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., 2020; Yan and Pei, 2022), existing methods face four key limitations (Fig. 1). (G1) Prior works (Jin et al., 2024; Luo et al., 2024; Wang et al., 2025) mainly

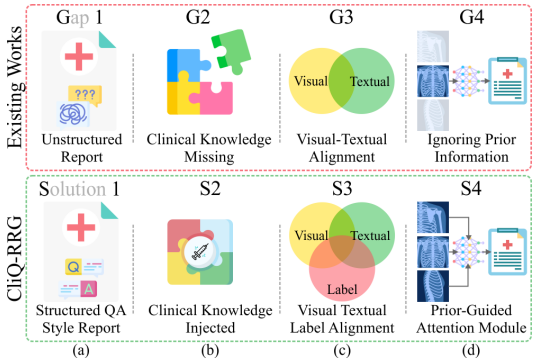

Figure 1: Comparison with existing methods. Existing approaches generate unstructured reports, neglect prior scans, and lacks in clinical knowledge, whereas CliQ-RRG produces QA-style reports, injects clinical knowledge, performs visual–textual–label alignment, and leverages prior scans via prior-guided attention.

generate long free-text narratives, in which key findings are embedded within paragraphs rather than presented in a structured form. Existing studies have noted limitations of lengthy and inconsistently 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 solely 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 more specific, 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 often 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 tri-channel 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 that integrates domain knowledge, leverages multi-view priors, and performs disease-aware alignment. 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 knowledge-guided generation pipeline that enriches intermediate reports with retrieved clinical knowledge and uses a large language model to restructure 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:

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

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 by appending $\text{top}{-}k_t$ clinically relevant knowledge tokens from a clinical knowledge. Finally, an LLM restructures the knowledge-injected report into a concise QA-style format with question–answer pairs and a clinical summary.

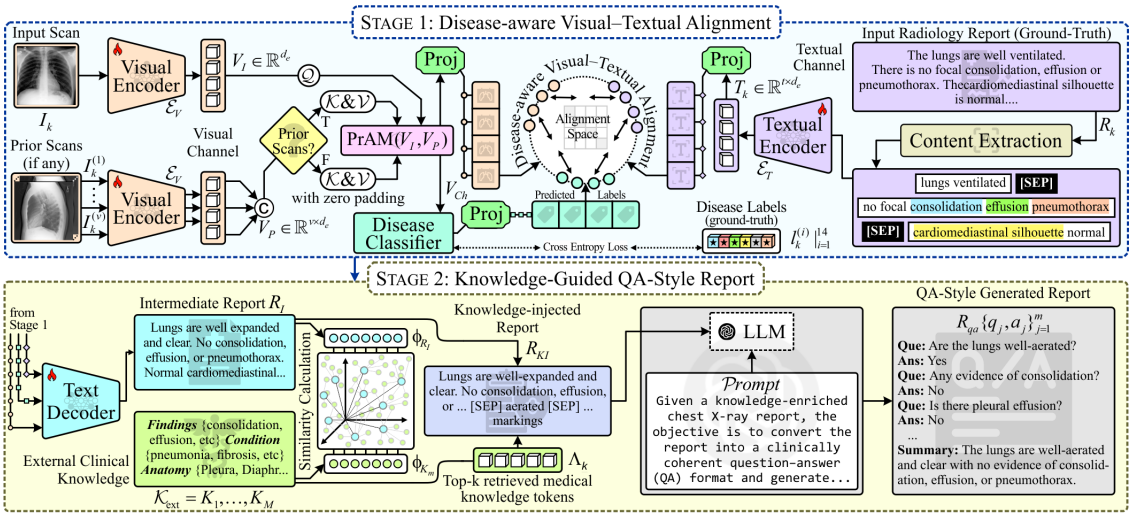

Figure 2: Overview of the proposed CliQ-RRG framework. STAGE 1: Disease-aware Visual-Textual Alignment aligns image and text features using predicted disease embeddings, while the Prior-Guided Attention Module (PrAM) integrates multi-view information from current and prior scans. STAGE 2: Knowledge-Guided QA-Style Report Generation enriches the intermediate report with retrieved clinical knowledge and employs a large language model to produce concise question–answer 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., 2025) 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.

