

Leveraging LLMs for Multimodal Retrieval-Augmented Radiology Report Generation via Key Phrase Extraction

Kyoyun Choi Byungmu Yoon Soobum Kim Jonggwon Park*

DEEPNOID Inc.

{kychoi, bmyoon, soobumk, jgpark}@deepnoid.com

Abstract

Automated radiology report generation (RRG) holds potential to reduce radiologists' workload, especially as recent advancements in large language models (LLMs) enable the development of multimodal models for chest X-ray (CXR) report generation. However, multimodal LLMs (MLLMs) are resource-intensive, requiring vast datasets and substantial computational cost for training. To address these challenges, we propose a retrieval-augmented generation approach that leverages multimodal retrieval and LLMs to generate radiology reports while mitigating hallucinations and reducing computational demands. Our method uses LLMs to extract key phrases from radiology reports, effectively focusing on essential diagnostic information. Through exploring effective training strategies, including image encoder structure search, adding noise to text embeddings, and additional training objectives, we combine complementary pre-trained image encoders and adopt contrastive learning between text and semantic image embeddings. We evaluate our approach on MIMIC-CXR dataset, achieving state-of-the-art results on CheXbert metrics and competitive RadGraph F1 metric alongside MLLMs, without requiring LLM fine-tuning. Our method demonstrates robust generalization for multi-view RRG, making it suitable for comprehensive clinical applications.

1. Introduction

Automated radiology report generation (RRG) can significantly alleviate the workload of radiologists. Recent advancements in large language models (LLMs) have rapidly enhanced AI's capability to assist in radiology, especially with multimodal models that interpret images and text, including chest X-rays (CXR) [4, 15, 49, 54]. However, applying multimodal LLMs (MLLMs) to CXR RRG is challenging due to high computational costs and vast data re-

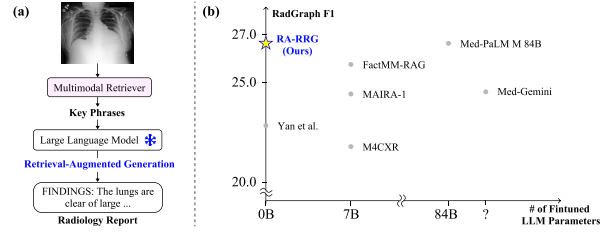


Figure 1. (a) A simplified illustration of our method. (b) Single-view RRG performance comparison between MLLMs and our model. The x-axis shows the number of parameters of fine-tuned LLMs, with 0B indicating no fine-tuning.

quirements. A promising solution to the challenges of LLM fine-tuning is retrieval-augmented generation (RAG) [26], which improves performance without training by incorporating text retrieval.

In the context of CXR RRG, multimodal retrieval-based models have been explored [10, 39], retrieving similar radiology reports based on the given image. However, due to the co-occurrence of diverse findings within radiology reports, retrieval models can introduce irrelevant information that does not align with the given image. Even when retrieval is performed at the sentence level [23, 59] rather than for the entire report, combining sentences from different reports can lead to contradictory information. Furthermore, CXR reports inherently include comparisons with prior studies, which can induce hallucinations in situation where only a single image is available.

To address these issues, we propose a novel RRG framework, RA-RRG (Retrieval-Augmented Radiology Report Generation), that combines LLMs with RAG. Building on RadGraph [18], which extracts clinical entities and relations as a knowledge graph, we use an LLM to extract key phrases, capturing only essential details. Moreover, by leveraging the instruction-following capability of LLMs when extracting key phrases, we can guide them to reflect desired qualities, such as removing references to past comparisons, thereby preventing hallucinations.

For retrieval, we apply TranSQ [23] to define seman-

*Corresponding author

tic queries, focusing on mapping meaningful semantic embeddings that align well with the text embeddings of key phrases. We introduce several training techniques to enhance the performance of the retrieval model. First, to determine the optimal choice of image encoder, we compare models of various structures and explore combining multiple image encoders. Next, to address the risk of overfitting caused by fixed text embeddings in TranSQ, we propose adding random noise during training. Finally, we incorporate in-batch contrastive loss to align text and semantic embeddings more accurately.

To generate reports from retrieval results, we utilize an LLM once again. Our approach hypothesizes that effective retrieval of accurate phrases, combined with proper prompting, enables the LLM to generate reliable radiology reports. Experimental results indicate that our method performs competitively with MLLMs while avoiding the need for vast data and computationally intensive MLLM training. Furthermore, this approach can extend to multi-view RRG, retrieving text for each image and collectively feeding the respective phrases to the LLM to produce comprehensive reports for multi-view studies.

Our main contributions are summarized as follows:

- We introduce a novel RA-RRG framework. Leveraging an LLM for key phrase extraction, our approach generates accurate reports while minimizing hallucinations.
- We enhance the retriever by combining image encoders, adding random noise, and applying contrastive learning.
- We propose a state-of-the-art multimodal RAG-based RRG approach that requires no additional LLM training.

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 [13]. Combining retrieval-augmented generation (RAG) with LLMs [11, 26] 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 [5, 56], image captioning [28, 40, 43], and video captioning [52]. This study applies 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 research has been ongoing, and MLLMs have broadened its potential applications, especially for CXR-focused models like LLaVA-Rad [4], CheXagent [7], MAIRA-1 [15], MAIRA-2 [1], and M4CXR [36]. Foundation models such as Med-Gemini [54] and MedPaLM-M

[49] also generate CXR reports as downstream tasks but come with high computational demands due to their size and data requirements.

Retrieval-based models are less affected by these issues. TranSQ [23] formulated report generation as a set prediction problem, generating semantic features to retrieve and compose sentences based on plausible clinical concerns. Teaser [59] introduced a topic-wise retrieval framework, employing a topic contrastive loss to effectively align queries with relevant report content. CXR-RePaiR [10] and CXR-ReDonE [39] aligned CXR images and report text embeddings with CLIP-based loss and ALBEF [27] models, respectively. CXR-RAG [41] combined retrieval with a pre-trained LLM for report generation, similar to our methodology. Liu et al. [31] leveraged retrieval-based in-domain adaptation and contrastive ranking, achieving precise and contextually grounded report generation through a structured coarse-to-fine decoding paradigm.

Another key approach in CXR RRG with RAG involves using RadGraph [18], which extracts report content as knowledge graphs. Yan et al. [53] serialized RadGraph outputs into text, allowing LLMs to learn radiologist-specific styles, while FactMM-RAG [47] extracted pathology-focused factual reports and applied contrastive learning. Our work builds on this by integrating an LLM to refine RadGraph’s outputs for enhanced key phrase extraction.

3. Methods

3.1. LLM-Based Key Phrase Extraction

Most previous studies that treat RRG as a retrieval task consider the entire radiology report as the target [10, 41] and/or divide the report into sentence-level segments [23, 59] for searching. However, both approaches are prone to co-occurrence issues, where multiple independent findings co-exist within a single text. Radiologist-written reports also often contain extraneous details, such as doctor names or user information.

To effectively use radiology report data for model training, we split reports into the smallest meaningful phrases and remove unnecessary information. We begin by applying RadGraph [18] to the *FINDINGS* section of radiology reports to extract entities and their relations. Through rule-based graph construction, we obtain RadGraph phrases that capture the core information of each report. For further details, refer to Appendix C.

