Fact-Aware Multimodal Retrieval Augmentation for Accurate Medical Radiology Report Generation

Liwen Sun*, **James Zhao***, **Megan Han, Chenyan Xiong**
School of Computer Science, Carnegie Mellon University
{liwens,jjzhao2,wenjingh,cx}@andrew.cmu.edu

Abstract

Multimodal foundation models hold significant potential for automating radiology report generation, thereby assisting clinicians in diagnosing cardiac diseases. However, generated reports often suffer from serious factual inaccuracy. In this paper, we introduce a fact-aware multimodal retrieval-augmented pipeline in generating accurate radiology reports (FactMM-RAG). We first leverage RadGraph to mine factual report pairs, then integrate factual knowledge to train a universal multimodal retriever. Given a radiology image, our retriever can identify high-quality reference reports to augment multimodal foundation models, thus enhancing the factual completeness and correctness of report generation. Experiments on two benchmark datasets show that our multimodal retriever outperforms state-of-the-art retrievers on both language generation and radiology-specific metrics, up to 6.5% and 2% score in F1CheXbert and F1RadGraph. Further analysis indicates that employing our factually-informed training strategy imposes an effective supervision signal without relying on explicit diagnostic label guidance, and successfully propagates fact-aware capabilities from the multimodal retriever to the multimodal foundation model in radiology report generation.

1 Introduction

Within hospitals worldwide, chest radiology serves as a critical technique in identifying cardiac diseases and abnormalities. Results of a chest radiograph are typically consolidated in a radiology report, including the source X-ray and a radiologist-produced findings section detailing clinical observations. Manually generating these reports, however, can be both time-consuming and potentially inaccessible in under-resourced hospitals [18, 39]. Recent multimodal foundation models have exhibited remarkable capabilities in challenging healthcare tasks, motivating an automation of this process to enhance physicians' efficiency on clinical decision-making and improve patient health outcomes [25, 31, 40, 43, 57].

Although prior medical multimodal foundation models have demonstrated promising capabilities on report generation given the radiology image, they still suffer from serious hallucinations by generating factually inaccurate reports [1, 32, 33]. Factual correctness is especially critical in chest radiology domains, as minute textual differences can drastically invert radiology report meaning and downstream prescribed treatments [7, 28, 46]. Retrieval-Augmented Generation (RAG) has emerged as a popular paradigm to address this issue by grounding text generation with retrieved relevant knowledge given a query [4, 11, 23]. However, developing medical multimodal retrievers remains challenging, requiring retrievers to bridge the gap between symptomatic image semantics and factually-equivalent report text.

*These authors contributed equally to this work.

GenAI for Health Workshop @ NeurIPS 2024, Vancouver.

To capture fine-grained details in chest radiographs and improve the factual completeness of generated reports, we introduce FactMM-RAG, a fact-aware multimodal retrieval-augmented pipeline for generating accurate radiology reports given a radiology image. By designing a novel report pair-mining procedure incorporating factual knowledge, we develop a fact-aware retriever to augment multimodal foundation models in generating accurate chest X-ray radiology reports. Specifically, we first leverage RadGraph [20] to mine factually-oriented report pairs by annotating consistent radiology entities and relations between query and reference reports with certain abnormalities. Next, we train a universal multimodal encoding architecture through mined report pairs to conduct multimodal dense retrieval. Given an unseen patient's radiology image, our retriever encodes it and searches for the most similar factually-informed reference report from an available report corpus. Passing them together into a multimodal foundation model unlocks its fact-aware potential to generate more accurate radiology reports.

Our experiments reveal that our retriever outperforms all state-of-the-art retrievers in both language generation and clinically relevant metrics on the MIMIC-CXR and CheXpert datasets, achieving up to 6.5% and 2% score in F1CheXbert and F1RadGraph for final RAG evaluation. We also investigate our retriever's fact-aware capability controlled by factual similarity thresholds and confirm that our factually-informed training strategy can impose a useful supervision signal without relying on explicit diagnostic label guidance. Further analysis through retrieval evaluation metrics shows that the fact-aware capability of our retriever can be effectively propagated to the multimodal foundation models. Lastly, our case study highlights that among reports describing the same symptom from different retrievers, those generated by our model are more accurate and achieve greater factual correctness.