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, 2024c), 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 $V_{ch} \in \mathbb{R}^{d_e}$. The classifier computes disease logits using a cross-attention mechanism (Vaswani et al., 2017) and is optimized with a cross-entropy loss:

$$L_{ch} = \text{Softmax}\left(\frac{V_{ch}\Phi^\top}{\sqrt{d_e}}\right) \quad (3)$$

$$\mathcal{L}_{CE} = \text{CrossEntropy}(L_{ch}, L_{gt}) \quad (4)$$

where $\Phi \in \mathbb{R}^{d \times d_e}$ 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. We retain the first 13 disease categories to identify reports with positive or uncertain findings.

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.

$$\min_{\{\theta_V, \theta_T\}} \mathcal{L}_{\text{Di-VTC}} = \mathcal{L}_{v \mapsto t} + \mathcal{L}_{t \mapsto v} \tag{5}$$

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

$$\mathcal{L}_{v \mapsto t} = -\sum_{g \in B} \frac{1}{|\mathrm{p}(g)|} \sum_{i \in \mathrm{p}(g)} \log \frac{\exp(\hat{v}_g^\top \hat{t}_i / \tau)}{\sum_{h \in B} \exp(\hat{v}_g^\top \hat{t}_h / \tau)} \tag{6}$$

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

where $\mathrm{p}(\cdot)$ 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.

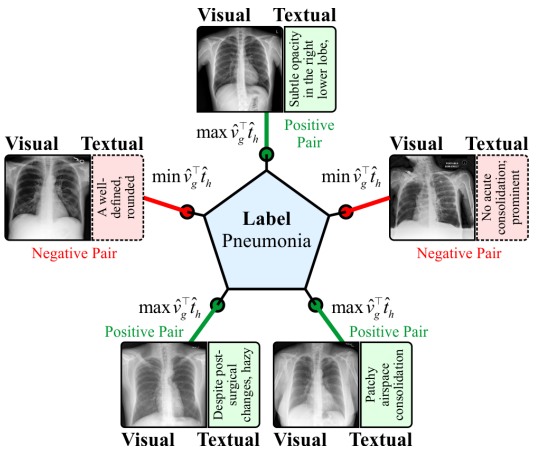

Figure 3: Conceptual illustration of the Disease-aware Visual–Textual Contrast (Di-VTC) framework. The model optimizes a "push–pull" dynamic where positive visual–textual pairs sharing disease labels (green lines) are pulled closer, while unrelated pairs (red lines) are pushed apart, strictly enforcing diagnostically 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 the visual channel V_{ch}, textual channel T_{ch}, and predicted diagnostic labels L_{ch} from Stage 1. For each k^{th} study, the decoder produces a sequence that closely matches the ground-truth (GT) report $R_k = R_k^1, \ldots, R_k^l$, where l is the report length. At each step t, it predicts token R_k^t conditioned upon all preceding tokens $R_k^1, \ldots, R_k^{t-1}$ and the integrated channel features (V_{ch}, T_{ch}, L_{ch}). We employ the negative log-likelihood as the generation loss $\mathcal{L}_{\text{G}}$ to optimize the decoder:

$$\mathcal{L}_{\text{G}} = -\sum_{t=1}^{l} \log p\left(R_k^t | R_k^1, \ldots, R_k^{t-1}, V_{ch}, T_{ch}, L_{ch}\right) \tag{8}$$

where $p(R_k^t|\cdot)$ denotes the probability of generating the t^{th} token given the preceding tokens and the multi-channel input. The loss $\mathcal{L}_{\text{G}}$ ensures the decoder generates accurate clinically reports aligned with (V_{ch}, T_{ch}, L_{ch}) information.

