

RA-RRG: Multimodal Retrieval-Augmented Radiology Report Generation with Key Phrase Extraction

Jongwon Park¹ Byungmu Yoon¹ Soobum Kim¹ Kyoyun Choi^{2*}

¹ DEEPNOID Inc., Seoul, South Korea

² Department of Artificial Intelligence and Data Science, Sejong University, Seoul, South Korea

jgpark@deepnoid.com, kychoi@sejong.ac.kr

Abstract

Automated radiology report generation (RRG) holds potential to reduce the workload of radiologists, and recent advances in multimodal large language models (MLLMs) have enabled multimodal chest X-ray (CXR) report generation. However, existing MLLMs are computationally expensive, require large-scale training data, and may produce hallucinated content, limiting their practical deployment. To address these limitations, we propose RA-RRG, a retrieval-augmented RRG framework that combines multimodal retrieval with large language models (LLMs) to generate radiology reports while reducing hallucinations and computational demands. RA-RRG uses LLMs to extract clinically essential key phrases from radiology reports and retrieves relevant phrases given an input image. By conditioning LLMs on the retrieved phrases, RA-RRG effectively suppresses hallucinations while maintaining strong report generation performance. Experiments on the MIMIC-CXR and IU X-ray datasets show state-of-the-art results on CheXbert metrics and competitive RadGraph F1 scores compared to MLLMs. Furthermore, RA-RRG naturally generalizes to multi-view RRG by aggregating phrases retrieved from multiple images, highlighting its broad applicability to real-world clinical scenarios. Code is available at <https://github.com/deepnoid-ai/RA-RRG>.

1 Introduction

Automated radiology report generation (RRG) has the potential to substantially reduce the workload of radiologists by translating medical images into textual descriptions. Recent advances in large language models (LLMs) have further expanded this potential, particularly through multimodal models that jointly process images and text, including chest X-rays (CXR) (Chaves et al., 2025; Hyland et al., 2023; Tu et al., 2024; Yang et al., 2024). Despite their strong performance, such multimodal

LLMs (MLLMs) typically require extensive computational resources and large-scale fine-tuning, which hinders their adoption in clinical settings.

Retrieval-augmented generation (RAG) (Lewis et al., 2020) offers a promising alternative by enhancing generation through external knowledge retrieval. In CXR RRG, prior work has explored multimodal retrieval-based approaches (Endo et al., 2021; Ramesh et al., 2022), which retrieve similar reports or sentences based on an input image. However, radiology reports often describe multiple co-occurring findings, and naively retrieved text may include information that is irrelevant or even contradictory to the given image. This issue is exacerbated when sentences from different reports are combined (Kong et al., 2022; Zhao et al., 2024). Moreover, radiology reports frequently contain comparative statements referring to prior examinations. When only a single image is available, we define such comparative content as *comparative hallucinations*, as it is unsupported by the input.

To address these challenges, we propose **RA-RRG**, a **Retrieval-Augmented RRG** framework that combines LLMs with multimodal retrieval without requiring any LLM fine-tuning. Building on RadGraph (Jain et al., 2021), we use an LLM to extract clinically essential key phrases from reports while explicitly excluding undesired content such as comparisons with prior studies, resulting in comparative hallucination-free key phrases aligned with the visual evidence. Given an input image, RA-RRG retrieves relevant key phrases and conditions an LLM on these image-consistent phrases to generate accurate and reliable reports without any LLM training. Experimental results on MIMIC-CXR and IU X-ray demonstrate that RA-RRG achieves strong performance on CheXbert metrics and competitive RadGraph F1 scores, while requiring only 18 GPU-hours for training, compared to over 200 GPU-hours for comparable MLLMs. Furthermore, the proposed framework naturally ex-

*Corresponding author.

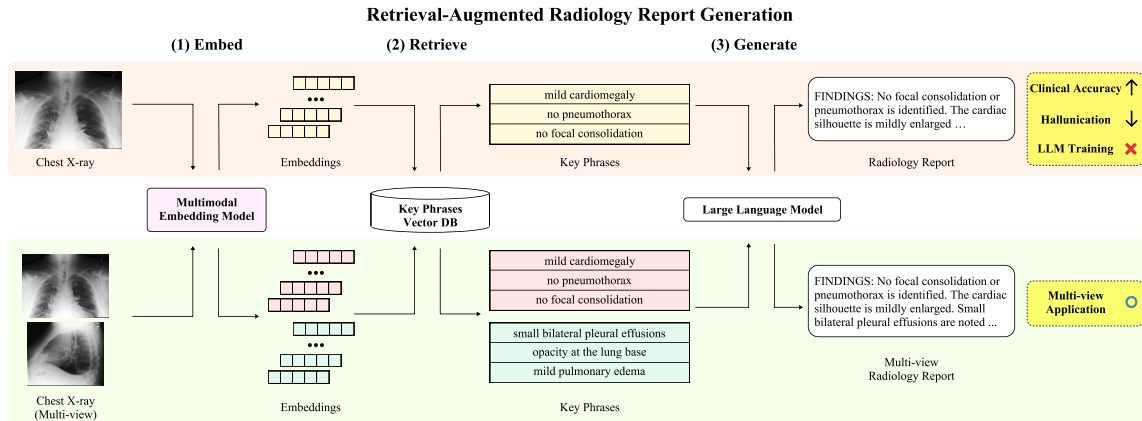


Figure 1: Overview of RA-RRG framework. Given a chest X-ray, a multimodal retriever selects clinically essential key phrases, which are then provided to an LLM to generate hallucination-suppressed reports without LLM training. The same pipeline naturally extends to multi-view inputs.

tends to multi-view RRG by aggregating phrases retrieved independently from multiple images. Figure 1 provides an overview of the RA-RRG framework.

Our main contributions are summarized as follows: (1) We propose RA-RRG, a retrieval-augmented RRG framework that produces clinically accurate radiology reports without LLM fine-tuning. (2) RA-RRG effectively suppresses both comparative and object-level hallucinations. (3) We demonstrate that RA-RRG generalizes well to multi-view settings, highlighting its broad applicability in real-world scenarios.

2 Related Works

2.1 Retrieval Augmented Generation

While LLMs have achieved human-level knowledge in various fields, they still suffer from outdated knowledge and hallucinations (Huang et al., 2025). Combining retrieval-augmented generation (RAG) with LLMs (Lewis et al., 2020; Gao et al., 2023) addresses these issues by retrieving information from an external database based on the query, allowing for updates without retraining the LLM.

Recent advances in MLLMs have expanded RAG to multimodal applications, including text-to-image generation (Chen et al., 2023; Yasunaga et al., 2023), image captioning (Sarto et al., 2022; Ramos et al., 2023; Li et al., 2024), and video captioning (Xu et al., 2024). In this study, we apply a multimodal RAG approach to generate radiology reports by retrieving text data with embeddings aligned to CXR images.

2.2 Radiology Report Generation

Automated RRG has been actively studied in recent years. With the advent of MLLMs, CXR-focused systems such as LLaVA-Rad (Chaves et al., 2025), CheXagent (Chen et al., 2024), MAIRA-1 (Hyland et al., 2023), MAIRA-2 (Bannur et al., 2024), and M4CXR (Park et al., 2025) have demonstrated report generation capabilities. General-purpose medical foundation models, including Med-Gemini (Yang et al., 2024) and MedPaLM-M (Tu et al., 2024), also support CXR report generation but require substantial computational resources and large-scale training data.

Retrieval-based approaches mitigate these limitations by leveraging existing reports. TranSQ (Kong et al., 2022) framed RRG as a set prediction problem, while Teaser (Zhao et al., 2024) introduced topic-wise retrieval with contrastive learning. CXR-RePaiR (Endo et al., 2021) and CXR-ReDonE (Ramesh et al., 2022) aligned CXR images and reports using CLIP- and ALBEF-based objectives, respectively, and CXR-RAG (Ranjit et al., 2023) combined retrieval with a pre-trained LLM for report generation. Liu et al. (2024) further improved retrieval-based RRG through in-domain adaptation and contrastive ranking with structured decoding.

Another line of work leverages RadGraph (Jain et al., 2021) to represent reports as structured clinical knowledge. Style-aware RRG (Yan et al., 2023) serialized RadGraph outputs to model radiologist-specific styles, while FactMM-RAG (Sun et al., 2025) focused on pathology-centric factual extraction with contrastive learning. In contrast, our approach integrates an LLM to refine RadGraph outputs into hallucination-suppressed key phrases for retrieval-augmented RRG.

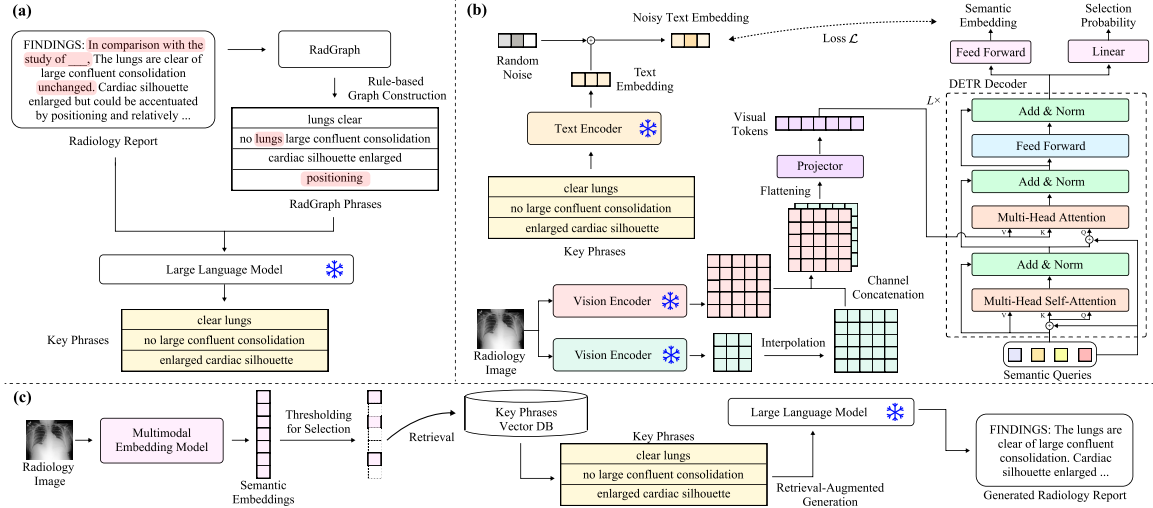


Figure 2: (a) Key phrase extraction using an LLM. (b) The multimodal retriever architecture. (c) Inference process of RA-RRG.

3 Methods

3.1 LLM-Based Key Phrase Extraction

Most prior retrieval-based RRG approaches treat the entire radiology report as the retrieval target (Endo et al., 2021; Ranjit et al., 2023) or split reports into sentence-level segments (Kong et al., 2022; Zhao et al., 2024). However, both strategies suffer from co-occurrence issues, as multiple independent findings may coexist within a single text, and reports often include extraneous information such as doctor names or user metadata. To better utilize radiology reports for training, we decompose reports into minimal clinically meaningful phrases while removing unnecessary content. Specifically, we apply RadGraph (Jain et al., 2021) to the *FINDINGS* section to extract entities and relations, and construct RadGraph phrases that represent the key clinical findings in the report (see Appendix B).