However, RadGraph results are not always as accurate as expected: sometimes, graphs that should be connected are fragmented. Additionally, we aim to exclude terms that may indicate hallucinations. In single-image report generation, words like ‘increased’ or ‘unchanged’ are also considered such terms. Inspired by Gutierrez et al. [12], which uses LLMs for knowledge graph extraction, we also explore

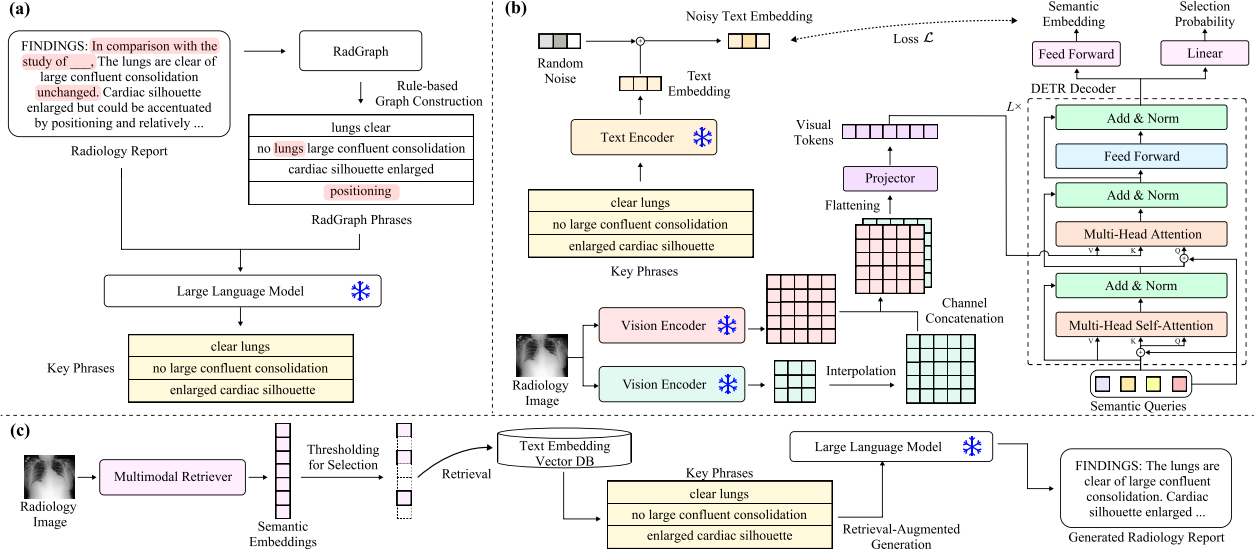


Figure 2. (a) Key phrase extraction using an LLM. (b) The multimodal retriever architecture. (c) Inference process of RA-RRG.

using LLMs for key phrase extraction. By incorporating an LLM trained on massive datasets, including medical knowledge, we can interpret reports and segment them into meaningful key phrases while filtering out hallucination-prone words associated with comparisons.

Since such robust LLMs are general-domain models and not specifically tailored for the medical domain, inputting only the original report may result in omission of essential information. Therefore, we provide both the original report and the graphs extracted by RadGraph as input to the LLM. The input prompt and examples of key phrases are shown in Figure 6 in the Appendix and Figure 2(a), respectively.

3.2. Multimodal Retriever

3.2.1. Model Architecture

To train a multimodal retrieval model using images and their corresponding lists of key phrases, we base our model on the architecture of TranSQ [23], which adapts the DETR [2] training approach 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 aim to leverage the full capabilities of pretrained vision encoders. Two common pretraining approaches are vision-language pretraining (e.g., CLIP [38]) and unimodal self-supervised learning (e.g., DINOv2 [34]). Since these image encoders have complementary advantages due to their distinct training approaches [19], we fuse the output features from various vision encoders rather than selecting a single encoder. Specifically, we use channel concatenation [45] to combine visual features. Each vision encoder follows a vision Transformer structure, so the output is a sequence of visual tokens. Since direct concatenation of

these sequences is not possible due to different lengths, we reshape the 1D visual token sequence into a 2D format and apply 2D interpolation to align it with other model sequence lengths. We concatenate these token sequences channel-wise to create a unified visual token sequence, leveraging the unique advantages of each image encoder.

Text encoder. Throughout training, for each image in the training dataset, key phrases extracted from the corresponding report are converted into text embeddings by the text encoder. Since we keep the text encoder frozen during training as Kong et al. [23], the text embeddings for a training image remain fixed, which can result in overfitting. Inspired by NEFTune [17], which adds random noise to embedding vectors when fine-tuning LLMs, we also apply random noise to the text embeddings only during training. This noise ϵ is sampled from a uniform distribution in the range $[-1, 1]$ and scaled by $1/\sqrt{d}$ for the embedding dimension d . Additionally, we apply L2 normalization to the text embeddings. For inference, we build a vector database of embeddings from all key phrases in the training dataset to facilitate retrieval.

DETR decoder. Similar to TranSQ, we use the original DETR decoder structure. The visual token sequence from the vision encoder serves as the encoder sequence, while N query embeddings are decoded in parallel via self-attention and encoder-decoder attention. To align the feature dimension of the visual token sequence with the decoder’s dimension, we apply a linear layer for projection. A selection classifier, consisting of a linear layer, computes selection logits, and semantic embeddings are calculated through a 3-layer feed-forward network with ReLU activation. Each semantic embedding is L2-normalized.

3.2.2. Loss Function

TranSQ loss. Similar to DETR, TranSQ applies the Hungarian algorithm [24] based on selection probability 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 configure N to exceed the number of key phrases, requiring the addition of empty elements $\emptyset$ to y to match the element count to N . For these two sets of N elements each, we use the Hungarian algorithm to find the permutation $\sigma \in \mathfrak{S}_N$ that minimizes the sum of matching costs:

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

The matching cost $\mathcal{L}_{\text{match}}(y_i, \hat{y}_{\sigma(i)})$ is calculated by multiplying -1 with the sum of the selection probability $\hat{p}_{\sigma(i)}$ and the cosine similarity $\mathcal{L}_{\text{sim}}$ between the text embedding v_i and the semantic embedding $\hat{v}_{\sigma(i)}$, where μ is a scaling factor for the selection probability term:

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

Based on the optimal assignment results, we compute the selection loss $\mathcal{L}_{\text{cls}}$ using binary classification labels $c_i = \mathbb{1}_{\{y_i \neq \emptyset\}}$, applying distribution-balanced loss [51] as $\mathcal{L}_{\text{cls}}$. Additionally, we compute a negative cosine similarity loss to bring the matched semantic and text embeddings closer.

In summary, the TranSQ loss is defined as follows:

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

In-batch semantic contrastive loss. The similarity loss $\mathcal{L}_{\text{sim}}$ only ensures that assigned semantic and text embeddings are similar, without explicitly preventing different semantic embeddings from becoming too similar. To address this, we adopt the CLIP loss [38], which pulls assigned embeddings closer while pushing non-assigned embeddings apart.

Within a mini-batch of size B , the set of embedding pairs E consists of text embeddings v_i^b and their corresponding semantic embeddings $\hat{v}_{\hat{\sigma}(i)}^b$, where b denotes the batch index. The set is defined as follows:

$$E = \{(v_i^b, \hat{v}_{\hat{\sigma}(i)}^b) \mid y_i^b \neq \emptyset, b = 1, \dots, B\}.$$

We treat these pairs as positives, while the remaining unmatched embeddings in E serve as negatives for the CLIP loss. However, treating negatives as hard negatives within the batch may unintentionally classify identical or similar

key phrases as negatives. To address this, we use similarity-based targets by calculating inner products in semantic and text embeddings, respectively, employing an open-source implementation of CLIP [44] as $\mathcal{L}_{\text{SC}}$.