Knowledge Token Retrieval: Unlike prior knowledge-injected RRG methods such as KiUT (Huang et al., 2023) and EKAGen (Bu et al., 2024a), which integrate external knowledge during decoding, CliQ-RRG integrates retrieved knowledge at the intermediate report level and utilizes LLM to restructure into QA-style pairs. Inspired by Liu et al. (2024a), we enrich the

intermediate report with external clinical knowledge to support QA-style synthesis. We construct a medical knowledge base $\mathcal{K}_{\text{ext}}{=}K_1, \ldots, K_M$, where each K_m is a structured phrase describing a symptom, anatomical structure, or disease manifestation, curated from trusted online medical sources (refer Implementation Details). Inspired by the bootstrapping strategy in Liu et al. (2024a), we sample external documents relevant to the diagnostic labels $L_k = [l_k^{(i)}]_{i=1}^{14}$ obtained in STAGE 1, ensuring that the retrieved knowledge is clinically aligned. Using BioWordVec (Zhang et al., 2019b), we embed the intermediate report R_I and each knowledge entry K_m into a shared space $\mathbb{R}^{d_w}$, yielding vectors ϕ_{R_I} and ϕ_{K_m}. We then identify the top-k_t most relevant knowledge tokens by ranking them according to their cosine similarity with the report.

$$\text{sim}(\phi_{R_I}, \phi_{K_m}) = \frac{\phi_{R_I} \cdot \phi_{K_m}}{\|\phi_{R_I}\| \|\phi_{K_m}\|}, \forall\, m : 1 \to M \tag{9}$$

Next, we retrieve the indices of the top$-k_t$ most relevant medical entries

$$\Lambda_{k_t} = \underset{m \in [1,M]}{\arg\ \text{top}-k_t}\ \text{sim}(\phi_{R_I}, \phi_{K_m}) \tag{10}$$

Finally, the selected knowledge tokens $K_m \mid m \in \Lambda_{k_t}$ are appended to the intermediate report to create a knowledge-injected report, R_{KI} for the QA-style report generation.

QA-Style Report Generation: We employ *gpt-3.5-turbo* model from OpenAI, denoted as $\mathcal{G}(\cdot)$, to restructure the knowledge-injected report $R_{KI}^{(k)}$ into a set of clinically meaningful question–answer pairs $R_{qa}^{(k)}$ and a concise summary, where k indexes the study. A specific $\mathcal{P}rompt$ guides $\mathcal{G}(\cdot)$ to elicit the desired output format, as shown below:

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 the QA-style generation as:

$$R_{qa}^{(k)} = \mathcal{G}\left(R_{KI}^{(k)}, \mathcal{P}rompt\right) \tag{11}$$

The output $R_{qa}^{(k)}$ contains m QA pairs, $\{(q_j, a_j)\}_{j=1}^{m}$, with q_j as the question and $a_j \in \{\text{Yes}, \text{No}\}$ as its concise answer. These are grouped thematically, with a final summary question encapsulating the key findings.

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:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{Di-VTC}} + \lambda_1 \cdot \mathcal{L}_{\text{G}} + \lambda_2 \cdot \mathcal{L}_{\text{CE}} \tag{12}$$

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 two public benchmarks: (1) MIMIC-CXR (Johnson et al., 2019), the largest RRG dataset with 337,110 scans and 227,835 reports, using the 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), we evaluate report quality using standard NLG metrics: BLEU (B) (Papineni et al., 2001), METEOR (MTR) (Denkowski and Lavie, 2011), ROUGE-L (RG-L) (Lin, 2004), and BERTScore (BERT) (Zhang et al., 2019a). For binary yes/no QA pairs, we follow recent studies (Manes et al., 2024; Kim et al., 2024) and conduct expert evaluation. Clinical Efficacy (CE) is assessed by labeling generated reports with CheXpert (Smit et al., 2020) and computing F1, Precision (Pre), and Recall (Rec). 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 Tesla T4 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. At inference, the *gpt-3.5-turbo* is employed in a few-shot setting to restructure the knowledge-injected report into QA pairs without fine-tuning. Implementation details are provided in appendix D.