Despite its effectiveness, RadGraph may produce fragmented graphs and does not explicitly address hallucination-prone expressions such as comparative statements involving prior studies (e.g., *improved* or *unchanged*), which are unsupported in single-image RRG. To address this limitation, we employ an LLM for key phrase extraction, inspired by prior work on LLM-based knowledge graph extraction (Gutierrez et al., 2024), and refine RadGraph outputs into key phrases while excluding hallucination-indicative terms. Because general-purpose LLMs may omit domain-specific clinical details when processing raw text alone, we provide both the original report and RadGraph-derived structures as input, with examples and prompt templates shown in Figure 2 (a) and Appendix Figure 7.

3.2 Multimodal Retriever

3.2.1 Model Architecture

To train a multimodal retrieval model using images and their corresponding lists of key phrases, our model builds on the architecture of TranSQ (Kong et al., 2022). TranSQ adapts the training approach of DETR (Carion et al., 2020) for sentence-level retrieval. Our model consists of a vision encoder, a DETR decoder, and a text encoder, as illustrated in Figure 2 (b).

Vision encoder We leverage complementary strengths of pre-trained vision encoders obtained from different pre-training paradigms, including vision-language pre-training (e.g., CLIP (Radford et al., 2021)) and unimodal self-supervised learning (e.g., DINOv2 (Oquab et al., 2024)). Rather than relying on a single encoder, we fuse visual features from multiple encoders using channel concatenation (Shi et al., 2025). All encoders follow a Vision Transformer (ViT) (Dosovitskiy et al., 2021) architecture and output sequences of visual tokens. To address differences in sequence length, we reshape the token sequences into 2D feature maps, apply spatial interpolation to align their resolutions, and concatenate them along the channel dimension to form a unified visual representation.

Text encoder During training, the text encoder converts key phrases, extracted from the report corresponding to each training image, into text embeddings. Since the text encoder is kept frozen as in (Kong et al., 2022), the resulting text embeddings for each training image remain fixed, which can lead to overfitting. Inspired by NEFTune (Jain

et al., 2024), which adds random noise to embeddings during LLM fine-tuning, we similarly inject noise into text embeddings during training. The noise ϵ is sampled from a uniform distribution in the range $[-1, 1]$ and scaled by $1/\sqrt{d}$, where d is the embedding dimension. We also apply L2 normalization to the text embeddings. For inference, we construct a vector database of embeddings from all key phrases in the training set to facilitate retrieval.

DETR decoder Similar to TranSQ, we adopt the original DETR decoder structure. The visual token sequence from the vision encoder serves as the encoder input, while N query embeddings are decoded in parallel through self-attention and encoder-decoder attention. To match the feature dimension of the visual token sequence with that of the decoder, we apply a linear projection layer. A selection classifier, implemented as a linear layer, produces selection logits, and semantic embeddings are generated by a three-layer feed-forward network with ReLU activation. Each semantic embedding is then L2-normalized.

3.2.2 Loss Function

Phrase matching loss We optimize the retriever using a phrase matching loss. Similar to DETR, TranSQ applies the Hungarian algorithm (Kuhn, 1955) based on selection probabilities and the similarity between semantic and text embeddings. Let y represent the ground truth set of key phrases. $\hat{y} = \{\hat{y}_i\}_{i=1}^N$ consists of N predictions. We set N to exceed the number of key phrases and pad y with empty elements (\emptyset) to form equal-sized sets. The Hungarian algorithm then finds an optimal permutation $\sigma \in \mathfrak{S}_N$ that minimizes the matching cost:

$$\hat{\sigma} = \arg \min_{\sigma \in \mathfrak{S}_N} \sum_i^N \mathcal{L}_{\text{match}}(y_i, \hat{y}_{\sigma(i)}) \quad (1)$$

Matching cost $\mathcal{L}_{\text{match}}(y_i, \hat{y}_{\sigma(i)})$ is computed as the negative sum of selection probability $\hat{p}_{\sigma(i)}$ (scaled by μ) and cosine similarity \mathcal{L}_{sim} between text embedding v_i and semantic embedding $\hat{v}_{\sigma(i)}$:

$$\begin{aligned} \mathcal{L}_{\text{match}}(y_i, \hat{y}_{\sigma(i)}) = \\ - \mu \mathbb{1}_{\{y_i \neq \emptyset\}} \hat{p}_{\sigma(i)} - \mathbb{1}_{\{y_i \neq \emptyset\}} \mathcal{L}_{\text{sim}}(v_i, \hat{v}_{\sigma(i)}) \end{aligned} \quad (2)$$

Given the optimal assignment, we compute the selection loss \mathcal{L}_{cls} using distribution-balanced loss (Wu et al., 2020) with binary classification labels $c_i = \mathbb{1}_{\{y_i \neq \emptyset\}}$, and a negative cosine similarity loss to align matched text and semantic embeddings. The overall phrase matching loss is:

$$\begin{aligned} \mathcal{L}_{\text{PM}}(y, \hat{y}) = \sum_{i=1}^N \left[\mathcal{L}_{\text{cls}}(c_i, \hat{p}_{\hat{\sigma}(i)}) \right. \\ \left. + \mathbb{1}_{\{y_i \neq \emptyset\}} (1 - \mathcal{L}_{\text{sim}}(v_i, \hat{v}_{\hat{\sigma}(i)})) \right] \end{aligned} \quad (3)$$

In-batch semantic contrastive loss \mathcal{L}_{sim} aligns matched semantic and text embeddings but does not discourage non-matching embeddings from becoming similar. Thus, we adopt a CLIP-style symmetric contrastive loss (Radford et al., 2021), which pulls matched embedding pairs closer while pushing mismatched pairs apart.

Given a mini-batch of size B , we construct a set of positive pairs $E = \{(v_i^b, \hat{v}_{\hat{\sigma}(i)}^b) \mid y_i^b \neq \emptyset, b = 1, \dots, B\}$ from the Hungarian matching, where v_i^b is the i -th text embedding and $\hat{v}_{\hat{\sigma}(i)}^b$ is its matched semantic embedding in batch b . We flatten all pairs across the batch and index them as $k = 1, \dots, |E|$. Let v_k and \hat{v}_k denote the text and semantic embeddings of the k -th pair, respectively. The cross-modal similarity matrix is defined as:

$$Z_{kl} = v_k^\top \hat{v}_l \quad (4)$$

Rather than using hard one-hot targets that treat all non-matched pairs as negatives, we construct soft contrastive targets from intra-modal self-similarities $S_{kl}^v = v_k^\top v_l$ and $S_{kl}^{\hat{v}} = \hat{v}_k^\top \hat{v}_l$:

$$q_{kl} = \text{softmax}_l \left((S_{kl}^v + S_{kl}^{\hat{v}}) / 2 \right) \quad (5)$$

This ensures that semantically similar pairs in both modalities are not heavily penalized. The loss is defined as a symmetric cross-entropy:

$$\begin{aligned} \mathcal{L}_{\text{SC}} = \frac{1}{2|E|} \sum_{k=1}^{|E|} \left[H(q_{k,:}, \text{softmax}(Z_{k,:})) \right. \\ \left. + H(q_{:,k}, \text{softmax}(Z_{:,k})) \right] \end{aligned} \quad (6)$$

where $H(p, r) = -\sum_l p_l \log r_l$ denotes the cross-entropy. The total training objective is:

$$\mathcal{L} = \sum_{b=1}^B \mathcal{L}_{\text{PM}}(y^b, \hat{y}^b) + \lambda \mathcal{L}_{\text{SC}}(E). \quad (7)$$

3.3 Multimodal Retrieval-Augmented RRG

3.3.1 Key Phrase Retrieval

To generate a radiology report from an image, we compute N selection probabilities and the corresponding semantic embeddings. Only embeddings with probabilities above a set threshold are used for key phrase retrieval. The retrieval target is a

vector database of text embeddings, built from the full set of key phrases gathered from the training dataset. Matching each semantic embedding to its nearest text embedding yields a list of key phrases that describe the image.

3.3.2 Radiology Report Generation with LLM

The final step of RA-RRG uses an LLM to generate a complete radiology report from retrieved key phrases. Since these phrases are not full sentences, the LLM integrates their content to produce a coherent and comprehensive report. To ensure desired report qualities, such as hallucination suppression, we provide task-specific instructions together with the key phrases as input to the LLM; the prompt is shown in Figure 8 in the Appendix.

Using an LLM broadens the applicability of RA-RRG beyond single-image report generation. The framework extends to multi-view and follow-up scenarios by extracting key phrases from each image independently and providing them, along with contextual information such as view position, to the LLM. Unlike MLLM-based approaches that require architectural modifications to handle multiple images (Bannur et al., 2024), RA-RRG enables unified report generation in a straightforward manner.

4 Experiments

4.1 Datasets

4.1.1 Training dataset

For training and validation, we use the **MIMIC-CXR** dataset (Johnson et al., 2019a,b), which contains paired chest X-ray images and radiology reports. Using the official MIMIC-CXR codebase¹, we extract only the *FINDINGS* section from each report and follow the official data split. We retain studies with non-empty RadGraph phrases and key phrases, excluding cases without clinically meaningful reports. After filtering, the dataset consists of 269,241 training images, 2,113 validation images, and 3,858 test images, using both frontal (PA/AP) and lateral views for training. On average, each image is associated with 7.16 key phrases, and the training set contains 243,064 unique key phrases, reflecting substantial redundancy across reports.

4.1.2 Test datasets

Following the official split of **MIMIC-CXR**, we use 3,858 images for single-view RRG evaluation. Unlike training and validation, we retain all 3,858

test images, including those with empty RadGraph or key phrases, to ensure fair comparison with prior work. For external evaluation, we use the IU X-Ray dataset (Demner-Fushman et al., 2016). Following the setting of PromptMRG (Jin et al., 2024), we evaluate on a publicly available² subset of 4,168 images, where frontal and lateral images are treated as independent samples and normal cases are down-sampled to a 10% ratio.

Multi-view RRG is evaluated using both frontal and lateral images, following the test protocol of MAIRA-2. Among the 3,858 MIMIC-CXR test images, 2,461 are frontal views; when multiple lateral images exist, one is selected randomly, and cases without a lateral view are evaluated using the frontal image only.

4.2 Evaluation Metrics

We evaluate both natural language generation (NLG) quality and clinical efficacy, using publicly available implementations³. For NLG evaluation, we report ROUGE-L (Lin, 2004) and BLEU scores (BLEU-1, BLEU-4) (Papineni et al., 2002). Clinical efficacy is assessed using CheXbert (Smit et al., 2020), which labels reports across 14 observation classes. We binarize the labels by treating all non-positive labels as negative and report micro-F1, macro-F1, and example-based F1 scores (Nicolson et al., 2023). We additionally report RadGraph F1 (Yu et al., 2023), which evaluates clinical correctness based on entities and relations extracted by RadGraph (Jain et al., 2021). To assess comparative hallucination, we follow (Ramesh et al., 2022) and measure the frequency of comparison-related keywords (e.g., unchanged, earlier, remain) and the proportion of reports containing them. We additionally evaluate object-level hallucination by computing RadGraph precision, recall, and F1 separately for entities and relations.