The total loss $\mathcal{L}$ is computed as the sum of the TranSQ loss and the in-batch semantic contrastive loss with ratio λ :

$$\mathcal{L} = \sum_{b=1}^B \mathcal{L}_{\text{TranSQ}}(y^b, \hat{y}^b) + \lambda \mathcal{L}_{\text{SC}}(E) \quad (4)$$

3.3. Multimodal RAG-Based 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 our RRG approach involves using an LLM to generate a complete radiology report from the retrieved key phrases. Although the retrieved phrases contain essential information about the given radiology image, they are not complete sentences suitable for the report. We leverage the LLM’s ability to integrate the content of these phrases and generate natural sentences, yielding a coherent and comprehensive radiology report. Additionally, to ensure desired report qualities, such as hallucination removal, we include specific instructions as part of the prompt, along with key phrases, as input to the LLM. The input prompt is shown in Figure 8 in the Appendix.

Moreover, using an LLM significantly broadens the range of tasks that the proposed RA-RRG can be applied to. Beyond writing report from a given single image, RA-RRG can be applied to analyzing both frontal and lateral images or comparing current images with prior radiology images and reports for follow-up assessments, in which real clinical scenarios often involve. Unlike models that rely solely on MLLMs and require structural design adjustments to process multiple images as input [1], our approach can easily handle the new task. We extract key phrases from each image individually, then input the retrieved key phrases along with contextual details such as view position into the LLM. This method allows for the straightforward generation of a unified report that includes descriptions for each image.

4. Experiments

4.1. Datasets

For training and evaluation, we use the MIMIC-CXR dataset [21, 22], which contains 227,835 studies with a to-

tal of 377,110 DICOM images, each study accompanied by a radiology report. We utilize the official MIMIC-CXR codebase to extract only the *FINDINGS* section from each report. Retaining only the studies with a *FINDINGS* section results in 270,790 training images, 2,130 validation images, and 3,858 test images, based on the official split. Both frontal images (PA, AP) and lateral images are used for training. Studies with empty RadGraph phrase or key phrase are excluded, as these are typically cases without clinically meaningful reports. After excluding these cases, we use 269,241 images for training and 2,113 for validation. To ensure fair comparisons with various models, we retain all 3,858 images for the test set. Additionally, we prepare separate test sets for two different settings: one that uses all 3,858 DICOM images without view position filtering (all image), and another that includes only 2,461 frontal DICOM images (frontal).

We evaluate multi-view RRG using a two-view setting that takes both frontal and lateral images as input. Out of the 2,461 frontal images in the test set, 1,116 have a corresponding lateral view, and 2,181 have an prior study [1]. For cases where multiple lateral images exist for a single frontal image, one is selected randomly. To analyze the results and facilitate comparison, we conduct evaluations on three test subsets: 1,116, 2,181, and 2,461 studies. If a lateral view is unavailable, we perform the evaluation in a single-view using only the frontal image.

The statistics of the extracted key phrases are as follows: each DICOM image in MIMIC-CXR is paired with an average of 7.16 key phrases. A total of 243,064 unique key phrases were extracted from the training dataset, indicating considerable redundancy, as many key phrases recur across images within the dataset.

For held-out external evaluation, we use the IU X-ray dataset [8]. (Details in Appendix B.)

4.2. Evaluation Metrics

Both natural language generation (NLG) and clinical efficacy metrics are used for evaluation. For NLG metrics, we employ ROUGE-L [30] and BLEU scores (BLEU-1, BLEU-4) [35] to assess lexical similarity. For clinical efficacy evaluation, we calculate the F1 score using CheXbert [46], which labels radiology reports across 14 observation classes as positive, negative, uncertain, or absent. We treat all labels except positive as negative, converting them into binary classes. To facilitate comparison with other models, we compute five F1 scores. Micro-averaged F1 (mF1) and macro-averaged F1 (MF1) scores are evaluated across either all 14 observations (mF1-14, MF1-14) or the five major observations (mF1-5, MF1-5): atelectasis, cardiomegaly, consolidation, edema, and pleural effusion. Example-based F1 (eF1) score is obtained by calculating the F1 score for each example and averaging these scores [33].

Building on the entities and relations extracted from radiology reports by RadGraph [18], Yu et al. [57] introduced two additional metrics for assessing clinical efficacy: RadGraph F1 and RadCliQ. RadGraph F1 is obtained by calculating the F1 scores for entities and relations and then averaging these scores. RadCliQ is a composite metric that combines BLEU, BERTScore, CheXbert vector similarity, and RadGraph F1. To compute the metrics, we use the official RadCliQ implementation code.

4.3. Implementation Details

To search for the best-performing vision encoder structure, we experiment with various CXR image encoders. These include BiomedCLIP [58] and XrayCLIP [3] for CLIP models, as well as RAD-DINO [37] and XrayDINOv2 [3] 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.

For the text encoder we employ MPNet (‘all-mpnet-base-v2’) [42] with an embedding dimension of 768. Both vision and text encoder parameters are frozen during training. The parameters of 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 use ‘Llama-3.1-70B-Instruct’, abbreviated as Llama 70B [9]. When generating radiology reports in the final step, we employ OpenAI’s GPT-4o [14] as the LLM.

5. Results

5.1. Single-View RRG

Table 1 presents the single-view RRG evaluation results on the MIMIC-CXR dataset. We compared our model with state-of-the-art RRG language models [4, 15, 20, 36, 49, 50, 54] and retrieval-based models [23, 29, 31, 47, 53]. Since many models were not publicly available, we used evaluation metric scores from the respective papers, which resulted in variations in test datasets.

Our model achieved state-of-the-art performance on all CheXbert metrics, with an MF1-14 score of 43.5, an improvement of 3.5 points over M4CXR’s 40.0. Both mF1-14 and eF1-14 scores were high, indicating improved

Type	Model	Sections	Test Images	CheXbert					RadGraph F1	RadCliQ _(↓)	NLG Metrics		
				mF1-14	mF1-5	MF1-14	MF1-5	eF1-14			ROUGE-L	BLEU-1	BLEU-4
Generation	METransformer [†] [50]	F	3,269	-	-	-	-	31.1	-	-	29.1	38.6	12.4
	PromptMRG [20]	F	3,858	-	-	38.1	-	47.6	-	-	26.8	39.8	11.2
	Med-PaLM M 84B [49]	F	4,834	53.6	57.9	39.8	51.6	-	26.7	-	27.3	32.3	11.3
	MAIRA-1 [15]	F	2,461	55.7	56.0	38.6	47.7	-	24.3	3.10	28.9	39.2	14.2
	LLaVA-Rad [4]	F	2,461	57.3	57.4	39.5	47.7	-	-	-	30.6	38.1	15.4
	Med-Gemini [54]	F + I	912	-	-	-	-	-	24.4	-	28.3	-	20.5
	M4CXR [36]	F	2,461	60.6	61.8	40.0	49.5	53.6	21.8	-	28.5	33.9	10.3
Retrieval	TransQ ^{†‡} [23]	-	5,159	51.9	-	-	-	-	-	-	28.6	42.3	11.6
	DCL* [29]	F	3,858	-	-	28.4	-	37.3	-	-	28.4	-	10.9
	Yan et al. [53]	F + I	2,799	-	-	-	-	-	22.8	3.53	-	-	-
	Liu et al. [†] [31]	-	5,159	-	-	-	-	47.3	-	-	29.1	40.2	12.8
	FactMM-RAG [47]	F + I	1,624	-	60.2	-	-	-	25.7	-	30.7	-	-
	RA-RRG	F	3,858	58.5	62.1	41.7	52.9	50.7	26.7	3.18	24.9	37.9	8.0
	RA-RRG	F	2,461	60.8	62.4	43.5	53.3	54.0	26.3	3.21	24.7	37.4	7.8