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	BERT$^{\#}$	B1	B4	MTR	RG-L	BERT$^{\#}$
I	R2GenRL* (Qin and Song, 2022)	ACL'22	0.392	0.113	0.149	0.275	0.852	0.487	0.177	0.205	0.377	0.835
	DCL* (Li et al., 2023)	CVPR'23	0.384	0.114	0.147	0.278	0.844	0.483	0.171	0.201	0.391	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	0.860	0.406	0.177	0.199	0.423	0.849
	SEI* (Liu et al., 2024b)	MICCAI'24	0.381	0.119	0.148	0.306	0.841	0.486	0.173	0.209	0.395	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	0.873	0.487	0.178	0.215	0.401	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	–	0.490	0.180	0.218	0.385	–
	LLM-RG4* (Wang et al., 2025)	AAAI'25	0.384	0.136	0.155	0.322	0.849	0.442	0.179	0.192	0.403	0.848
IV	MPO† (Xiao et al., 2025)	AAAI'25	0.416	0.139	0.162	0.309	–	0.548	0.209	0.224	0.415	–
	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	–
	CliQ-RRG (Knowledge-Injected)	Ours	0.498	0.275	0.182	0.324	0.897	0.553	0.216	0.221	0.434	0.878
	CliQ-RRG (QA Pair)	Ours	**0.516**	**0.284**	0.191	**0.343**	**0.911**	**0.561**	**0.227**	0.230	**0.442**	**0.895**

I: Contrastive-based; II: Knowledge Injected-based; III: LLM-based, IV: Multi View-based; V: QA Style-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
EKAGen† (Bu et al., 2024a)	0.517	0.483	0.499
PromptMRG† (Jin et al., 2024)	0.501	0.509	0.476
KiUT† (Huang et al., 2023)	0.371	0.318	0.321
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 (Knowledge-Injected)	0.591	0.602	0.596
CliQ-RRG (QA Pair)	0.605	**0.618**	**0.611**
Disease Classifier	*0.496*	*0.515*	*0.505*

4 Result and Discussion

Quantitative Analysis: We compare CliQ-RRG with five categories (Appendix B) on MIMIC-CXR and IU X-Ray. All reproduced baselines (⋆) follows original configurations (i.e., without prior scans and external knowledge). For fair comparison, we evaluate two settings: (i) Knowledge-Injected (R_{KI}), before QA restructuring, and (ii) QA-Pair (R_{qa}), which aggregates the diagnostic summary and generated QA-pairs into a single sequence and compares it with GT reports. Although the GT and QA-pair format differs, the diagnostic summary and findings produce a high density of clinically relevant terms, ensures strong n-gram overlap and fair comparison with baselines. We further report format-independent metrics (BERTScore, CE) to verify diagnostic grounding improvements.

Comparison with State-of-the-Art Methods: As detailed in Table 1, CliQ-RRG achieves SOTA performance across most metrics. Compared to contrastive and LLM-based baselines, our framework demonstrates stronger semantic alignment with the GT. On the multi-view MIMIC-CXR and IU X-Ray datasets, CliQ-RRG outperforms all competitors, including MPO (Xiao et al., 2025), highlighting the effectiveness of the PrAM for multi-view integration. Knowledge-injected methods such as RADAR (Hou et al., 2025) on MIMIC-CXR and KiUT (Huang et al., 2023) on IU X-Ray remain competitive. REVTAF (Zhou et al., 2025) achieves the highest MTR and second-highest RG-L on MIMIC-CXR. We outperform CoD (Jin et al., 2025) QA-style baselines in both evaluation settings, confirming the efficacy of our tri-channel alignment and knowledge injection.

Evaluation of Clinical Efficacy Metrics: Table 2 reports CE metric on MIMIC-CXR, where CliQ-RRG achieves the superior performance and outperforms REVTAF (Zhou et al., 2025) in Pre. We also evaluate the disease classifier, confirming that classifier provides reliable supervisory signals to accurately guide multimodal alignment.

Table 3: Ablation analysis on MIMIC-CXR showing incremental performance gains by integrating the PrAM, Di-VTC (across V_{ch}–T_{ch} and L_{ch}), $\mathcal{K}_{\text{ext}}$, and LLM component into the base model. Results are reported as mean $\pm$ 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}	$\mathcal{K}_{\text{ext}}$	LLM	BL-1	BL-4	MTR	RG-L	∇_{N}	P	R	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)	×	✓	✓	✓	×	0.472±0.004	0.256±0.005	0.178±0.003	0.319±0.006	20.3%	0.556	0.588	0.572	23.5%
(g)^{y}	✓	✓	✓	✓	×	0.498±0.005*	0.275±0.004*	0.182±0.004*	0.324±0.005*	25.6%	0.591	0.602	0.596	28.8%
(h)^{z}	✓	✓	✓	✓	✓	0.516±0.003*	0.284±0.007*	0.191±0.005*	0.343±0.006*	31.0%	0.605	0.618	0.611	32.1%