4.3 Implementation Details

Based on an exploration of various vision encoder structures (see Appendix A for details), we combine two image encoders made available by Chambon et al. (2024), namely XrayDINOv2 and Xray-CLIP. For the text encoder we employ MPNet ('all-mpnet-base-v2') (Reimers and Gurevych, 2019) with embedding dimension $d = 768$. During multimodal retriever training, both vision and text encoder parameters are frozen. The parameters of

¹<https://github.com/MIT-LCP/mimic-cxr>

²<https://github.com/jhb86253817/PromptMRG>

³Links to the evaluation tools are provided in Appendix D

Table 1: Results of single-view RRG evaluation on the MIMIC-CXR test set (*FINDINGS* section). Compared methods include METransformer (Wang et al., 2023), PromptMRG (Jin et al., 2024), Med-PaLM M (Tu et al., 2024), MAIRA-1 (Hyland et al., 2023), LLaVA-Rad (Chaves et al., 2025), M4CXR (Park et al., 2025), TranSQ (Kong et al., 2022), DCL (Li et al., 2023), I3+C2FD (Liu et al., 2024), and MCA-RG (Xing et al., 2025). * denotes results from PromptMRG; † uses CheXpert labeling; ‡ treats uncertain as positive. Best values are in bold.

Type	Model	CheXbert			RadGraph	NLG Metrics		
		micro-F1	Macro-F1	example-F1	F1	ROUGE-L	BLEU-1	BLEU-4
Generation	METransformer [†]	-	-	31.1	-	29.1	38.6	12.4
	PromptMRG	-	38.1	47.6	-	26.8	39.8	11.2
	Med-PaLM M 84B	53.6	39.8	-	26.7	27.3	32.3	11.3
	MAIRA-1	55.7	38.6	-	24.3	28.9	39.2	14.2
	LLaVA-Rad	57.3	39.5	-	-	30.6	38.1	15.4
	M4CXR	58.1	38.8	50.2	21.7	28.4	33.3	10.2
Retrieval	TranSQ [‡]	51.9	-	-	-	28.6	42.3	11.6
	DCL*	-	28.4	37.3	-	28.4	-	10.9
	I3+C2FD [†]	-	-	47.3	-	29.1	40.2	12.8
	MCA-RG	-	33.5	40.8	-	30.0	41.1	12.8
	RA-RRG	58.5	41.7	50.7	26.7	24.9	37.9	8.0

Table 2: Results of single-view RRG evaluation on the IU X-Ray dataset. The test setting follows PromptMRG, and evaluation results of other models are referenced from the same source. Best values are highlighted in bold.

Type	Model	CheXbert			RadGraph	NLG Metrics		
		micro-F1	Macro-F1	example-F1	F1	ROUGE-L	BLEU-1	BLEU-4
Generation	R2Gen (Chen et al., 2020)	-	7.1	13.6	-	25.3	32.5	5.9
	CvT2DistilGPT2 (Nicolson et al., 2023)	-	15.5	16.8	-	27.7	38.3	8.2
	RGRG (Tanida et al., 2023)	-	18.7	18.0	-	18.0	26.6	6.3
	PromptMRG (Jin et al., 2024)	-	24.6	21.1	-	28.1	40.1	9.8
Retrieval	M2KT (Yang et al., 2023)	-	15.1	14.5	-	26.1	37.1	7.8
	DCL (Li et al., 2023)	-	17.7	16.2	-	26.7	35.4	7.4
	RA-RRG	36.5	26.6	24.4	30.8	27.2	36.3	6.7

DETR decoder are randomly initialized, with the number of query embeddings N set to 50 and the number of decoder layers L set to 6. The model dimension of the DETR decoder and the dimension of the semantic embeddings are set to the same value of 768. In the Hungarian algorithm, we set the selection probability ratio μ to 0.5. We set the in-batch contrastive loss ratio λ to 0.1, and the selection probability threshold for semantic embedding retrieval to 0.4. For key phrase extraction described in Section 3.1, we utilize ‘Llama-3.1-70B-Instruct’, abbreviated as Llama 70B (Dubey et al., 2024). When generating radiology reports in the final step, we employ OpenAI’s GPT-4o (Hurst et al., 2024) as the LLM.

5 Results

5.1 Single-View RRG

Table 1 reports the single-view RRG results on MIMIC-CXR. We compared our method with state-of-the-art generative RRG models and retrieval-based approaches. Our method achieved state-of-the-art performance on all CheXbert metrics, with

a Macro-F1 of 41.7, outperforming Med-PaLM M 84B (39.8) by 1.9 points, while also yielding strong micro-F1 and example-F1 scores. Notably, these results were obtained without LLM fine-tuning, surpassing even fine-tuned MLLMs. On RadGraph F1, our model achieved 26.7, matching the previous best result. This indicated that RA-RRG effectively captured clinically relevant entities and relations through key phrase retrieval. We also evaluate RA-RRG with open-source LLMs (Llama-3.1-8B and 70B) for report generation; results in Appendix Table 8 show comparable performance, confirming that the framework is not dependent on proprietary LLMs.

Table 2 shows the held-out evaluation results on the IU X-Ray dataset. We followed the test setting of Jin et al. (2024), referencing evaluation results of other models from the same source. RA-RRG achieved the highest CheXbert Macro-F1 (26.6) and example-F1 (24.4) scores, indicating stronger generalization than prior methods. Consistent with the MIMIC-CXR results (Table 1), RA-RRG yielded lower NLG metric scores. While com-

Table 3: Results of multi-view RRG evaluation on the MIMIC-CXR test set. The test setting follows MAIRA-2.

Model	CheXbert			RadGraph	NLG Metrics		
	micro-F1	Macro-F1	example-F1	F1	ROUGE-L	BLEU-1	BLEU-4
Med-PaLM M 84B <small>Zero-shot</small>	50.5	37.8	-	28.3	28.7	34.6	12.4
MAIRA-2 <small>Infer: No Prior No Comp</small>	-	35.8	-	-	27.3	-	-
MAIRA-2 <small>Train: No Prior No Comp</small>	-	39.3	-	-	33.9	-	-
RA-RRG	60.6	42.2	54.3	25.8	24.2	34.1	7.0

parable BLEU-1 scores suggested preservation of key content, ROUGE-L and BLEU-4 were lower, reflecting their reliance on exact lexical overlap. The limited correlation between NLG metrics and clinical quality is further examined in Section 6.1.

Comparing CheXbert results across datasets revealed a substantial drop in Macro-F1 from 41.7 on MIMIC-CXR to 26.6 on IU X-Ray. This discrepancy was consistent with prior observations that CheXpert labels are tailored to MIMIC-CXR and may be less suitable for IU X-Ray (Irvin et al., 2019; Chen et al., 2020). Moreover, since the RadGraph model used for key phrase extraction was trained on MIMIC-CXR and CheXpert data, reduced generalization on the held-out dataset was observed. For the benefit of future research, Table 2 reports all clinical efficacy metrics for RA-RRG, including those not provided by Jin et al. (2024).

5.2 Multi-View RRG

Table 3 presents multi-view performance results on the MIMIC-CXR test set. We evaluated RA-RRG in a two-view setting alongside two comparison models. Med-PaLM M 84B, a single-view model, reported zero-shot performance in a two-view setting, although the exact test configuration was unspecified. MAIRA-2, designed for multi-study inputs, was evaluated using its reported ablations on 2,181 studies with prior information.

Med-PaLM M 84B achieved reasonable RadGraph F1 scores, likely due to its large model size, but showed substantially lower CheXbert performance than RA-RRG. MAIRA-2 performed poorly without prior information (Macro-F1 35.8), and even the trained version (39.3) fell below our RA-RRG, which achieved 42.2. Despite being trained only on single-image retrieval, RA-RRG showed effective generalization to multi-view inputs.

5.3 Hallucination Analysis

5.3.1 Comparative hallucination suppression

Table 4 compares three models—PromptMRG, a sentence-level extraction ablation (E1), and RA-

Table 4: Frequency of comparative hallucination-related terms and the percentage of model-generated reports containing each keyword.

Keyword	Keyword Count ↓			Report Inclusion Rate (%) ↓		
	PromptMRG	E1	RA-RRG	PromptMRG	E1	RA-RRG
Change	559	160	16	9.95	4.04	0.41
Unchanged	1187	5992	1	24.44	73.46	0.03
Prior	405	1141	38	7.70	24.31	0.98
Stable	431	101	4	10.91	2.49	0.10
Interval	359	69	0	6.53	1.79	0.00
Previous	920	264	20	22.65	6.17	0.52
Again	600	1280	0	12.93	27.94	0.00
Increased	280	1125	869	6.95	23.43	19.10
Improve	0	0	0	0.00	0.00	0.00
Remain	104	25	76	2.62	0.65	1.97
Worse	18	9	0	0.41	0.23	0.00
Persistent	34	449	0	0.83	11.25	0.00
Removal	49	6	8	1.11	0.16	0.21
Similar	729	180	5	15.66	4.54	0.13
Earlier	84	3	3	2.07	0.08	0.08
Decreased	516	131	31	11.28	3.21	0.80
Recurrence	0	0	0	0.00	0.00	0.00
Redemonstrate	0	0	0	0.00	0.00	0.00

RRG—using two metrics: the frequency of comparative hallucination-related keywords and the percentage of reports containing them. E1 segments reports at the sentence level rather than extracting key phrases, with details provided in Appendix A. Unlike PromptMRG and E1, both trained on sentences referencing prior studies, RA-RRG exhibited lower keyword frequency and fewer affected reports. With RadGraph and LLM-extracted key phrases, RA-RRG effectively removed unnecessary comparisons and irrelevant content. Figure 3 illustrates a representative example: while PromptMRG generated comparative expressions such as ‘compared to’ and ‘unchanged’, RA-RRG produced an accurate, concise, and hallucination-free report.

In Table 4, most comparative hallucination-related keywords appeared in less than 1% of the reports generated by RA-RRG, except for the term ‘increased’ which was found in 19.1%. While we directly adopted the term set from Ramesh et al. (2022), ‘increased’ does not necessarily imply comparison with prior studies. For example, in the phrase ‘increased interstitial marking,’ the term refers to a deviation from the normal state, which is identifiable from a single image. Our analysis

Original report	PromptMRG	RA-RRG
FINDINGS: There has been placement of an OG feeding tube which is coiled within the stomach with the tip pointing towards the fundus. Compared to the most recent prior radiograph there has been no significant change. Moderate localized right pleural effusion is unchanged. Left mid and lower lung opacities are stable. There is no pneumothorax. Cardiac silhouette is enlarged but stable.	FINDINGS: as compared to the previous radiograph the patient has received a nasogastric tube, the course of the tube is unremarkable the tip of the tube projects over the middle parts of the stomach, there is no evidence of complication notably no pneumothorax, the right pleural effusion with subsequent atelectasis has minimally increased in extent, otherwise the radiograph is unchanged.	FINDINGS: the left lung appears clear, while there is a small right pleural effusion present, atelectasis is noted at the base of the left upper lobe, a dobhoff tube is seen with the tip positioned in the stomach, no pneumothorax is detected.
	micro-F1 : 57.1 BLEU-1 : 45.4 Macro-F1 : 14.3 BLEU-4 : 10.8 RadGraph F1 : 45.9 ROUGE-L : 26.1	micro-F1 : 57.1 BLEU-1 : 35.2 Macro-F1 : 14.3 BLEU-4 : 9.0 RadGraph F1 : 49.4 ROUGE-L : 22.5

Figure 3: Example reports generated by PromptMRG (Jin et al., 2024) and RA-RRG for a single MIMIC-CXR test image. Hallucinated expressions are highlighted in red, with clinical accuracy and NLG metrics reported.