Our main contributions can be summarized as follows:

- We propose a fact-aware medical multimodal retriever to augment multimodal foundation models in generating accurate chest X-ray radiology reports.
- We design a method for mining factually-informed radiology report pairs that trains multimodal encoders to retrieve high-quality reference reports.
- We demonstrate that on two benchmark datasets, our medical multimodal retriever outperforms state-of-the-art medical multimodal retrievers on both language generation and clinically relevant metrics.

The rest of this paper is organized as follows. We review related work in in Section 2. We discuss the pipeline of FactMM-RAG in Sections 3. Section 4 and 5 discuss our experimental setup and results.

2 Related Work

Retrieval Augmented Generation. Retrieval Augmented Generation, utilizing external knowledge to enhance language models, has shown great promise in text-generation performance on factual accuracy especially for Open-Domain QA. [2, 19]. Guu et al. [12], Lewis et al. [23] involve end-to-end training through both generators and retrievers; Shi et al. [37], Yu et al. [51] adapt the end-to-end pattern by employing black-box LLM training signal propagation for retriever tuning. Further works have expanded RAG to multiple modalities, employing unified image-text encoders [34] or separate pretrained encoders [9, 35] and plugging retrieved documents into multimodal foundation models [4, 15]. Yasunaga et al. [47] similarly integrates multimodal retrieval with both text and image generation capabilities.

Medical Multimodal Retriever. Joint training of image-text pairs in a shared embedding space, as exemplified by CLIP [34], facilitates visual and textual modality interactions, providing flexible representations for general-domain downstream tasks. Adapting general-domain multimodal retrievers to medical domains, however, is non-trivial due to the necessity of specialized knowledge. Zhang et al. [55] introduces an unsupervised approach for radiology image representation learning from paired text descriptions. Huang et al. [16] leverages global image-report and local sub-region features for multimodal retrieval and classification. Wang et al. [44], You et al. [48] propose medical knowledge extraction for constructing contrastive learning image-text pairs. Zhang et al. [52] addresses the limited diversity within medical datasets, curating a

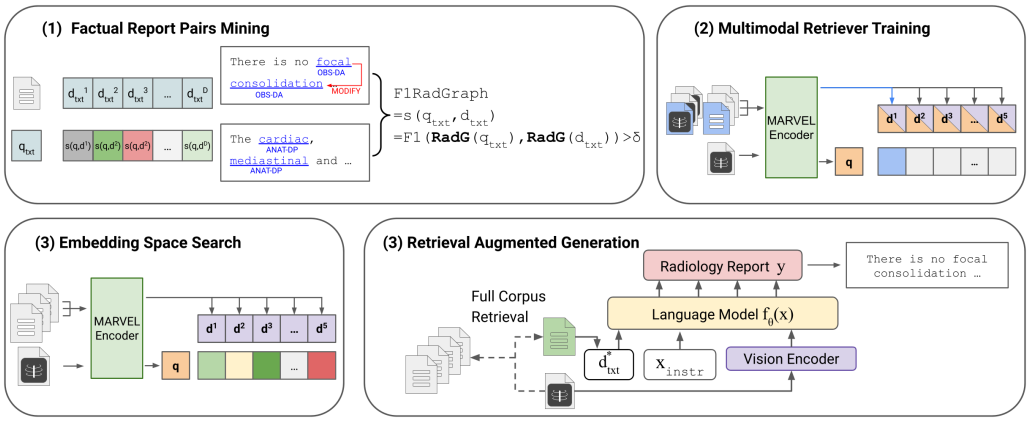


Figure 1: An overview of the FactMM-RAG system. It mainly contains three stages: (1) Leveraging RadGraph to characterize each radiology report and mine factually-informed report pairs; (2) Integrating factual knowledge into the training of the universal multimodal retriever; (3) Given the radiology image, employing the fact-aware multimodal retriever to search for factually-informed reference reports and augmenting the multimodal foundation model in generating accurate radiology reports.

large biomedical image-text collection towards a biomedical multimodal foundation model. Nevertheless, these existing medical multimodal retrievers neglect specific image information and do not adequately emphasize factual accuracy, resulting in imprecision when retrieving radiology reports.