PrAM: Prior-Guided Attention Module; V_{ch}: Visual Channel; T_{ch}: Textual Channel; L_{ch}: Predicted diagnostic label; $\mathcal{K}_{\text{ext}}$: External clinical knowledge.
LLM: Large Language Model; ∇_{N} and ∇_{C}: Average improvement over the base configuration for NLG and CE metrics;
x: CliQ-RRG (Base Model); y: CliQ-RRG (Knowledge-Injected); z: CliQ-RRG (QA Pair)

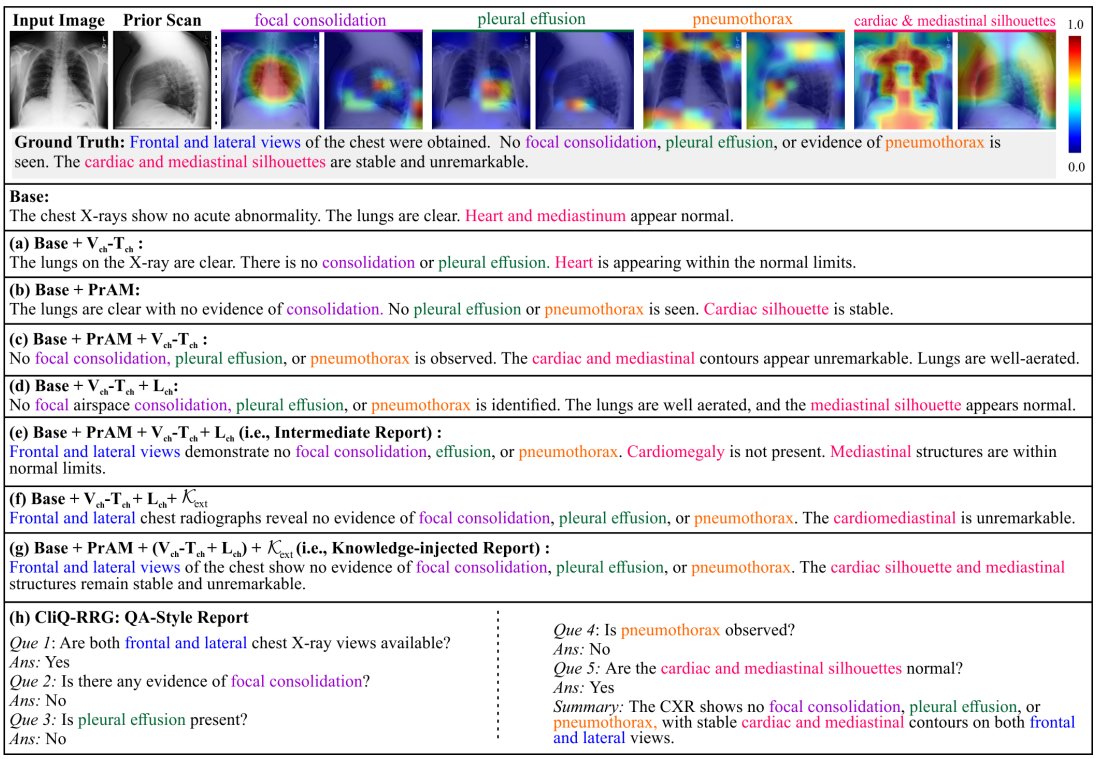

Figure 4: Qualitative analysis of CliQ-RRG on MIMIC-CXR. Report comparison across models given input and prior scans, with ground truth (GT). Shared medical terms use consistent color coding, and Grad-CAM highlights anatomically and pathologically relevant regions.