Table 5: Entity- and relation-level analysis comparing the sentence-level baseline (E1) and RA-RRG. CheXbert scores are example-based, and RadGraph scores are computed at the entity and relation levels.

Metric	E1			RA-RRG		
	Precision	Recall	F1	Precision	Recall	F1
CheXbert	0.492	0.557	0.489	0.491	0.599	0.507
RadGraph entity	0.311	0.402	0.339	0.334	0.416	0.360
RadGraph relation	0.133	0.183	0.147	0.160	0.209	0.173

revealed that 32% (278 out of 869) of ‘increased’ appeared in this specific phrase. To enable more rigorous comparative hallucination detection, the term set should be carefully curated.

5.3.2 Entity- and relation-level analysis

While keyword-based analysis captures comparative hallucinations, it fails to address fabricated clinical findings. To evaluate object-level hallucinations, we compare E1 and RA-RRG using entity- and relation-level RadGraph metrics alongside CheXbert example-based precision, recall, and F1 (Table 5). RA-RRG improved CheXbert example-based recall (0.557 \rightarrow 0.599) while maintaining comparable precision (0.492 \rightarrow 0.491), leading to a higher F1 (0.489 \rightarrow 0.507). At both the RadGraph entity and relation levels, RA-RRG improved precision, recall, and F1, indicating that RA-RRG not only suppresses comparative expressions but also reduces false-positive fabrications while improving factual coverage.

5.4 Retrieval Error Analysis

Table 6: Comparison of micro-averaged CheXbert and RadGraph F1 before and after LLM-based generation.

Stage	CheXbert (micro-averaged)				RadGraph F1
	F1	Recall	Precision	Specificity	
Retrieval-only	0.588	0.681	0.517	0.878	0.257
After Generation	0.585	0.671	0.519	0.881	0.267

To analyze how retrieval errors propagate

Table 7: Comparison of training computational cost. GPU-hours are computed as the number of GPUs multiplied by training time.

Method	Hardware	Time (hours)	Total GPU-hours
M4CXR	2 \times H100	108	208
LLaVA-Rad	8 \times A100	28	224
RA-RRG	1 \times H100	18	18

through the pipeline, we compare micro-averaged CheXbert performance and RadGraph F1 before and after LLM-based generation (Table 6). In the retrieval-only setting, retrieved key phrases are simply concatenated with period delimiters without LLM generation. After generation, recall decreased slightly, whereas precision and specificity increased marginally. The higher RadGraph F1 scores after generation indicate that the reports generated by the LLM better preserve entity-relation structures. Overall, LLM-based generation mildly pruned over-predicted findings rather than introducing additional fabrications, although retrieval errors largely propagated to the final output. A label-level analysis is provided in Appendix E.

5.5 Computational Cost

Table 7 compares the training computational cost of RA-RRG with comparable MLLM baselines. RA-RRG required only 18 GPU-hours on a single H100, compared to 208 GPU-hours for M4CXR and 224 GPU-hours for LLaVA-Rad. While H100-hours and A100-hours are not directly interchangeable, the comparison highlights that RA-RRG requires substantially fewer GPU-hours and fewer devices than multimodal LLM training pipelines.

6 Discussion

6.1 Limitations of NLG Metrics for RRG

RA-RRG consistently reports lower NLG scores than other models (Sections 5.1, 5.2) because it focuses on key phrase extraction, omitting irrelevant details such as view positions or comparisons with prior images, which reduces lexical overlap while preserving clinical relevance. We argue that clinical accuracy should be prioritized over NLG metrics for RRG. NLG metrics rely on surface-level similarity and can be misleading. For example, given a ground truth report ‘Bilateral pleural effusions are present’, a generated report ‘Bilateral pleural effusions are not present’ receives high NLG scores (ROUGE-L: 0.9242, BLEU-1: 0.8333, BLEU-4: 0.5373) despite contradicting the reference text.

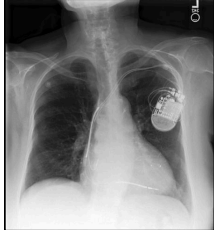
	Original report	FINDINGS: Left-sided AICD/pacemaker device is noted with leads terminating in the right atrium, right ventricle, and coronary sinus, unchanged. Mild enlargement of the cardiac silhouette is stable, with aortic knob calcifications re-demonstrated. The pulmonary vascularity is normal, and the lungs are clear. No pleural effusion or pneumothorax is present. There are mild degenerative changes in the thoracic spine with anterior bridging osteophytes.
	Baseline (E1)	FINDINGS: The cardiac silhouette is mildly enlarged and stable since . Mild degenerative changes are seen throughout the thoracic spine. A left-sided pacemaker/AICD is in unchanged position with the leads projecting over the right atrium right ventricle and coronary sinus. PA and lateral views of the chest were obtained. Calcifications are noted at the aortic arch. Mediastinal and hilar contours are normal and unchanged. The lungs appear clear. There is no pulmonary edema. There is no pleural effusion pneumothorax or acute consolidation. Osseous structures demonstrate no acute abnormalities (RadGraph F1: 57.4)
	RA-RRG	FINDINGS: There are calcifications at the aortic arch observed. A left-sided pacemaker device/AICD is present with leads positioned in the right atrium and right ventricle, specifically noted in the right ventricle. The lungs are clear, and the cardiac silhouette exhibits mild enlargement, indicating mild cardiomegaly. Thoracic spine shows mild degenerative changes. The pleural spaces are free from effusion, and there is no evidence of pulmonary edema or pneumothorax. Mediastinal and hilar contours remain unremarkable, with no focal consolidation identified. (RadGraph F1: 59.8)

Figure 4: Example of single-view RRG. The baseline is sentence-level ablation (E1) from Table 8 in Appendix. Positive findings are highlighted in yellow, and hallucinations are marked in red.

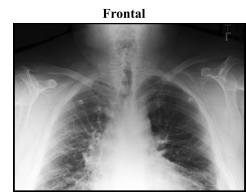
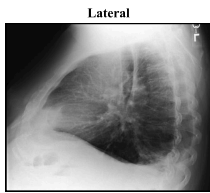
	Retrieved Key Phrases		Retrieved Key Phrases
	<ul style="list-style-type: none"> mild cardiomegaly hyperinflated lung volumes no acute osseous abnormalities moderately enlarged cardiomeastinal silhouette no pleural effusion mild interstitial pulmonary edema mild pulmonary vascular congestion no pneumothorax no focal consolidation 		<ul style="list-style-type: none"> mild cardiomegaly no acute osseous abnormalities small bilateral pleural effusions linear opacities in the right lower lobe suggestive of atelectasis opacity at the left lower lung base patchy opacity suggestive of pneumonia in the right lower lobe moderately enlarged cardiomeastinal silhouette opacity at the left lower lobe mild pulmonary edema mild pulmonary vascular congestion no pneumothorax mediastinal contours unremarkable
Original report	FINDINGS: There are low lung volumes. This accentuates the size of the cardiac silhouette which is likely top normal. There is crowding of the bronchovascular structures but no evidence of pulmonary edema. The mediastinal and hilar contours are otherwise within normal limits. Previously described subpleural left lower lobe opacity seen on prior chest radiograph which corresponds to an area of pleural fat on CT appears more prominent on the current exam. Bilateral patchy opacities in the lung bases may reflect areas of infection or atelectasis. There are small bilateral pleural effusions. No pneumothorax is identified and there are no acute osseous abnormalities.		
RA-RRG (Frontal view)	FINDINGS: The chest x-ray reveals mild cardiomegaly and a moderately enlarged cardiomeastinal silhouette. There are hyperinflated lung volumes, with mild interstitial pulmonary edema and pulmonary vascular congestion present. No pleural effusion, pneumothorax, or focal consolidation is noted, and no acute osseous abnormalities are identified. (RadGraph F1: 20.5)		
RA-RRG (Multi-view)	FINDINGS: The imaging shows mild cardiomegaly with a moderately enlarged cardiomeastinal silhouette. There is evidence of mild interstitial pulmonary edema and mild pulmonary vascular congestion. The lungs appear hyperinflated, and there are no acute osseous abnormalities visible. There is no pneumothorax present. The frontal view indicates no pleural effusion, however, the lateral view notes small bilateral pleural effusions. The lateral view also reveals linear opacities in the right lower lobe and patchy opacity suggestive of pneumonia, as well as an opacity at the left lower lung base. Opacity in the left lower lobe is further seen, likely suggestive of atelectasis. The mediastinal contours appear unremarkable. (RadGraph F1: 35.1)		

Figure 5: Example of multi-view RRG. At the top are the frontal and lateral images with their predicted key phrases. Below the original report, two radiology reports are generated: 1) using only the frontal view, and 2) using both the frontal and lateral views (multi-view). Content present in the original report but visible only in the lateral view is highlighted in yellow.

RA-RRG rephrases retrieved key phrases using an LLM, which can lower NLG scores despite clinical correctness. As shown in Figure 3, RA-RRG achieved lower NLG scores than PromptMRG while matching its CheXbert scores, indicating that both models captured key findings and that the NLG gap mainly reflects structural variation rather than clinical errors. Moreover, its higher RadGraph F1 score suggests better preservation of clinical relations despite rephrased text.

6.2 Qualitative Analysis

Figure 4 presents an example of single-view RRG. For comparison, we also include the output of E1, a sentence-level extraction baseline described in Section 5.3. Both E1 and RA-RRG accurately predicted positive findings highlighted in yellow, such as enlarged cardiac silhouette and calcification. However, E1 exhibited hallucinations by generating comparative expressions such as “unchanged” and referencing multiple views, despite being given only a single frontal image. In contrast, RA-RRG avoided such hallucinations and produced a concise, focused report consistent with the input.

Figure 5 shows an example of multi-view RRG, comparing reports generated using only the frontal-view key phrases with those incorporating both frontal and lateral views. When using only the

frontal view, the model missed pleural effusion and opacity-related findings. By incorporating lateral-view key phrases, the model correctly identified bilateral pleural effusion, opacity, and suspected atelectasis. Although the multi-view report missed suspected pneumonia, it showed improved diagnostic performance over the frontal-only report.