Medical Multimodal Foundation Model. Significant efforts have been made in applying multimodal foundation models to the medical imaging domain [25, 31, 40, 43]. As chest X-ray radiology is the most commonly performed imaging examination, tailored medical multimodal foundation models for this critical area has gathered much attention [3, 5, 6, 42, 45]. Jain et al. [20] advances this area by designing a novel information extraction schema to structure radiology reports from chest radiographs; Delbrouck et al. [7], Miura et al. [30] take a step forward, using reinforcement learning from semantic rewards to improve the factual quality of generated radiology reports; Chen et al. [6] recently has also developed an instruction-tuned multimodal foundation model capable of sophisticated interpretation and analysis of chest X-rays.

One closely related line of work to ours is retrieval-based radiology report generation given only radiology images. For instance, Li et al. [24] proposes a retrieval policy module to update radiology reports via hierarchical reinforcement learning; Endo et al. [10] employs image-text embeddings from contrastive learning for retrieval-augmented radiology report generation; Ramesh et al. [36] proposes synthesizing additional reports and reducing hallucinations from reference report priors to improve report generation.

3 Methodology

In this section, we present the overall methodology of FactMM-RAG. We first detail the training procedure of our fact-aware medical multimodal retriever in Section 3.1. We then provide the pipeline for retrieval-augmented radiology report generation with our multimodal retriever in section 3.2. The overview is illustrated in Figure 1.

3.1 Fact-aware Multimodal Retrieval

This section discusses the training process of the multimodal retriever with factual knowledge. Each patient in the corpus has a chest X-ray radiology image along with its corresponding report. We begin by annotating each report using RadGraph [20], then constructing factual report pairs to train our multimodal retriever. We describe these steps as follows.

Chest Radiograph Annotation. Since radiology reports are free-text, we utilize the RadGraph information extraction tool to extract structured knowledge graphs from them. Specifically, RadGraph employs named entity recognition and relation extraction models to identify radiological entities (e.g. carina, lungs, abnormalities) and the clinical relations between them (e.g. modify, located at, suggestive of). Each radiology report is then segmented into distinct regions and stored as $[(\texttt{entity}_1, \texttt{entity label}_1, \texttt{relation}_1), (\texttt{entity}_2, \texttt{entity label}_2, \texttt{relation}_2), \ldots]$. After characterizing the chest radiograph for each report in the training corpus, we construct factual report pairs.

Factual Report Pairs Mining. Each report has an associated medical label describing the symptom. We first utilize the query report to search for other reports with the same symptom, aiming to eliminate false negatives when constructing report pairs. Rather than solely relying on the diagnostic labels, we further capture the factually-oriented pathology semantics between different reports. Following F1RadGraph [20], we calculate the factual similarity $s(q_{txt}, d_{txt})$ between query report q_{txt} and other reports d_{txt} in the annotated format as follows,

$$s(q_{txt}, d_{txt}) = \frac{2 \cdot (\hat{q}_{txt} \cap \hat{d}_{txt})}{\text{length}(\hat{q}_{txt}) + \text{length}(\hat{d}_{txt})}, \tag{1}$$

where $\hat{q}_{txt}, \hat{d}_{txt}$ denotes reports with only annotated entities and relations in RadGraph structured form. We then set a strict threshold δ to filter out searched reports with low similarity score:

$$N_{q_{txt}} = \{d_{txt} \in D | s(q_{txt}, d_{txt}) > \delta\}. \tag{2}$$

where $N_{q_{txt}}$ denotes factual positive report pairs for q_{txt} and D is the total training corpus. Since each query report is associated with a corresponding radiology image, these factual report pairs can also be applied to the query report's radiology image. Next, we train our multimodal retriever with mined factual report pairs.