Table 1. Results of single-view RRG evaluation on the MIMIC-CXR test set. Results are reported separately for cases where the generation target is the *FINDINGS* section only (F) and for both the *FINDINGS* and *IMPRESSION* sections (F+I). The frontal test set, which allows direct comparison under the same setting, is shaded in the table. * indicates results taken from the Jin et al. [20]. † refers to CheXpert labeling, and ‡ treats uncertain as positive. ↓ indicates that lower values are better. Best values are highlighted in bold.

class-wise F1 scores across the board. Notably, this result suggests that our model generated clinically accurate reports through RAG without LLM fine-tuning, outperforming even fine-tuned MLLMs. For RadGraph F1, our model scored 26.7 on the all image test set, matching the previous state-of-the-art, Med-PaLM M 84B. Our model also performed well on the frontal test set, scoring 26.3 and surpassing all other generation and retrieval models except Med-PaLM M 84B. Since RadGraph F1 considers both entities and relationships in reports, this high score suggests that our RA-RRG effectively captured essential report information through key phrase retrieval.

RA-RRG scored lower on NLG metrics because it does not replicate the exact phrasing of ground-truth reports. Unlike models trained on full reports, our model extracts key phrases, omitting irrelevant details like view position or past comparisons. While this approach ensures the generation of clinically relevant reports, it reduces the lexical overlap. Despite preserving key content, as evidenced by the comparable BLEU-1 score, our model appears to underperform on ROUGE-L and BLEU-4 due to their reliance on exact matches. The inferior RadCliQ score of our model compared to MAIRA-1 (3.21 vs. 3.10) can likewise be attributed to the same reason, as RadCliQ incorporates the NLG metric BLEU-2.

5.2. Multi-View RRG

Table 2 presents multi-view performance results on the MIMIC-CXR test set. We evaluated RA-RRG in a two-view setting alongside three comparison models. Med-PaLM M 84B, a single-view model, reported zero-shot generalization performance to a two-view setting, though the exact test set used is unspecified. MAIRA-2 [1] was designed to operate in a multi-study setting, and therefore demonstrates strong performance when prior study information is available. We brought the results of MAIRA-2’s two ablations for two-

view evaluation. MAIRA-2 used a test set of 2,181 studies with prior information. M4CXR [36] evaluated on all views of 2,461 studies by treating each frontal image independently and using all images within a study as input.

To analyze multi-view RRG performance, we evaluated RA-RRG on 1,116 studies with both frontal and lateral views. We compared single-view frontal, single-view lateral, and multi-view (both views). CheXbert F1 scores were ranked in the order of frontal, multi-view, and lateral: suggesting that critical findings are more prevalent in the frontal view, while the lateral view provides limited information. In the multi-view, key phrases from both views are combined, but errors from the lateral view can propagate, slightly lowering performance compared to the frontal view alone.

Med-PaLM M 84B showed lower CheXbert scores compared to RA-RRG’s two-view (1,116) setup, as expected because it was a zero-shot generalization. However, it achieved RadGraph F1 of 28.3, higher than RA-RRG’s 27.7. MAIRA-2’s performance without prior study information is low, with an MF1-14 of 35.8, and even the trained version with 39.3 falls below our RA-RRG, which achieves 42.2. Consistent with single-view results, the RadCliQ scores of our model are inferior to those of other models. Comparing M4CXR and our two-view (2,461) results, we find that M4CXR has a 1.1 higher mF1-14, while our MF1-14 score is 1.8 higher. This demonstrates that, although trained only on single-image retrieval, our model can also be easily applied to multi-image inputs.

5.3. Ablation Study

To assess the effectiveness of our proposed method, we conducted ablation studies, as summarized in Table 3. 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,

Model	View (Test Studies)	CheXbert					RadGraph F1	RadCliQ _(↓)	NLG Metrics		
		mF1-14	mF1-5	MF1-14	MF1-5	eF1-14			ROUGE-L	BLEU-1	BLEU-4
Med-PaLM M 84B Zero-shot [49]	two-view (-)	50.5	56.4	37.8	51.2	-	28.3	-	28.7	34.6	12.4
MAIRA-2 Infer:No Prior No Comp [1]	two-view (2,181)	-	-	35.8	-	-	-	3.18	27.3	-	-
MAIRA-2 Train:No Prior No Comp [1]	two-view (2,181)	-	-	39.3	-	-	-	2.89	33.9	-	-
M4CXR Multi-image [36]	all view (2,461)	61.1	-	41.0	-	-	-	-	-	-	-
RA-RRG	frontal view (1,116)	56.0	62.2	42.7	52.5	47.5	28.6	3.06	25.4	39.0	8.6
	lateral view (1,116)	52.2	60.4	35.6	50.3	43.3	27.7	3.10	25.4	39.4	8.7
	two-view (1,116)	54.3	60.5	41.3	51.6	46.1	27.7	3.19	24.5	32.9	7.2
	two-view (2,181)	60.6	62.3	42.2	53.0	54.3	25.8	3.28	24.2	34.1	7.0
	two-view (2,461)	60.0	61.8	42.8	53.0	53.3	25.9	3.27	24.3	34.5	7.1

Table 2. Performance evaluation results for multi-view RRG. The ‘two-view’ refers to using both frontal and lateral views, while only the frontal view is used if no lateral view is available. ‘all view’ utilizes all images within a study. Single-image RRG results for frontal and lateral views are included for comparison and shaded for distinction.

Method				Experiment	CheXbert			RadGraph F1	RadCliQ _(↓)	NLG Metrics		
Extraction Level	Image Encoder	Extended	RAG		mF1-14	MF1-14	eF1-14			ROUGE-L	BLEU-1	BLEU-4
Sentence	XrayDINOv2	-	-	E1	57.3	37.7	48.9	24.3	3.26	26.1	37.9	10.1
RadGraph Phrase	XrayDINOv2	-	-	E2	56.7	40.1	49.7	23.6	3.58	18.9	27.7	4.2
Key Phrase	XrayDINOv2	-	-	E3	57.2	41.2	49.5	25.6	3.30	22.3	36.0	7.3
	RAD-DINO	-	-	E4	57.7	40.1	49.7	25.1	3.30	22.4	36.1	7.2
	XrayCLIP	-	-	E5	57.4	41.0	49.5	25.7	3.29	23.0	36.2	7.3
	BiomedCLIP	-	-	E6	47.0	26.0	38.6	20.8	3.57	20.9	30.6	4.9
	XrayDINOv2 + XrayCLIP	-	-	E7	57.6	42.0	49.3	25.5	3.29	22.0	36.8	7.3
	XrayDINOv2 + XrayCLIP	$\mathcal{L}_{SC}$	-	E8	57.7	41.7	50.3	25.7	3.31	22.4	36.9	7.5
		ϵ	-	E9	58.3	42.5	50.8	25.9	3.30	21.6	37.3	7.6
		$\mathcal{L}_{SC}, \epsilon$	-	E10	58.8	<u>42.3</u>	51.1	25.7	3.28	23.5	36.2	7.4
	XrayDINOv2 + XrayCLIP	$\mathcal{L}_{SC}, \epsilon$	Llama 70B	E11	<u>58.6</u>	41.9	50.7	<u>26.6</u>	3.18	<u>25.4</u>	38.4	<u>8.2</u>
			GPT-4o	E12 (RA-RRG)	58.5	41.7	50.7	26.7	3.18	24.9	<u>37.9</u>	8.0

Table 3. Results of the ablation study on 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.