7 Conclusion

In this study, we introduced RA-RRG, a retrieval-augmented framework for RRG that leveraged LLMs. By extracting clinically essential key phrases and retrieving image-consistent phrases, RA-RRG effectively suppressed hallucinations and generated clinically faithful reports. Experimental results demonstrated that RA-RRG achieved strong performance on standard clinical metrics, remaining competitive with fine-tuned multimodal LLMs without requiring any LLM fine-tuning, while using substantially fewer computational resources. Analysis at the entity and relation level further confirmed that RA-RRG reduces hallucinations beyond comparative expressions. The proposed framework naturally generalized to multi-view RRG by aggregating phrases retrieved from multiple images. Accordingly, this retrieval-based paradigm enabled the extension of RRG to broader clinical settings without additional model training.

Limitations

Despite the robust performance of RA-RRG, the proposed framework has several limitations. First, RA-RRG relies on LLMs for key phrase extraction, and the quality of the extracted phrases can be sensitive to the capability of the underlying LLM. Our ablation analysis in Appendix A indicates that the RRG stage itself does not heavily depend on LLM performance; however, errors or omissions in key phrase extraction may propagate to subsequent retrieval and report generation stages. In addition, retrieval errors cannot be corrected during report generation, since the LLM conditions its output on the retrieved key phrases. While this design helps suppress hallucinations, it also limits the model’s ability to recover from retrieval-stage errors. Future work could explore lightweight refinement mechanisms to mitigate error propagation without reintroducing hallucinations.

Our evaluation is subject to limitations regarding coverage and clinical assessment. The phrase-level vector database is constructed from key phrases extracted solely from the training set, which may limit the system’s ability to handle out-of-vocabulary findings. In addition, our evaluation lacks human assessment by radiologists; while we report standard clinical and automatic metrics, expert evaluation remains necessary to rigorously assess clinical correctness, usefulness, and readability in real-world settings. Incorporating broader phrase coverage and human evaluation will be essential for validating the framework in real clinical practice.

Acknowledgements

This work was supported by the Technology Innovation Program (RS-2025-02221011, Development of Medical-Specialized Multimodal Hyper-scale Generative AI Technology for Global Integration) funded by the Ministry of Trade Industry & Energy (MOTIE, South Korea), and by the faculty research fund of Sejong University in 2026.

References

Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, and 1 others. 2024. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey

Zagoruyko. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P Langlotz. 2024. Chexpert plus: Hundreds of thousands of aligned radiology texts, images and patients. *arXiv preprint arXiv:2405.19538*.

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, Hany Awadalla, Julia Gong, Houdong Hu, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Yu Gu, Cliff Wong, Mu Wei, and 8 others. 2025. A clinically accessible small multimodal radiology model and evaluation metric for chest x-ray findings. *Nature Communications*, 16(1):3108.

Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W. Cohen. 2023. Re-imagen: Retrieval-augmented text-to-image generator. In *The Eleventh International Conference on Learning Representations*.

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)*, pages 1439–1449. Online. Association for Computational Linguistics.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, Emily B. Tsai, Andrew Johnston, Cameron Olsen, Tanishq Mathew Abraham, Sergios Gatidis, Akshay S. Chaudhari, and Curtis Langlotz. 2024. Chexagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Steven E Shooshan, Louis Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, and 1 others. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

- Mark Endo, Rayan Krishnan, Viswesh Krishna, Andrew Y Ng, and Pranav Rajpurkar. 2021. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In *Machine Learning for Health*, pages 209–219. PMLR.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.
- Bernal Jimenez Gutierrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. **HippoRAG: Neurobiologically inspired long-term memory for large language models**. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. **A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions**. *ACM Trans. Inf. Syst.*, 43(2).
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Stephanie L. Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Mercy Ranjit, Anton Schwaighofer, Fernando Pérez-García, Valentina Salvatelli, Shaury Srivastav, Anja Thieme, Noel Codella, Matthew P. Lungren, Maria Teodora Wetscherek, Ozan Oktay, and Javier Alvarez-Valle. 2023. Maira-1: A specialised large multimodal model for radiology report generation. *arXiv preprint arXiv:2311.13668*.
- Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. **Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison**. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI’19/IAAI’19/EAAI’19. AAAI Press.
- Neel Jain, Ping yeh Chiang, Yuxin Wen, John Kirchenbauer, Hong-Min Chu, Gowthami Somepalli, Brian R. Bartoldson, Bhavya Kailkhura, Avi Schwarzschild, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2024. **NEFTune: Noisy embeddings improve instruction finetuning**. In *The Twelfth International Conference on Learning Representations*.
- Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, Du Nguyen Duong Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew Lungren, Andrew Ng, Curtis Langlotz, Pranav Rajpurkar, and Pranav Rajpurkar. 2021. **Radgraph: Extracting clinical entities and relations from radiology reports**. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.
- 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 EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019a. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.
- Alistair EW 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. 2019b. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.
- Ming Kong, Zhengxing Huang, Kun Kuang, Qiang Zhu, and Fei Wu. 2022. Transq: Transformer-based semantic query for medical report generation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 610–620, Cham. Springer Nature Switzerland.
- Harold W Kuhn. 1955. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.
- Jiaxuan Li, Duc Minh Vo, Akihiro Sugimoto, and Hideki Nakayama. 2024. Evcap: Retrieval-augmented image captioning with external visual-name memory for open-world comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13733–13742.
- 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3334–3343.

- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Chang Liu, Yuanhe Tian, Weidong Chen, Yan Song, and Yongdong Zhang. 2024. [Bootstrapping large language models for radiology report generation](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):18635–18643.
- Aaron Nicolson, Jason Dowling, and Bevan Koopman. 2023. [Improving chest X-ray report generation by leveraging warm starting](#). *Artificial Intelligence in Medicine*, 144:102633.
- Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, and 7 others. 2024. [DINOv2: Learning robust visual features without supervision](#). *Transactions on Machine Learning Research*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Jonggwon Park, Soobum Kim, Byungmu Yoon, Jihun Hyun, and Kyoyun Choi. 2025. [M4cxr: Exploring multitask potentials of multimodal large language models for chest x-ray interpretation](#). *IEEE Transactions on Neural Networks and Learning Systems*, 36(10):17841–17855.
- Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Matthew P. Lungren, Maria Teodora Wetscherek, Noel Codella, Stephanie L. Hyland, Javier Alvarez-Valle, and Ozan Oktay. 2025. [Exploring scalable medical image encoders beyond text supervision](#). *Nature Machine Intelligence*, 7(1):119–130.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Vignav Ramesh, Nathan A. Chi, and Pranav Rajpurkar. 2022. [Improving radiology report generation systems by removing hallucinated references to non-existent priors](#). In *Proceedings of the 2nd Machine Learning for Health symposium*, volume 193 of *Proceedings of Machine Learning Research*, pages 456–473. PMLR.
- Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjieva. 2023. Smallcap: lightweight image captioning prompted with retrieval augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2840–2849.
- Mercy Ranjit, Gopinath Ganapathy, Ranjit Manuel, and Tanuja Ganu. 2023. Retrieval augmented chest x-ray report generation using openai gpt models. In *Machine Learning for Healthcare Conference*, pages 650–666. PMLR.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2022. Retrieval-augmented transformer for image captioning. In *Proceedings of the 19th international conference on content-based multimedia indexing*, pages 1–7.
- Min Shi, Fuxiao Liu, Shihao Wang, Shijia Liao, Subhashree Radhakrishnan, Yilin Zhao, De-An Huang, Hongxu Yin, Karan Sapra, Yaser Yacoob, Humphrey Shi, Bryan Catanzaro, Andrew Tao, Jan Kautz, Zhiding Yu, and Guilin Liu. 2025. [Eagle: Exploring the design space for multimodal LLMs with mixture of encoders](#). In *The Thirteenth International Conference on Learning Representations*.
- Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. 2020. Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. In *EMNLP 2020-2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 1500–1519.
- Liwen Sun, James Jialun Zhao, Wenjing Han, and Chenyan Xiong. 2025. [Fact-aware multimodal retrieval augmentation for accurate medical radiology report generation](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 643–655, Albuquerque, New Mexico. Association for Computational Linguistics.
- Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. 2023. [Interactive and explainable region-guided radiology report generation](#). In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 7433–7442. IEEE.
- Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaeckermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutarō Tanno, Ira Ktena, and 1 others. 2024. Towards generalist biomedical ai. *NEJM AI*, 1(3):AIoa2300138.
- Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2023. Metransformer: Radiology report generation by transformer with multiple learnable expert

- tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11558–11567.
- Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. 2020. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *Computer Vision – ECCV 2020*, pages 162–178, Cham. Springer International Publishing.
- Qilong Xing, Zikai Song, Youjia Zhang, Na Feng, Junqing Yu, and Wei Yang. 2025. Mca-rg: Enhancing llms with medical concept alignment for radiology report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 380–390. Springer.
- Jilan Xu, Yifei Huang, Junlin Hou, Guo Chen, Yuejie Zhang, Rui Feng, and Weidi Xie. 2024. Retrieval-augmented egocentric video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13525–13536.
- Benjamin Yan, Ruochen Liu, David Kuo, Subathra Adithan, Eduardo Reis, Stephen Kwak, Vasanth Venugopal, Chloe O’Connell, Agustina Saenz, Pranav Rajpurkar, and 1 others. 2023. Style-aware radiology report generation with radgraph and few-shot prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14676–14688.
- Lin Yang, Shawn Xu, Andrew Sellergren, Timo Kohlberger, Yuchen Zhou, Ira Ktena, Atilla Kiraly, Faruk Ahmed, Farhad Hormozdiari, Tiam Jaroensri, and 1 others. 2024. Advancing multimodal medical capabilities of gemini. *arXiv preprint arXiv:2405.03162*.
- 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.
- Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Richard James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-Tau Yih. 2023. Retrieval-augmented multimodal language modeling. In *International Conference on Machine Learning*, pages 39755–39769. PMLR.
- Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Uruahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y Ng, and 1 others. 2023. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9).
- 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):AIoa2400640.
- Junting Zhao, Yang Zhou, Zhihao Chen, Huazhu Fu, and Liang Wan. 2024. Topicwise separable sentence retrieval for medical report generation. *IEEE Transactions on Medical Imaging*.

A Ablation Study

To assess the effectiveness of our proposed method, we conducted ablation studies, as summarized in Table 8. For the experiments where RAG was not applied (from E1 to E10), the retrieved phrases were simply concatenated to form a single report for evaluation. First, we examined the impact of text extraction levels used for training and retrieval. E1, E2, and E3 were configured to segment a report by sentences, RadGraph extraction followed by rule-based graph construction, and the proposed key phrase extraction with an LLM, respectively. E1, using full sentences, achieved the highest NLG metrics. However, CheXbert Macro-F1 scores were the lowest for E1 and improved progressively across E2 and E3. E3 also achieved the best RadGraph F1 among the three experiments, suggesting that the proposed key phrase extraction was effective in enhancing clinical efficacy metrics.

Next, we examined the impact of different image encoders. Experiments E3 to E6 employed single image encoders, while E7 applied multiple image encoders. Comparing DINOv2-based E3 and E4, other metrics showed minimal differences, but E3 (XrayDINOv2) demonstrated a noticeably higher Macro-F1. Examining CLIP-based E5 and E6, E5 (XrayCLIP) showed significantly superior results. Consequently, in configuring the multiple image encoder for E7, we combined XrayDINOv2 and XrayCLIP using channel concatenation, yielding the highest Macro-F1 of 42.0.