Ablation Study: We analyze the contribution of each component in CliQ-RRG on MIMIC-CXR (Table 3) and IU X-Ray (appendix E). The **base** model uses a visual encoder and transformer decoder, without alignment or knowledge injection. 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. Conversely, adding PrAM to the base **(model b)** enhances clinical efficacy (F1: 0.493) by focusing on relevant regions. Combining both **(model c)** further improves the 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), demonstrates label-based grounding superior to image-text pairing. Integrating PrAM with Di-VTC **(model e)** improves NLG and CE metrics. Adding $\mathcal{K}_{ext}$ into Di-VTC **(model f)** improves semantic completeness (F1: 0.572). The knowledge-enhanced framework **(model g)** outperforms the base model (NLG: +25.6%, CE: +28.8%), validating that $\mathcal{K}_{ext}$ complements $V_{ch}-T_{ch}-L_{ch}$ channel alignment. Our full framework with LLM-based QA restructuring **(model h)** attains the best perfor-

mance, with improvements of 31.0% (NLG) and 32.1% (CE). Results are reported as mean ± std over five runs, and models (g) and (h) show statistically significant gains ($p \leq 0.05$).

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, capturing terms such as *"cardiomegaly"*. Enriching the report with external clinical knowledge further refines phrasing (*"stable and unremarkable"*), aligning 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.

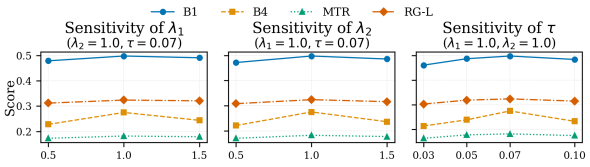

Figure 5: Parameter sensitivity analysis on MIMIC-CXR showing the effect of generation loss weight (λ_1), classification loss weight (λ_2), and temperature (τ).

Human 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 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 4 (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 200 QA pairs randomly from MIMIC-CXR and compare with corresponding GT reports, categorizing each pair as fully acceptable, clinically acceptable, or hallucinated. As shown in Table 4 (ii), 189 pairs are fully acceptable, and 9 are clinically consistent, yielding a 99.0% reliability rate, with 1.0% hallucinated content. The results validate that restructuring into a QA pair preserves diagnostic correctness.

Table 4: 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.

(i)

Dataset	R_0	R_1	R_2	R_3	R_4	PA	κ_F
MIMIC-CXR	4.03	3.96	4.25	4.09	4.23	0.79	0.76
IU X-Ray	4.13	4.09	4.43	3.91	4.11	0.82	0.78

(ii)

Outcome	Count	%
Fully acceptable	189	94.5%
Clinically acceptable (minor linguistic deviations)	9	4.5%
Hallucinated (contradictory/unsupported findings)	2	1.0%
Total Reliable (Fully + Clinically) Pairs	**198**	**99.0%**

5 Related Work

Due to space constraints in the main text, we provide a comprehensive discussion of related work and prior approaches 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 a prior-guided attention for integrating multi-view chest X-rays. Additionally, injecting domain-specific clinical information enriches the semantic depth of the generated reports and supports the structured QA formulation. Qualitative analysis shows that CliQ-RRG effectively aligns the 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 proposed framework over state-of-the-art methods.

7 Limitations

While CliQ-RRG demonstrates strong performance, several limitations suggest directions for future improvement. First, the QA formulation is grounded in a predefined diagnostic label space (i.e., CheXpert categories), which ensures systematic coverage of clinically important findings but may limit open-set detection of rare or previously unseen diseases. Second, 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. Third, the large language model is used as a fixed restructuring component to convert knowledge-enriched reports into structured yes or no QA pairs, rather than being jointly optimized for multimodal reasoning. Fourth, evaluation is limited to MIMIC-CXR and IU X-Ray, which follow specific acquisition protocols and population characteristics. Fifth, the framework relies on supervised learning with paired images, reports, and diagnostic labels, which may constrain scalability in data-scarce settings. Finally, external knowledge is retrieved through semantic similarity, prioritizing accurate grounding while leaving explicit modeling of disease interdependencies unaddressed.

References

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, Yx L, 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.

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 *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 10348–10359.

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. 2024c. Factual serialization enhancement: A key innovation for chest x-ray report generation. *Preprint*, arXiv:2405.09586.

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*.

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, page 311. 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 *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 15580–15589.

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.

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.

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.

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. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 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. 2025. 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. 2019a. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. 2019b. Biowordvec, improving biomedical word embeddings with subword information and mesh. *Scientific Data*, 6(1).

Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. 2020. 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22529–22538.

Appendix

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 five 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.

IV: Multi View-based Methods

- *MPO* (Xiao et al., 2025) adapts report generation to different user preferences by optimizing weighted objectives through reinforcement learning.
- *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.

C Medical Knowledge Base Construction

We construct an external knowledge base $\mathcal{K}_{ext} = \{K_m\}_{m=1}^{M}$ for knowledge-guided generation used in STAGE 2.

Step 1: Source Collection

We curated approximately 5,000 reliable medical documents from PubMed abstracts related to thoracic imaging and chest radiology. Documents were filtered to retain content relevant to radiographic findings, disease descriptions, and anatomical structures.

Step 2: Knowledge Unit Extraction
We processed the raw text to extract atomic clinical facts rather than full paragraphs. Using a biomedical Named Entity Recognition (NER) model instantiated with BioClinicalBERT, we identified sentences containing at least one radiology-relevant entity (e.g., symptom, anatomical region, imaging finding). Each sentence is converted into a concise atomic phrase through normalization. The normalization step removes redundant modifiers, standardizes terminology, and retains clinically meaningful descriptors. Near-duplicate entries are filtered using cosine similarity thresholding.

Step 3: Label Mapping
Each knowledge entry $\mathcal{K}_m$ is mapped to disease category $y_m \in \mathcal{Y}$. This mapping ensures that the retrieved knowledge is consistent with the predicted diagnostic labels (L_{ch}) used during the generation process.

Step 4: Embedding and Retrieval Setup
We embedded each knowledge entry $\mathcal{K}_m$ into a vector space using BioWordVec to create a searchable index. During the report generation phase, we calculate the cosine similarity between the embedding of the generated intermediate report (ϕ_{R_I}) and the knowledge entries (ϕ_{K_m}) to retrieve the top-k most relevant clinical facts, which are then appended to the report context.

D 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. To inject external clinical knowledge, we construct an external knowledge base aligned with the predicted disease label set L_k (refer to Appendix C). During generation, we retrieve the top-$k_t = 10$ most relevant knowledge tokens via semantic similarity.[1] All experiments were supported by an Intel(R) Xeon(R) Silver 4215R CPU with 256 GB RAM, and inference was set with $m = 5$.

E Ablation Analysis on IU Dataset

We further validate each component on the IU X-Ray dataset as detailed in Table 5. The base encoder–decoder achieves a BL-1 score of 0.435. Adding standard visual–textual alignment in model (a) improves average performance by 5.4%, while incorporating the Prior-Guided Attention Module in model (b) yields a performance gain of 8.2%, highlighting the benefit of multi-view context. Introducing disease-aware supervision further improves performance. Replacing standard alignment with Disease-aware Visual–Textual Contrast in model (d) increases BL-1 to 0.498, surpassing the prior-aware model (c) and demonstrating stronger grounding from disease labels. Combining PrAM with Di-VTC in model (e) increases the score to 0.531. Injecting external clinical knowledge into the model (g) further refines performance to 0.553, confirming the value of domain-specific context. The full framework in model (h), which applies LLM-based QA restructuring, achieves the best results across all NLG metrics with a 28.1% average improvement over the base model. These results are consistent with our MIMIC-CXR experiments and confirm the robustness of CliQ-RRG.

Table 5: Ablation analysis on IU dataset showing incremental performance gains by integrating the PrAM, Di-VTC (across V_{ch}–T_{ch} and L_{ch}), $\mathcal{K}_{\text{ext}}$, and LLM-based components into the base configuration.

Model	STAGE 1			STAGE 2		NLG Metrics				
	PrAM	$V_{ch}-T_{ch}$	L_{ch}	$\mathcal{K}_{\text{ext}}$	LLM	BL-1	BL-4	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)	×	✓	✓	✓	×	0.544	0.211	0.217	0.429	22.9%
(g)	✓	✓	✓	✓	×	**0.553**	**0.216**	0.221	**0.434**	**24.9%**
(h)	✓	✓	✓	✓	✓	**0.561**	**0.227**	0.230	**0.442**	**28.1%**

PrAM: Prior-Guided Attention Module; V_{ch}: Visual Channel; T_{ch}: Textual Channel; L_{ch}: Predicted diagnostic label; $\mathcal{K}_{\text{ext}}$: External clinical knowledge. LLM: Large Language Model; ∇_{N}: Average improvement across all NLG metrics over the base configuration.