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, including RadCliQ, the second-best overall. However, CheXbert MF1-14 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 MF1-14. 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 MF1-14 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 eF1-14 by 1.0 and RadGraph F1 by 0.2. E9 achieved the highest MF1-14 (42.5) and increased RadGraph F1 by 0.4. Applying both the loss and noise, E10 achieved the highest average CheXbert metrics

and a RadCliQ of 3.28, indicating overall improvement.

Overall, E10 showed the best performance across most metrics. Experiments E11 and E12 shared the same retrieval model as E10, with the only difference being the LLM used for report generation. Applying RAG with two LLM models—Llama 70B for E11 and GPT-4o for E12—yielded even higher RadGraph F1 scores, showing an increase of over 0.9 points over E10, along with improvements in NLG metrics and RadCliQ. This improvement likely arises from increased lexical similarity due to the LLM’s ability to generate natural sentences from key phrases. Comparison between E11 and E12 showed minimal performance differences, indicating that both LLMs were able to handle the essential information in key phrases well. Based on the ablation studies, we selected E12 with a RadGraph F1 score of 26.7 as our final model.

5.4. Qualitative Analysis

Figure 3 presents an example of single-view RRG results. For comparison, we also include the output of model E1 from Table 3 as a baseline model. The positive findings highlighted in yellow include four observations: AICD/pacemaker device, enlargement of the cardiac silhouette, calcification, and degenerative changes in the thoracic spine. Both E1 and RA-RRG models accurately predicted these findings. However, E1’s report contains two instances of “unchanged” due to inherent hallucinations from com-

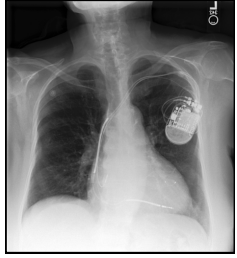
	Original report	FINDINGS: Left-sided AICD/pacemaker device is noted with leads terminating in the right atrium, right ventricle, and coronary sinus, unchanged. Mild enlargement of the cardiac silhouette is stable, with aortic knob calcifications re-demonstrated. The pulmonary vascularity is normal, and 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 (RadCliQ: 2.05)
	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. (RadCliQ: 1.95)

Figure 3. Example of single-view RRG. The baseline is model E1 from Table 3. Positive findings are highlighted in yellow, and hallucinations are marked in red.

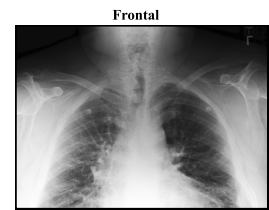
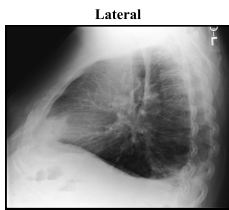
Frontal  Retrieved Key Phrases - mild cardiomegaly - hyperinflated lung volumes - no acute osseous abnormalities - moderately enlarged cardiomedistinal silhouette - no pleural effusion - mild interstitial pulmonary edema - mild pulmonary vascular congestion - no pneumothorax - no focal consolidation		Lateral  Retrieved Key Phrases - 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 cardiomedistinal 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 cardiomedistinal 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. (RadCliQ: 4.14)		
RA-RRG (Multi-view)	FINDINGS: The imaging shows mild cardiomegaly with a moderately enlarged cardiomedistinal 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. (RadCliQ: 3.06)		

Figure 4. 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.

parative language in the retrieved text. It additionally references both PA and lateral views, despite only a single frontal image being provided. In contrast, RA-RRG produced a coherent description of the findings without any hallucinations. By using RadGraph and LLM-based key phrase extraction, the model effectively removed unnecessary comparisons or irrelevant content, yielding a clean, hallucination-free report.

Figure 4 shows an example of multi-view RRG results, comparing reports generated using only the frontal view key phrases with those incorporating both frontal and lateral key phrases. Content in the original report that was retrieved as key phrases from only the lateral view image is highlighted in yellow. From the frontal view image, the model failed to capture pleural effusion and did not detect opacity-related findings. Conversely, with the lateral view, the model accurately predicted bilateral pleural effusion, retrieved key phrases indicating the presence of opacity, and suspected atelectasis. Although the resulting multi-view report is not fully comprehensive, as it lacks a mention of suspected pneumonia, it nonetheless demonstrates improved diagnostic performance compared to the frontal view report. This is further supported by the RadCliQ score, where the multi-view report scored 3.06, showing a marked improvement over the frontal view’s score of 4.14.

6. Conclusion

In this study, we introduced a novel multimodal RAG framework for RRG, RA-RRG. Leveraging the strengths of LLMs, our approach extracts essential key phrases, which are then utilized in retriever training and RAG. We identified an effective combination of image encoders for multimodal retriever, and further enhanced retrieval by incorporating noisy text embeddings and contrastive loss into training. RA-RRG achieved state-of-the-art results on clinical metrics such as CheXbert and RadGraph F1 on the MIMIC-CXR dataset, performing competitively with fine-tuned MLLMs without requiring any LLM fine-tuning. Rigorous evaluation in multi-view RRG demonstrated that our method performs comparably to MLLMs, highlighting its strong generalization capability.

Since our method functions as a RAG system without LLM fine-tuning, it can be extended to RRG scenarios that incorporate prior studies for follow-up. Additionally, this approach enables applications beyond report generation, such as LLM-based interactions for report summarization, modifications, and follow-up recommendations. Future work could include human evaluations of reports generated by the RAG system and an exploration of further applications of this method in the medical domain.

References

- [1] Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, et al. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*, 2024. 2, 4, 5, 6, 7, 8
- [2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 3
- [3] Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P Langlotz. Chexpert plus: Hundreds of thousands of aligned radiology texts, images and patients. *arXiv preprint arXiv:2405.19538*, 2024. 5
- [4] 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, Tristan Naumann, Muhao Chen, Matthew P. Lungren, Akshay Chaudhari, Serena Yeung-Levy, Curtis P. Langlotz, Sheng Wang, and Hoi-fung Poon. Towards a clinically accessible radiology foundation model: open-access and lightweight, with automated evaluation. *arXiv preprint arXiv:2403.08002*, 2024. 1, 2, 5, 6
- [5] Wenhui Chen, Hexiang Hu, Chitwan Saharia, and William W. Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022. 2
- [6] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 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, 2020. Association for Computational Linguistics. 2, 3
- [7] 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. Chexagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*, 2024. 2
- [8] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Steven E Shooshan, Louis Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016. 5, 2
- [9] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 5
- [10] Mark Endo, Rayan Krishnan, Viswesh Krishna, Andrew Y Ng, and Pranav Rajpurkar. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In *Machine Learning for Health*, pages 209–219. PMLR, 2021. 1, 2
- [11] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023. 2
- [12] Bernal Jimenez Gutierrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. HippoRAG: Neurobiologically inspired long-term memory for large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2
- [13] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023. 2
- [14] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 5
- [15] 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. Maira-1: A specialised large multimodal model for radiology report generation. *arXiv preprint arXiv:2311.13668*, 2023. 1, 2, 5, 6, 7, 8
- [16] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghighi, 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. 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 Press, 2019. 2
- [17] 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. NEFTune: Noisy embeddings improve instruction finetuning. In *The Twelfth International Conference on Learning Representations*, 2024. 3
- [18] 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. Radgraph: Extracting clinical entities and relations from radiology reports. In *Proceedings of the Neural Informa-*