Finally, we assessed the impact of semantic contrastive loss and noise addition to text embedding, from E8 to E10. Compared to E7, E8 improved example-F1 by 1.0 and RadGraph F1 by

0.2. E9 achieved the highest Macro-F1 (42.5) and increased RadGraph F1 by 0.4. Applying both the loss and noise, E10 achieved the highest average CheXbert metrics, indicating overall improvement.

Experiments from E11 to E14 used various sizes of Llama and GPT-4o as the LLM model in RAG, sharing the same retrieval model as E10. Compared to E10, all experiments yielded higher RadGraph F1 scores, even for the smallest LLM (Llama 3B, E11). NLG metrics improved as well, likely due to increased lexical similarity resulting from the LLM’s ability to generate natural sentences from key phrases. Comparing across E11 to E14, neither the size nor the type of LLM had a significant impact on RRG performance, suggesting that the critical factor is likely the key phrase retrieval rather than LLM performance. Based on the ablation studies, we selected E14 with the best RadGraph F1 score of 26.7 as our final model. Since we have empirically observed that randomness does not significantly impact the generation results, likely because all essential information is included in the prompt, we executed LLM inference only once.

B Key Phrase Extraction Details

B.1 RadGraph Phrase Extraction

RadGraph extracts clinical entities and relations as a knowledge graph. It captures three types of relations: ‘located_at’, ‘suggestive_of’, and ‘modify’. To organize the extracted entities and these relations into graphs representing minimal meaningful units, we apply the following rules: entities connected by ‘modify’, which adds contextual meaning to another entity, are grouped within the same graph, while entities linked by ‘located_at’ and

Table 8: Results of the ablation study on the MIMIC-CXR all-image test set. The table summarizes experiment outcomes based on text extraction level, image encoder, extended settings, and RAG application. \mathcal{L}_{SC} represents semantic contrastive loss, and ϵ denotes text embedding noise. Best values are highlighted in bold, and second-best values are underlined.

Extraction Level	Method			Experiment	CheXbert			RadGraph	NLG Metrics			
	Image Encoder	Extended	RAG		micro-F1	Macro-F1	example-F1	F1	ROUGE-L	BLEU-1	BLEU-4	
Sentence RadGraph Phrase	XrayDINOv2	-	-	E1	57.3	37.7	48.9	24.3	26.1	37.9	10.1	
		-	-	E2	56.7	40.1	49.7	23.6	18.9	27.7	4.2	
Key Phrase	XrayDINOv2	-	-	E3	57.2	41.2	49.5	25.6	22.3	36.0	7.3	
		-	-	E4	57.7	40.1	49.7	25.1	22.4	36.1	7.2	
	XrayCLIP	-	-	E5	57.4	41.0	49.5	25.7	23.0	36.2	7.3	
		-	-	E6	47.0	26.0	38.6	20.8	20.9	30.6	4.9	
	XrayDINOv2 + XrayCLIP	-	-	E7	57.6	42.0	49.3	25.5	22.0	36.8	7.3	
		\mathcal{L}_{SC}	-	-	E8	57.7	41.7	50.3	25.7	22.4	36.9	7.5
	XrayDINOv2 + XrayCLIP	ϵ	-	-	E9	58.3	42.5	<u>50.8</u>	25.9	21.6	37.3	7.6
		$\mathcal{L}_{SC}, \epsilon$	-	-	E10	58.8	<u>42.3</u>	51.1	25.7	23.5	36.2	7.4
	XrayDINOv2 + XrayCLIP	$\mathcal{L}_{SC}, \epsilon$	Llama 3B	E11	58.2	41.7	50.5	25.8	25.3	<u>38.3</u>	8.0	
			Llama 8B	E12	58.5	42.0	50.7	26.3	24.7	36.7	7.2	
Llama 70B			E13	<u>58.6</u>	41.9	50.7	<u>26.6</u>	<u>25.4</u>	38.4	<u>8.2</u>		
GPT-4o			E14 (RA-RRG)	58.5	41.7	50.7	26.7	24.9	37.9	8.0		

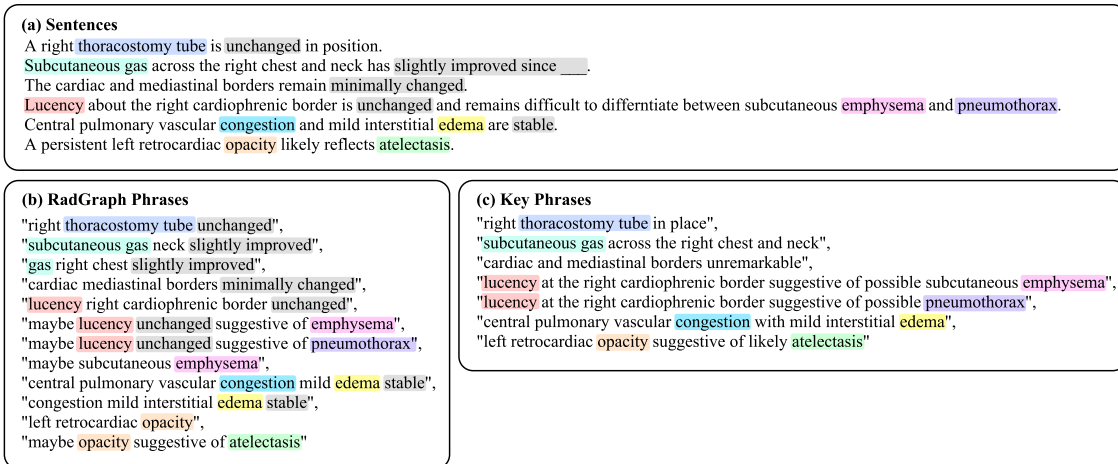


Figure 6: Example of retrieval target extraction from same radiology report as (a) sentences, (b) RadGraph phrases, and (c) key phrases. Key findings are highlighted using multiple colors, with the same color applied to identical findings. Phrases that may induce hallucinations are shown in gray.

‘suggestive_of’ are grouped into separate graphs. Each graph is then converted into a phrase. For the three types of observation-related entities (‘OBS-DA’: observation definitely absent, ‘OBS-DP’: observation definitely present, and ‘OBS-U’: observation uncertain), we prepend ‘no’ for ‘OBS-DA’ and ‘maybe’ for ‘OBS-U’. Examples of the resulting phrases, referred to as ‘RadGraph phrases’, can be found in Figure 6 (b).

B.2 LLM Prompt for Key Phrase Extraction

The input prompt to the LLM for key phrase extraction is designed to accurately extract clinically significant findings from radiology reports. These findings are then organized into natural phrases that reflect the current state. As shown in Figure 7, the input prompt instructs the LLM to identify key phrases based on the following guidelines.

First, the LLM is prompted to eliminate comparative expressions such as “improved”, “unchanged”, “worsened” and “consistent”, ensuring that the extracted key phrases reflect only information directly inferable from the current image and thereby minimizing hallucinations. Since the LLM is a general-purpose model not specialized in the medical domain, it may miss clinically important details. To mitigate this, RadGraph phrases are included in the input prompt alongside the original *FINDINGS* section. While these phrases may contain fragmented structures, they help the LLM better capture clinically meaningful content. Finally, representative examples of well-extracted key phrases are provided to more effectively guide the model in extracting clinically relevant findings.

B.3 Key Phrase Extraction Example

Figure 6 shows three possible options for retrieval targets in retrieval-based RRG: (a) sentences from the *FINDINGS* section, (b) RadGraph phrases refined with rule-based processing after RadGraph extraction and (c) the key phrases extracted from the proposed LLM prompting. A comparison between Figure 6 (b) and Figure 6 (c) highlights the effectiveness of the LLM prompting described in Section B.2.

Figure 6 (b) includes past comparative expressions such as “unchanged” and “improved” (highlighted in gray) because these expressions appear in the original report, as shown in Figure 6 (a). In contrast, Figure 6 (c) excludes such expressions, as the LLM was instructed to remove them. Additionally, Figure 6 (b) contains multiple overlapping phrases representing a single finding, such as “emphysema” (highlighted in pink) and “edema” (highlighted in yellow). In Figure 6 (c), these overlapping phrases are combined into a single key phrase that integrates all the scattered information, resulting in greater semantic clarity. These observations demonstrate that LLM prompting is effective in minimizing potential hallucinations by removing past comparative expressions and in extracting clear and concise key phrases.

C LLM Prompt for RRG

In the final step of generating the report with the LLM, an effective input prompt design is required to utilize the retrieved key phrases efficiently. Figures 8 and 9 illustrate the input prompts for the LLM in different contexts: Figure 8 shows the prompt used when a single image, either frontal or

[System Prompt]

You are an expert medical assistant AI specializing in understanding and analyzing chest x-ray radiology reports.

Your task is to extract the medically significant and meaningful findings from the given chest x-ray report, focusing on identifying phrases or expressions that describe notable conditions or abnormalities. Note that the report may reference previous studies, but we only need an interpretation based on the current chest x-ray. Therefore, remove and rewrite terms like "new", "improved", "unchanged", "worsened" or "consistent" to reflect the current status in a way that indicates the condition exists as observed in this image, without implying any comparison to prior images or studies.

Additionally, you are provided with findings generated by rule-based methods. These findings may be incomplete and may miss clinically significant observations. Your task is to review the given chest x-ray report in detail and generate a comprehensive description of the findings that includes every clinically significant observation without omitting any key observations.

Adhere strictly to the following JSON format for the final output, using examples as a guideline for the desired analysis structure. Do not provide any explanations; output only in JSON format.

[Example 1]
 INPUT:
 Cardiomegaly is accompanied by improving pulmonary vascular congestion and decreasing pulmonary edema. Left retrocardiac opacity has substantially improved, likely a combination of atelectasis and effusion. A more confluent opacity at the right lung base persists, and could be due to asymmetrically resolving edema, but pneumonia should be considered in the appropriate clinical setting. Small right pleural effusion is likely unchanged, with pigtail pleural catheter remaining in place and no visible pneumothorax.

rule-based findings:
 ["cardiomegaly", "improving pulmonary vascular congestion", "decreasing pulmonary edema", "left retrocardiac opacity substantially improved", "maybe opacity substantially improved suggestive of atelectasis", "maybe effusion", "maybe more confluent opacity suggestive of resolving edema", "maybe more confluent opacity suggestive of pneumonia", "right lung base", "maybe asymmetrically", "maybe small right pleural effusion unchanged", "pigtail pleural catheter in place", "no pneumothorax"]

OUTPUT:
 {
 "key_phrase": [
 "cardiomegaly with pulmonary vascular congestion", "pulmonary edema", "left retrocardiac opacity", "left retrocardiac opacity suggestive of likely atelectasis", "left retrocardiac opacity suggestive of likely effusion", "right lung base opacity", "right lung base opacity suggestive of possible pneumonia", "maybe small right pleural effusion", "pigtail pleural catheter in place", "no pneumothorax",
]
 }
]

[Example 2]

[User Prompt]

INPUT:
 {original report}

rule-based findings:
 {RadGraph phrases}

OUTPUT:

Figure 7: LLM prompt for key phrase extraction. The LLM extracts key phrases as a list by leveraging the original radiology report and RadGraph phrases.

lateral, is provided, whereas Figure 9 illustrates the prompt for a two-view setting with both frontal and lateral images as input. For both prompts, the LLM is required to remove any comparative expressions or references to prior study, as such expressions are definitively hallucinations given that only the current radiology data is provided.