F Impact of LLMs on QA Restructuring

We evaluate the impact of different LLMs on restructuring the intermediate report into QA format. Fig. 6 reports results on MIMIC-CXR and IU X-Ray using BLEU-4, CE (F1), and BERTScore. We compare closed-source models (GPT-3.5-turbo, GPT-5, Gemini) and an open-source model (Llama-3-8B-Instruct) under zero-shot, few-shot, and QLoRA fine-tuned settings. Few-shot prompting consistently improves over zero-shot, while models such as GPT-5 and Gemini further improve restructuring quality. Fine-tuned Llama-3-8B-Instruct achieves the best overall performance, showing

[1] We tested top-$k_t \in [1, 15]$ and selected 10 for its optimal performance.

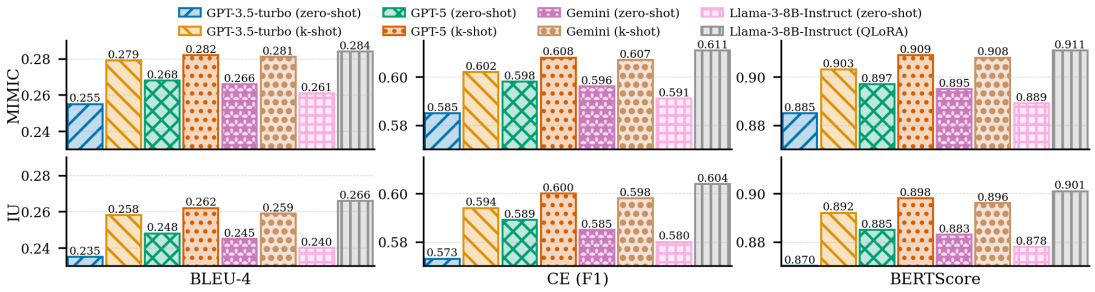

Figure 6: QA restructuring comparison across LLMs under zero-shot, few-shot, and fine-tuned settings on MIMIC-CXR and IU X-Ray using BLEU-4, CE (F1), and BERTScore.

that lightweight task-specific adaptation is more effective than prompt design alone.

Since all LLMs operate on the same knowledge-injected intermediate report, diagnostic content remains unchanged; performance differences mainly reflect linguistic restructuring ability, indicating robustness to LLM choice.

G Computational Efficiency

We analyze the computational efficiency of CliQ-RRG on a single NVIDIA Tesla T4 GPU, as summarized in Table 6. The framework consists of a two-stage architecture with $\sim$ 170M trainable parameters, including STAGE 1 multimodal alignment and STAGE 2 report decoding. The average inference time of the pipeline is 150 ms per study, excluding the final QA restructuring step. Knowledge token retrieval introduces negligible overhead due to the fixed vocabulary and simple similarity search. The external LLM is invoked only once per study for QA formatting via the GPT-3.5-turbo API and is not included in the local inference cost. This final restructuring step introduces a variable latency of 1.2 to 2.5 s and runs independently of the local inference hardware.

Table 6: Computational efficiency of CliQ-RRG

Method	Parameters	Inference Time
CliQ-RRG (Knowledge-Injected)	$\sim$ 170M	150–250 ms / study*
CliQ-RRG (QA Pair)		1.35–2.75 s / study*

$*$: For Single-scan and prior scan.

H 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.

I 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. Third, we plan to explore vision–language models that directly perform multimodal question answering in an interactive manner. Fourth, future evaluations will include multi-center and cross-institutional datasets to examine robustness across diverse clinical environments. Fifth, we aim to investigate semi-supervised and weakly supervised learning strategies to reduce dependence on fully annotated data. Sixth, incorporating structured medical knowledge graphs will allow the model to capture disease relationships while preserving semantic knowledge retrieval.

J Reproducibility

To support reproducibility, we release our code at: https://anonymous.4open.science/r/CliQ-RRG