- tion Processing Systems Track on Datasets and Benchmarks, 2021. 1, 2, 5
- [19] Dongsheng Jiang, Yuchen Liu, Songlin Liu, Jin'e Zhao, Hao Zhang, Zhen Gao, Xiaopeng Zhang, Jin Li, and Hongkai Xiong. From clip to dino: Visual encoders shout in multi-modal large language models. *arXiv preprint arXiv:2310.08825*, 2023. 3
 - [20] Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. Promptmrg: Diagnosis-driven prompts for medical report generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(3):2607–2615, 2024. 5, 6, 2, 3
 - [21] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019. 4
 - [22] 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. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019. 4
 - [23] Ming Kong, Zhengxing Huang, Kun Kuang, Qiang Zhu, and Fei Wu. Transq: Transformer-based semantic query for medical report generation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 610–620, Cham, 2022. Springer Nature Switzerland. 1, 2, 3, 5, 6
 - [24] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 4
 - [25] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023. 1
 - [26] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020. 1, 2
 - [27] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 2
 - [28] Jiaxuan Li, Duc Minh Vo, Akihiro Sugimoto, and Hideki Nakayama. 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, 2024. 2
 - [29] Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. 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, 2023. 5, 6, 3
 - [30] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 5
 - [31] Chang Liu, Yuanhe Tian, Weidong Chen, Yan Song, and Yongdong Zhang. Bootstrapping large language models for radiology report generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):18635–18643, 2024. 2, 5, 6
 - [32] Aaron Nicolson, Jason Dowling, and Bevan Koopman. Improving chest X-ray report generation by leveraging warm starting. *Artificial Intelligence in Medicine*, 144:102633, 2023. 3
 - [33] Aaron Nicolson, Jason Dowling, and Bevan Koopman. Improving chest x-ray report generation by leveraging warm starting. *Artificial intelligence in medicine*, 144:102633, 2023. 5
 - [34] 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, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 3
 - [35] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 2002. 5
 - [36] Jonggwon Park, Soobum Kim, Byungmu Yoon, Jihun Hyun, and Kyoyun Choi. M4cxr: Exploring multi-task potentials of multi-modal large language models for chest x-ray interpretation. *arXiv preprint arXiv:2408.16213*, 2024. 2, 5, 6, 7
 - [37] Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Matthew P Lungren, et al. Rad-dino: Exploring scalable medical image encoders beyond text supervision. *arXiv preprint arXiv:2401.10815*, 2024. 5
 - [38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 4
 - [39] Vignav Ramesh, Nathan A Chi, and Pranav Rajpurkar. Improving radiology report generation systems by removing hallucinated references to non-existent priors. In *Machine Learning for Health*, pages 456–473. PMLR, 2022. 1, 2
 - [40] Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedzhieva. Smallcap: lightweight image captioning prompted with retrieval augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2840–2849, 2023. 2

- [41] Mercy Ranjit, Gopinath Ganapathy, Ranjit Manuel, and Tanuja Ganu. Retrieval augmented chest x-ray report generation using openai gpt models. In *Machine Learning for Healthcare Conference*, pages 650–666. PMLR, 2023. 2
- [42] N Reimers. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019. 5
- [43] Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Retrieval-augmented transformer for image captioning. In *Proceedings of the 19th international conference on content-based multimedia indexing*, pages 1–7, 2022. 2
- [44] M. Moein Shariatnia. Simple CLIP, 2021. <https://github.com/moein-shariatnia/OpenAI-CLIP>. 4
- [45] 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. Eagle: Exploring the design space for multimodal LLMs with mixture of encoders. In *The Thirteenth International Conference on Learning Representations*, 2025. 3
- [46] Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. 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, 2020. 5
- [47] Liwen Sun, James Zhao, Megan Han, and Chenyan Xiong. Fact-aware multimodal retrieval augmentation for accurate medical radiology report generation. *arXiv preprint arXiv:2407.15268*, 2024. 2, 5, 6
- [48] Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-guided radiology report generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 7433–7442. IEEE, 2023. 3
- [49] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaeckermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *NEJM AI*, 1(3):AIoa2300138, 2024. 1, 2, 5, 6, 7
- [50] Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. 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, 2023. 5, 6
- [51] Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *Computer Vision – ECCV 2020*, pages 162–178, Cham, 2020. Springer International Publishing. 4
- [52] Jilan Xu, Yifei Huang, Junlin Hou, Guo Chen, Yuejie Zhang, Rui Feng, and Weidi Xie. Retrieval-augmented egocentric video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13525–13536, 2024. 2
- [53] Benjamin Yan, Ruochen Liu, David Kuo, Subathra Adithan, Eduardo Reis, Stephen Kwak, Vasantha Venugopal, Chloe O’Connell, Agustina Saenz, Pranav Rajpurkar, et al. Style-aware radiology report generation with radgraph and few-shot prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14676–14688, 2023. 2, 5, 6
- [54] Lin Yang, Shawn Xu, Andrew Sellergrén, Timo Kohlberger, Yuchen Zhou, Ira Ktena, Atilla Kiraly, Faruk Ahmed, Farhad Hormozdiari, Tiam Jaroensri, et al. Advancing multimodal medical capabilities of gemini. *arXiv preprint arXiv:2405.03162*, 2024. 1, 2, 5, 6, 8
- [55] Shuxin Yang, Xian Wu, Shen Ge, Zhuozhao Zheng, S. Kevin Zhou, and Li Xiao. Radiology report generation with a learned knowledge base and multi-modal alignment. *Medical Image Analysis*, 86:102798, 2023. 3
- [56] Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Richard James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-Tau Yih. Retrieval-augmented multimodal language modeling. In *International Conference on Machine Learning*, pages 39755–39769. PMLR, 2023. 2
- [57] Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y Ng, et al. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9), 2023. 5
- [58] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023. 5
- [59] Juntong Zhao, Yang Zhou, Zhihao Chen, Huazhu Fu, and Liang Wan. Topicwise separable sentence retrieval for medical report generation. *IEEE Transactions on Medical Imaging*, 2024. 1, 2

Leveraging LLMs for Multimodal Retrieval-Augmented Radiology Report Generation via Key Phrase Extraction

Supplementary Material

A. Implementation Details

A.1. External Source Codes

We provide the sources of all external codes referenced in Section 4 as footnotes, along with the hyperparameters if specifically configured:

- The official MIMIC-CXR codebase used to extract the *FINDINGS* section from each report.¹
- The pycocoeval python package for computing NLG metrics (e.g., BLEU, ROUGE-L).²
- The official implementation code for RadCliQ.³
- Hyperparameters for the distribution-balanced loss $\mathcal{L}_{cls}$ based on COCO-MLT experimental settings.⁴ 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 $(N - 7.16)$.
- The publicly available MPNet for text encoder.⁵
- XrayCLIP and XrayDINOv2 for vision encoders.⁶
- The Llama-3.1-70B-Instruct model.⁷ Sampling parameters are set with a temperature of 0.6 and a top P probability of 0.9, which are the default settings. The vllm python package [25] is used with 4-bit quantization for inference.
- ‘gpt-4o-2024-08-06’ as GPT-4o through the OpenAI API.

A.2. LLM Selection

The LLMs used in this work are Llama 70B and GPT-4o, both with the default sampling parameters. 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.

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 the reports are segmented and contain no patient-specific information, allowing RAG experiments to be conducted using OpenAI’s API.