In Figure 9, additional instructions are provided to integrate the retrieved key phrases from the two different view images into a cohesive and natural report. The system prompt directs the LLM to mention duplicate findings retrieved from both images only once. For any conflicting phrases between the

[System Prompt]

You are an expert medical assistant AI specializing in understanding and analyzing chest x-ray radiology reports.

Your task is to generate a coherent radiology report using key phrases describing findings from a single chest X-ray image as input. Please combine the phrases naturally into a comprehensive, well-phrased interpretation. Since only one image is provided, avoid any comparative expressions or mentions of previous imaging.

Adhere strictly to the following JSON format for the final output. Do not provide any explanations; output only in JSON format.

[Example]
 INPUT:
 ["cardiomegaly with pulmonary vascular congestion", "left retrocardiac opacity", "left retrocardiac opacity suggestive of likely atelectasis", "left retrocardiac opacity suggestive of likely effusion", "right lung base opacity", "right lung base opacity suggestive of possible pneumonia", "small right pleural effusion", "pigtail pleural catheter in place", "no pneumothorax",
]

OUTPUT:
 {"report": "Cardiomegaly is accompanied by pulmonary vascular congestion. There is an opacity in the left retrocardiac region, likely indicative of a combination of atelectasis and effusion. A opacity at the right lung base, potentially due to possible pneumonia. A small right pleural effusion is noted, with a pigtail pleural catheter in place, and no visible pneumothorax."}

[User Prompt]

INPUT:
 {key phrases}

OUTPUT:

Figure 8: Single-view RAG prompt for RRG. Key phrases are provided as input to generate a radiology report.

frontal and lateral view images, the retrieval result from the frontal view image takes priority. This prioritization is based on the conventional perspective that the frontal view provides more critical infor-

[System Prompt]

You are an expert medical assistant AI specializing in understanding and analyzing chest x-ray radiology reports.

Your task is to generate a coherent radiology report using key phrases describing findings from both a single frontal and a single lateral chest x-ray image as input. Please combine the phrases naturally into a comprehensive, well-phrased interpretation, reflecting findings from each view. If there are overlapping findings between the frontal and lateral views, mention such findings only once to avoid redundancy. If there is any incoherence between findings from the frontal and lateral views, prioritize findings from the frontal view as more accurate. Since only two images (one frontal and one lateral) are provided, avoid any comparative expressions or mentions of previous imaging.

Adhere strictly to the following JSON format for the final output. Do not provide any explanations; output only in JSON format.

[Example]
 INPUT:
 {"frontal": ["cardiomegaly with pulmonary vascular congestion", "left retrocardiac opacity", "right lung base opacity", "small right pleural effusion", "no pneumothorax",],
 "lateral": ["posterior lower lobe opacity suggestive of atelectasis", "no pneumothorax", "retrosternal clear space",]}

OUTPUT:
 {"report": "Cardiomegaly is accompanied by pulmonary vascular congestion. The left retrocardiac opacity is observed, with an opacity at the right lung base that may indicate a small pleural effusion. There is no visible pneumothorax. The lateral view shows a posterior lower lobe opacity, likely suggestive of atelectasis, with a clear retrosternal space."}

[User Prompt]

INPUT:
 {
 "frontal": key phrases,
 "lateral": key phrases
 }
 }

OUTPUT:

Figure 9: Multi-view RAG prompt for RRG. Key phrases retrieved from the frontal and lateral images are separately provided as input to generate a radiology report.

mation about the chest condition and includes more comprehensive diagnostic details compared to the lateral view.

D Additional Implementation Details

Vision Encoders To search for the best-performing vision encoder structure, we experiment with various CXR image encoders. These include BiomedCLIP (Zhang et al., 2025) and XrayCLIP⁴ (Chambon et al., 2024) for CLIP models, as well as RAD-DINO (Pérez-García et al., 2025) and XrayDINOv2⁴ (Chambon et al., 2024) for DINOv2 models. Although XrayDINOv2 was originally trained at an image resolution of 224, we use a resolution of 518, interpolating positional embeddings as needed. For the final model, we combine multiple image encoders, specifically XrayDINOv2 and XrayCLIP. Since XrayDINOv2 has a longer visual token sequence, we interpolate XrayCLIP’s output and concatenate them channel-wise.

Text Encoder and Retriever For the text encoder we employ MPNet (‘all-mpnet-base-v2’)⁵ (Reimers and Gurevych, 2019) with embedding dimension $d = 768$. For the distribution-balanced loss \mathcal{L}_{cls} , the hyperparameters are based on COCO-MLT experimental settings.⁶ The selection process is treated as single-label binary classification, with the positive class size set to 7.16 (the average number of key phrases as described in Section 4.1) and the negative class size fixed at $N - 7.16 = 42.84$.

Optimization We use a learning rate of 0.0002 with a cosine decay scheduler and 50 warm-up steps. The retriever is trained with a batch size of 128 for a maximum of 10 epochs, with the best model determined by validation loss. Weight decay is set to 0.05, and gradient clipping is applied with a maximum value of 1.0. The optimizer is AdamW. We train the model on a single H100 GPU for 18 hours, utilizing automatic mixed precision with bfloat16.

Large Language Models The LLMs used in this work are ‘Llama-3.1-70B-Instruct’⁷ (abbreviated

as Llama 70B (Dubey et al., 2024)) and GPT-4o⁸ (Hurst et al., 2024). For key phrase extraction (Section 3.1), radiology reports from the training data must be input into the LLM. However, licensing restrictions for the training dataset (MIMIC-CXR) explicitly prohibits sharing access to the data with third parties including sending it through APIs. To address this, we setup the open-source Llama 70B model locally to generate LLM responses instead. The sampling parameters are set to their default values: a temperature of 0.6 and a top P probability of 0.9. The vllm python package (Kwon et al., 2023) is used with 4-bit quantization for inference. In contrast, the final RRG step (Section 3.3.2) inputs general medical key phrases (e.g., ‘no pleural effusion,’ ‘mild cardiomegaly’) into the LLM rather than full reports. The key phrases extracted from reports are segmented and contain no patient-specific information, allowing RAG experiments to be conducted using GPT-4o. From a cost perspective, approximately 485 reports are generated for \$1, averaging \$0.002 per report.

Evaluation Tools We use publicly available implementations of standard evaluation metrics for the NLG metrics⁹ and RadGraph.¹⁰

E Label-Level Retrieval Error Analysis

To further analyze retrieval errors at the label level, we evaluate the retrieval-only setting (without LLM generation) by computing CheXbert metrics for each of the 14 observation labels (Table 9). Common findings such as cardiomegaly and pleural effusion tend to have higher recall but lower precision, whereas rarer findings such as lung lesion and fracture exhibit high specificity but low recall, suggesting a tendency toward under-retrieval.

F RAG Analysis

F.1 Retrieval Augmented RRG

Figure 10 visualizes the process of RA-RRG leveraging a pre-trained LLM to generate a report from the key phrases retrieved through multimodal retrieval. Phrases that correspond to the same finding are highlighted in the same color. Figure 10 (b) demonstrates that the key phrases derived from multimodal retrieval generally reflect the major findings in the original report shown in Figure 10

⁴<https://github.com/Stanford-AIMI/chexpert-plus>

⁵<https://huggingface.co/sentence-transformers/all-mpnet-base-v2>

⁶https://github.com/wutong16/DistributionBalancedLoss/blob/master/configs/coco/LT_resnet50_pfc_DB.py

⁷<https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct>

⁸‘gpt-4o-2024-08-06’ through the OpenAI API

⁹<https://pypi.org/project/pycocoevalcap>

¹⁰<https://github.com/rajpurkarlab/CXR-Report-Metric>

Table 9: Label-level CheXbert metrics in the retrieval-only setting (without LLM generation) on the MIMIC-CXR test set.

Label	F1	Recall	Precision	Specificity	Accuracy
Enlarged Cardiomediastinum	0.140	0.138	0.142	0.923	0.857
Cardiomegaly	0.703	0.844	0.602	0.684	0.742
Lung Opacity	0.612	0.681	0.556	0.675	0.678
Lung Lesion	0.237	0.191	0.313	0.969	0.915
Edema	0.512	0.744	0.390	0.782	0.776
Consolidation	0.176	0.176	0.176	0.957	0.918
Pneumonia	0.301	0.418	0.235	0.928	0.902
Atelectasis	0.551	0.713	0.448	0.688	0.695
Pneumothorax	0.382	0.615	0.277	0.967	0.960
Pleural Effusion	0.721	0.895	0.604	0.724	0.779
Pleural Other	0.235	0.204	0.277	0.978	0.948
Fracture	0.196	0.191	0.201	0.954	0.911
Support Devices	0.784	0.789	0.779	0.878	0.847
No Finding	0.371	0.397	0.349	0.953	0.920

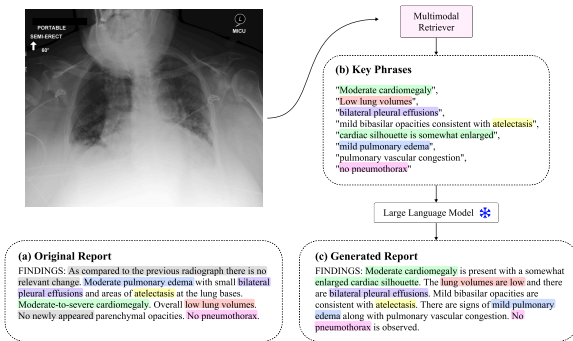


Figure 10: An example of key phrase retrieval results and the generated radiology report. Descriptions with the same meaning are highlighted in the same color, while content unsuitable for single-view RRG is shown in gray. The sample is sourced from the MIMIC-CXR test set.

(a). However, the phrase “pulmonary vascular congestion,” which is not explicitly mentioned in the original report, is added during the retrieval process, indicating false positive. Figure 10 (c) illustrates how the LLM integrates the relationships between findings naturally and generates a structured and contextually coherent radiology report based on the input key phrases. The generated report effectively incorporates the detailed information from the key phrases and preserves the major findings, consistent with the original report. Notably, the false positive phrase “pulmonary vascular congestion” was subsequently incorporated into the generated report. This reveals a limitation of RA-RRG: it inherently propagates retrieval errors into the generated report, as its quality depends on the multimodal retriever.

F.2 Effect of In-Context Example Quantity

In Figure 8, one in-context example is included in the prompt given as input to the LLM. To assess the impact of the number of in-context examples,

Table 10: Number of in-context examples for RRG.

In-context examples	micro-F1	Macro-F1	RadGraph F1	ROUGE-L	BLEU-1
0 example	58.4	41.6	26.5	24.5	35.2
1 example	58.5	41.7	26.7	24.9	37.9
3 examples	58.2	41.6	26.7	25.2	38.2

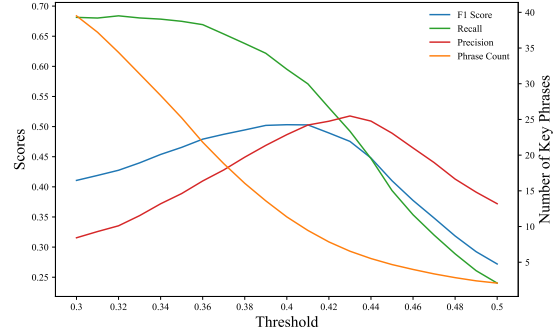


Figure 11: Impact of threshold on example-based average CheXbert scores and the number of key phrases.

we vary the number of examples provided (0, 1, and 3). Table 10 shows the results. While more context examples for RRG slightly improved the NLP metrics, there was no gain in clinical efficacy. Considering the higher cost of longer prompts, we concluded that one example suffices for a clinically accurate report.

F.3 Semantic Embedding Retrieval Threshold

The number of retrieved key phrases in the inference stage is a crucial factor that directly influences the generated report. This number varies for each image and is determined by the semantic embedding retrieval threshold. Figure 11 illustrates the average number of retrieved key phrases and the corresponding CheXbert example-based F1 score, precision, and recall across different thresholds. The semantic embedding retrieval threshold of 0.4, which we set in Section 4.3, is the value at which the example-based F1 score is maximized.

G Qualitative Examples

Figure 12 presents a comparison of the radiology reports generated by RA-RRG, MAIRA-1, and MedPaLM M 84B based on the findings in the original report. RA-RRG generally captured the findings mentioned in the original report well, particularly by providing clear descriptions of the positions of the “endotracheal tube” and “nasogastric tube” and addressing “atelectasis” appropriately. However, it omitted phrases such as “the aorta is tortuous” and introduced details absent from the original report, such as “subtle increased opacity at the left lung base may indicate possible pneumonia”. This

demonstrates RA-RRG’s ability to reflect key findings while occasionally including unnecessary details. MAIRA-1 also performed well in addressing the findings from the original report but missed “atelectasis” and inaccurately described the side port location of the “nasogastric tube,” showing limitations in certain details. Med-PaLM M 84B generally addressed most findings accurately but incorrectly described the position of the “NG tube” as extending beyond the film.

Figure 13 illustrate how accurately RA-RRG and MAIRA-1 identify the key findings from the given CXR image. RA-RRG missed findings such as “opacification likely reflects atelectasis” and “calcification”. However, it generally captured other key findings appropriately. In contrast, MAIRA-1 effectively captured the key findings but shared the same limitation in failing to mention “calcification.” Additionally, it exhibited hallucinations, such as including comparisons to prior studies that do not align with the single-view RRG or referencing unnecessary changes.

Figure 14 compares the results of RA-RRG, Med-Gemini, and MAIRA-2 for the same study, with each model performing RRG under different input scenarios. Figure 14 (a) compares the outcomes of RA-RRG and Med-Gemini on a single frontal view image. RA-RRG generally reflected the original report’s key findings, but it also added observations not present in the source, such as “moderate enlargement of the right hilus” and “prominent enlargement of the pulmonary arteries.” It also showed inconsistency with the original report by describing the location of “pleural effusion” as “bilateral,” whereas the original report indicated “right-sided.” In contrast, Med-Gemini failed to mention key findings such as “opacity in the right lower lobe” and “aortic calcifications,” which are interpreted as significant

omissions of critical pathological information. Additionally, Med-Gemini introduced unnecessary details not included in the original report, such as “mild pulmonary vascular congestion.”

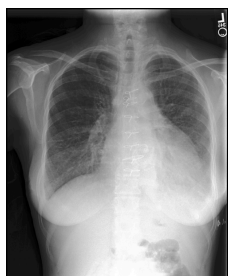
Figure 14 (b) displays the comparison of RA-RRG and MAIRA-2 after adding the lateral view from the same study as the frontal view in Figure 14 (a). It is worth noting that the radiology report of MAIRA-2 was generated using multi-view inputs along with additional prior study data. As a result, the generated results of MAIRA-2 in Figure 14 (b) include comparative expressions referencing the past, but these are not considered hallucinations and are therefore not highlighted in gray in the figure. RA-RRG, similar to its result in Figure 14 (a), exhibited errors in the location of “pleural effusion” and generated additional details absent from the original report. Meanwhile, MAIRA-2 failed to mention “right lower lobe opacity” and “aortic calcification” and was observed adding extra content not included in the original report, such as “pulmonary vascular congestion” and “mild-to-moderate pulmonary edema.”

RA-RRG demonstrated competitive performance with state-of-the-art MLLMs without requiring LLM fine-tuning and showed seamless adaptability to multi-view RRG. Additionally, the use of key phrase extraction and RAG appears to effectively suppress hallucinations. However, compared to the original reports, some false positives with additional descriptions and false negatives from missed findings were observed, highlighting the need for further improvements.



Original report	FINDINGS: Endotracheal tube terminates 6 cm from the carina. Nasogastric tube loops within the stomach with side port within the gastric body. There are no significant pleural effusions or pneumothorax. Mild retrocardiac atelectasis is present. Heart size is normal. The aorta is tortuous.
RA-RRG	FINDINGS: The lungs appear clear overall with linear opacities in the right lower lobe and retrocardiac regions, suggestive of possible atelectasis. Subtle increased opacity at the left lung base may indicate possible pneumonia. The endotracheal tube is appropriately positioned, terminating 2.3 cm above the carina. The nasogastric tube is positioned with its tip below the stomach. The heart size and cardiomeastinal silhouette are normal. There is no evidence of pleural effusion or pneumothorax.
MAIRA-1	FINDINGS: Endotracheal tube terminates 5 cm above the carina. Nasogastric tube loops in the stomach, with side port near the GE junction. Heart size is normal. Aorta is tortuous and unfolded. There are no pleural effusions or pneumothorax. Lungs are clear.
Med-Palm M 84B	FINDINGS: Endotracheal tube terminates 5 cm above the carina. Nasogastric tube courses below the diaphragm and inferiorly beyond the film. There is retrocardiac opacity which may represent atelectasis or aspiration. The heart size is normal. The aorta is tortuous and unfolded. No pleural effusions or pneumothorax.

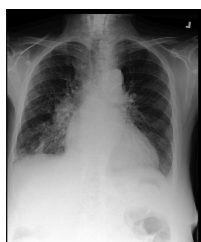
Figure 12: Comparison of single-view RRG results. Positive findings are highlighted with different colors. The sample is sourced from the MIMIC-CXR test set. Results for MAIRA-1 and Med-PaLM M 84B are referenced from Hyland et al. (2023) and Tu et al. (2024), respectively.



Original report	FINDINGS: The patient is status post median sternotomy CABG and mitral valve replacement. The heart is mildly enlarged. The mediastinal contours are unchanged with calcification of the aortic knob again noted. Mild pulmonary edema appears progressed compared to the prior exam with small bilateral pleural effusions also minimally increased compared to the prior exam. Left basilar opacification likely reflects atelectasis. There is no pneumothorax. No acute osseous abnormalities are identified.
RA-RRG	FINDINGS: The patient is status post median sternotomy with intact median sternotomy wires and a prosthetic heart valve in place, indicating status post mitral valve replacement. The cardiac silhouette is somewhat enlarged, with findings of mild to moderate cardiomegaly. There are small bilateral pleural effusions present. Mild interstitial pulmonary edema is noted. The bozny structures appear unremarkable. There is no evidence of focal consolidation or pneumothorax.
MAIRA-1	FINDINGS: The patient is status post median sternotomy and mitral valve replacement. Heart size is moderately enlarged but unchanged. The mediastinal contours are stable. There is mild interstitial pulmonary edema, new compared to the prior study. Small bilateral pleural effusions are also new. Retrocardiac opacity likely reflects atelectasis. There is no pneumothorax. No acute osseous abnormalities are visualized.

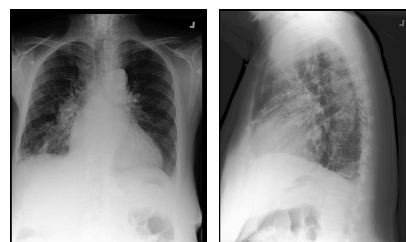
Figure 13: Comparison of single-view RRG results. Positive findings are highlighted with different colors, and phrases considered to be hallucinations are shown in gray. The sample is sourced from the MIMIC-CXR test set. MAIRA-1's result is from Hyland et al. (2023).

(a) Single-View



Original report	FINDINGS: PA and lateral chest radiographs were obtained. There is an ill-defined opacity in the right lower lobe that does not obscure the right heart border. A right-sided pleural effusion is small. There is no pneumothorax. Cardiomegaly is mild. Aortic calcifications are minimal.
RA-RRG	FINDINGS: The chest X-ray reveals a tortuous aorta with calcifications at the aortic arch. The heart appears mildly enlarged, with descriptions of mild cardiomegaly and a slightly enlarged cardiac silhouette. There is moderate enlargement of the right hilus and prominent enlargement of the pulmonary arteries. Bilateral pleural effusions are present, and mild pulmonary vascular congestion and edema are noted. The right lung base shows mild increased patchy opacity, possibly suggestive of infection, accompanied by an opacity within the right lower lobe and in the right perihilar areas. No pneumothorax is observed.
Med-Gemini	FINDINGS: There is a right lower lobe consolidation. there is also a small right pleural effusion. The heart is mildly enlarged. There is mild pulmonary vascular congestion. there is no pneumothorax.

(b) Multi-View



Original report	FINDINGS: PA and lateral chest radiographs were obtained. There is an ill-defined opacity in the right lower lobe that does not obscure the right heart border. A right-sided pleural effusion is small. There is no pneumothorax. Cardiomegaly is mild. Aortic calcifications are minimal.
RA-RRG	FINDINGS: The imaging demonstrates a tortuous aorta with calcifications at the aortic arch. Cardiomegaly is noted, with a slightly enlarged cardiac silhouette. Mild pulmonary vascular congestion is present, alongside mild pulmonary edema. Mild increased patchy opacity is observed at the right lung base, suggestive of possible infection, and there is opacity within the right lower lobe. A moderate enlargement of the right hilus and prominent enlargement of the pulmonary arteries are evident. Small bilateral pleural effusions are noted. There is no pneumothorax. The lateral view confirms the tortuous aorta, calcifications at the aortic arch, and the absence of a pneumothorax, as well as the presence of mild interstitial pulmonary edema.
MAIRA-2 (with prior study)	FINDINGS: PA and lateral views of the chest were obtained. Cardiomeastinal silhouette including mild cardiomegaly is unchanged. There is increased pulmonary vascular congestion and mild-to-moderate pulmonary edema. A small right pleural effusion is present. There is no focal consolidation or pneumothorax.

Figure 14: Comparison of (a) single-view and (b) multi-view RRG results for the same study. The report for MAIRA-2 was generated using multi-view inputs along with additional prior study information. Positive findings are highlighted with different colors. The sample is sourced from the MIMIC-CXR validation set. Results for Med-Gemini (Yang et al., 2024) and MAIRA-2 (Bannur et al., 2024) are referenced from their respective papers.