LLM	mF1-14	MF1-14	RadGraph F1	ROUGE-L	BLEU-1
Llama 70B (E11)	58.6	41.9	26.6	25.4	38.4
Llama 8B	58.5	42.0	26.3	24.7	36.7
Llama 3B	58.2	41.7	25.8	25.3	38.3
GPT-4o (E12)	58.5	41.7	26.7	24.9	37.9

Table 4. Impact of various sizes of LLMs on RRG performance. E11 and E12 denote the ablation study settings in Table 3.

Table 4 presents a comparative analysis of the results obtained by employing GPT-4o and various sizes of Llama in the final RRG step. Reducing the LLM size did not significantly affect RRG performance, suggesting that the critical factor is likely the key phrase retrieval rather than LLM performance. Additionally, we have empirically observed that randomness does not significantly impact the generation results, likely because all essential information is included in the prompt. Therefore, we executed LLM inference only once. Since Llama 70B and GPT-4o (Table 3, E11 and E12) revealed no significant differences, based on the best RadGraph score, we selected GPT-4o as the LLM for RAG. Considering the API cost for GPT-4o, approximately 485 reports were generated for \$1, averaging \$0.002 per report.

A.3. Training Details for Multimodal Retriever

During multimodal retriever training, we freeze the pre-trained parameters of the vision encoders and the text encoder, while only the parameters of the DETR decoder are randomly initialized and trained. 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.

A.4. 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 5 illustrates the average number of retrieved key phrases and the corresponding CheXbert example-based F1 score, precision, and recall across different thresholds. The semantic

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

²<https://pypi.org/project/pycocoevalcap/>

³<https://github.com/rajpurkarlab/CXR-Report-Metric>

⁴https://github.com/wutong16/DistributionBalancedLoss/blob/master/configs/coco/LT_resnet50_pfc_DB.py

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

⁶<https://github.com/Stanford-AIMI/chexpert-plus>

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

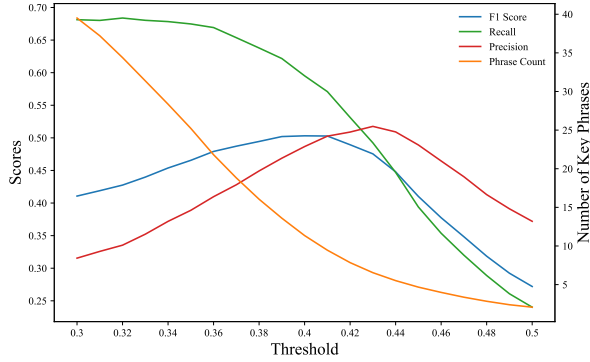


Figure 5. Impact of threshold on example-based average CheXbert scores and the number of key phrases.

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.

B. Held-out External Evaluation

B.1. Dataset

For held-out external evaluation, we use the full IU X-Ray dataset [8]. This dataset contains 2,955 studies, with each study comprising one frontal and one lateral image. Jin et al. [20] used the entire IU X-Ray dataset as the test set, treating each frontal and lateral image as an independent sample and excluding a portion of normal images to maintain a 10% normal image ratio. This subset of 4,168 images is publicly available⁸ and is also used in our evaluation to assess the performance of single-view RRG.

B.2. Results

Table 5 shows our evaluation results on the held-out IU X-Ray dataset. We followed the same test setup as Jin et al. [20], with evaluation results for the other models referenced from the same source. RA-RRG achieved the highest score on CheXbert MF1-14 and eF1-14, with values of 26.6 and 24.4, respectively, suggesting stronger generalization compared to other models. However, similar to the results on MIMIC-CXR, the proposed model exhibited relatively lower NLG metric scores, likely due to its RAG-based sentence reconstruction.

Comparing CheXbert scores between Tables 1 and 5 revealed a sharp drop in MF1-14 from 41.7 on MIMIC-CXR to 26.6 on IU X-Ray. Previous studies [6, 16] have noted that CheXpert classes are designed for MIMIC-CXR and may be less suited to IU X-Ray. Nonetheless, experimental results indicated that RA-RRG may lack generalization ability on held-out datasets, warranting further research. For the benefit of future research, Table 5 presents the evaluation results of RA-RRG across all clinical efficacy metrics, including those not reported by Jin et al. [20].

⁸<https://github.com/jhb86253817/PromptMRG>

C. Key Phrase Extraction Details

C.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 ‘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 7(b).

C.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 6, the input prompt instructs the LLM to identify key phrases based on the following guidelines.

First, the LLM is tasked with eliminating comparative expressions such as “new”, “improved”, “unchanged”, “worsened” and “consistent”. This ensures that the extracted key phrases contain only information directly inferred from the given current image, thereby minimizing hallucinations. Since the LLM is a general language model not specialized in the medical domain, it may overlook clinically important information. To address this limitation, RadGraph phrases are included in the input prompt along with the original *FINDINGS* section of the report. Although RadGraph phrases may include fragmented graphs, they remain capable of sufficiently assisting the LLM in capturing clinically meaningful findings. Finally, to ensure clarity in the output format, some well-extracted key phrase examples are provided. These examples guide the model in extracting clinically relevant findings more effectively.

C.3. Key Phrase Extraction Example

Figure 7 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 7(b) and Figure 7(c) highlights the effectiveness of the LLM prompting described in Section C.2.

Figure 7(b) includes past comparative expressions such as “unchanged” and “improved” (highlighted in gray) be-

Type	Model	CheXbert				RadGraph F1	RadCliQ _(t)	NLG Metrics		
		mF1-14	mF1-5	MF1-14	MF1-5	eF1-14		ROUGE-L	BLEU-1	BLEU-4
Generation	R2Gen [6]	-	-	7.1	-	13.6	-	25.3	32.5	5.9
	CvT2DistilGPT2 [32]	-	-	15.5	-	16.8	-	27.7	38.3	8.2
	RGRG [48]	-	-	18.7	-	18.0	-	18.0	26.6	6.3
	PromptMRG [20]	-	-	24.6	-	21.1	-	28.1	40.1	9.8
Retrieval	M2KT [55]	-	-	15.1	-	14.5	-	26.1	37.1	7.8
	DCL [29]	-	-	17.7	-	16.2	-	26.7	35.4	7.4
	RA-RRG	36.5	43.7	26.6	32.8	24.4	30.8	27.2	36.3	6.7

Table 5. Evaluation results of single-view RRG on the IU X-Ray dataset. The test setting follows PromptMRG [20], and evaluation results of other models are referenced from the same source. Best values are highlighted in bold.

cause these expressions appear in the original report, as shown in Figure 7(a). In contrast, Figure 7(c) excludes such expressions, as the LLM was instructed to remove them. Additionally, Figure 7(b) contains multiple overlapping phrases representing a single finding, such as “emphysema” (highlighted in pink) and “edema” (highlighted in yellow). In Figure 7(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.

D. LLM Prompt for RAG

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 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 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 information about the chest condition and includes more comprehensive diagnostic details compared to the lateral view. This assumption is also supported by our experimental results in Table 2, as discussed in Section 5.2.

In Figure 8, one in-context example is included in the prompt given as input to the LLM. To examine the performance differences based on the number of in-context exam-

In-context examples	mF1-14	MF1-14	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

Table 6. Number of in-context examples for RRG

ples, we conducted additional experiments by varying the number of context examples to 0, 1, and 3. Table 6 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.

E. Qualitative Examples

E.1. Key Phrase Retrieval to RAG

Figure 10 visualizes the process of RA-RRG leveraging LLM from the key phrases retrieved through multimodal retrieval. In Figure 10(a), (b), and (c), 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(a). However, the phrase “pulmonary vascular congestion,” which is not explicitly mentioned in the original report, is added during the retrieval process, suggesting the possibility of hallucination. 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.

E.2. Comparison with MLLMs

Figure 11 presents a comparison of the radiology reports generated by RA-RRG, MAIRA-1, and Med-PaLM 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

[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]

INPUT:

Frontal and lateral radiographs of the chest redemonstrate a round calcified pulmonary nodule in the posterior right lung base, unchanged from multiple priors and consistent with prior granulomatous disease. A known enlarged right hilar lymph node seen on CT of ___ likely accounts for the increased opacity at the right hilum. A known right mediastinal lymph node conglomerate accounts for the fullness at the right paratracheal region. No pleural effusion, pneumothorax or focal consolidation is present. The patient is status post median sternotomy and CABG with wires intact. The cardiac silhouette is normal in size. The mediastinal and hilar contours are unchanged from the preceding radiograph.

rule-based findings:

["round calcified pulmonary nodule unchanged", "round calcified nodule posterior right lung base unchanged", "consistent granulomatous disease", "enlarged", "right hilar", "increased opacity right hilum", "right mediastinal node conglomerate", "fullness right paratracheal region", "no pleural effusion", "no pneumothorax", "no focal consolidation", "status post median sternotomy cabg", "wires intact", "cardiac silhouette normal size", "mediastinal hilar contours unchanged"]

OUTPUT:

```
{
  "key_phrase": [
    "round calcified pulmonary nodule in the posterior right lung base", "granulomatous disease", "right hilar lymph node", "opacity right hilum", "right mediastinal lymph node conglomerate", "fullness at the right paratracheal region", "status post median sternotomy", "status post CABG with wires intact", "no pleural effusion", "no pneumothorax", "no focal consolidation", "cardiac silhouette normal in size", "mediastinal hilar contours unremarkable"
  ]
}
```

[User Prompt]

INPUT:

{original report}

rule-based findings:

{RadGraph phrases}

OUTPUT:

Figure 6. LLM prompt for key phrase extraction. The LLM extracts key phrases as a list by leveraging the original radiology report and RadGraph phrases.

(a) Sentences

A right thoracostomy tube is unchanged in position.
 Subcutaneous gas across the right chest and neck has slightly improved since ____.
 The cardiac and mediastinal borders remain minimally changed.
 Lucency about the right cardiophrenic border is unchanged and remains difficult to differentiate between subcutaneous emphysema and pneumothorax.
 Central pulmonary vascular congestion and mild interstitial edema are stable.
 A persistent left retrocardiac opacity likely reflects atelectasis.

(b) RadGraph Phrases

"right thoracostomy tube unchanged",
 "subcutaneous gas neck slightly improved",
 "gas right chest slightly improved",
 "cardiac mediastinal borders minimally changed",
 "lucency right cardiophrenic border unchanged",
 "maybe lucency unchanged suggestive of emphysema",
 "maybe lucency unchanged suggestive of pneumothorax",
 "maybe subcutaneous emphysema",
 "central pulmonary vascular congestion mild edema stable",
 "congestion mild interstitial edema stable",
 "left retrocardiac opacity",
 "maybe opacity suggestive of atelectasis"

(c) Key Phrases

"right thoracostomy tube in place",
 "subcutaneous gas across the right chest and neck",
 "cardiac and mediastinal borders unremarkable",
 "lucency at the right cardiophrenic border suggestive of possible subcutaneous emphysema",
 "lucency at the right cardiophrenic border suggestive of possible pneumothorax",
 "central pulmonary vascular congestion with mild interstitial edema",
 "left retrocardiac opacity suggestive of likely atelectasis"

Figure 7. 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.

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 12 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 13 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 13(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 “pleu-

ral 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 13(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 13(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 13(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 13(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.

[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.

[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.

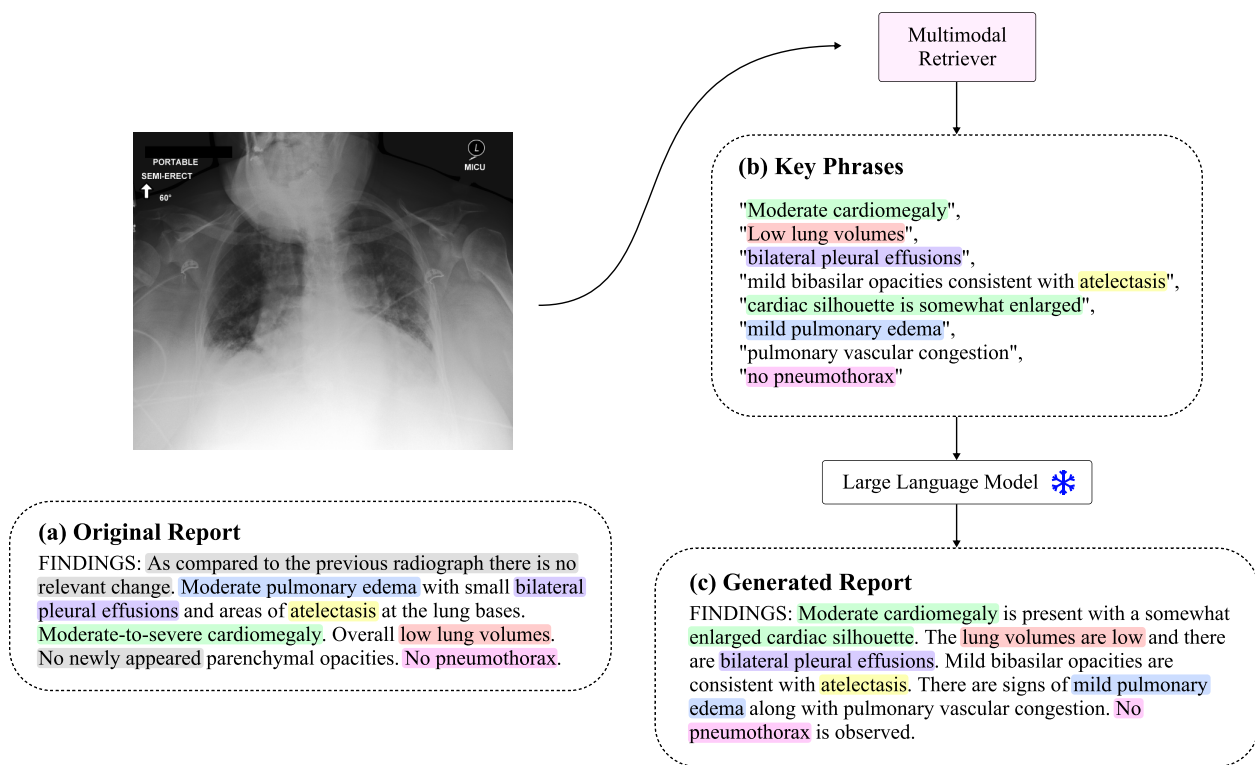


Figure 10. An example of key phrase retrieval results and the generated radiology report. Descriptions with the same meaning are highlighted in the same color, while content unsuitable for single-view RRG is shown in gray. The sample is sourced from the MIMIC-CXR test set.

	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 11. 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. [15] and Tu et al. [49], 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 bony 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 12. 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. [15].

(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. Cardiomediastinal 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 13. 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 [54] and MAIRA-2 [1] are referenced from their respective papers.