Multimodal Dense Retrieval. Following previous work [56], we universally encode each query image q_{img} and other image-text pairs (d_{txt}, d_{img}) in the training corpus, using one encoder, MARVEL:

$$\mathbf{q} = \text{MARVEL}(q_{img}); \tag{3}$$

$$\mathbf{d} = \text{MARVEL}(d_{txt}, d_{img}), \tag{4}$$

where each image-text pair is represented as a single embedding. We then model the relevance score $f(q, d)$ between the query image and other image-text pairs by cosine similarity:

$$f(q, d) = \cos(\mathbf{q}, \mathbf{d}). \tag{5}$$

To inject factually-oriented medical knowledge into multimodal retrieval, we train the encoder to minimize the following loss,

$$\mathcal{L} = -\sum_{q_{img} \in D} \sum_{d^+ \in N_{q_{img}}} \log \frac{e^{f(\mathbf{q}, \mathbf{d}^+)/\tau}}{e^{f(\mathbf{q}, \mathbf{d}^+)/\tau} + \sum_{\mathbf{d}^-} e^{f(\mathbf{q}, \mathbf{d}^-)/\tau}}, \tag{6}$$

where d^+ are obtained through factual report pair mining and d^- are in-batch negative samples [22]. Then, we use our multimodal retriever and foundation model to perform retrieval-augmented radiology finding generation.

3.2 Retrieval Augmentation for Accurate Radiology Report Generation

Given our trained fact-aware multimodal retriever, we encode the query image and each report in the training corpus. Then, we retrieve the report with the highest relevance score to the query image as the factually-informed relevant report. Subsequently, we pass the query image along with the relevant report into a multimodal foundation model to perform retrieval-augmented generation training. The multimodal foundation model is finetuned by standard autogressive loss,

$$\mathcal{L} = -\frac{1}{n} \log \prod_{i}^{n} p_\theta(y_i | q_{img}, d^*_{txt}, x_{\texttt{instr}}, y_{<i}), \tag{7}$$

where q_{img} is the query image, d^*_{txt} is the retrieved factually-informed relevant report, $x_{\texttt{instr}}$ is the prompt instruction, and y is the ground-truth report. During inference, we retrieve a relevant report from the training corpus using an unseen patient X-ray image, and pass them into the multimodal foundation model to generate findings with higher factual accuracy.

4 Experimental Setup

Dataset. Following Delbrouck et al. [8], we use the processed MIMIC-CXR [21] to train both retriever and foundation model. This dataset contains 125,417 training radiology image-report pairs, 991 validation pairs, and 1,624 test pairs. They are sourced from the Beth Israel Deaconess Medical Center. CheXpert [17] is another chest X-ray dataset from Stanford Health Care. Since it contains complete finding reports only for a testing dataset containing 1000 pairs, we use it as zero-shot evaluation.

Evaluation Metrics. We evaluate our proposed system using both natural language generation and medically-tailored evaluation metrics. For language fluency measures, we use ROUGE-L [26] to evaluate the longest common subsequence overlap between the generated and reference findings, and BERTScore [53] to evaluate non-clinical semantic sentence similarity.

For clinical accuracy measures, we employ CheXbert [38] to generate the ground-truth diagnostic labels for finding reports. Following Delbrouck et al. [8], we then calculate the F1CheXbert [54], which is the F1-score for 5 observations (Cardiomegaly, Edema, Consolidation, Atelectasis, Pleural Effusion) by comparing the generated report with the reference report's classifications. Beyond using the limited diagnostic labels for evaluation, we also adopt F1RadGraph [20] to measure factual correctness by calculating the overlap in radiological entities and clinical relations between the generated report and the reference report. See Appendix A for more details.

Baselines. We mainly compare our retriever with other baselines under the multimodal RAG setting. We include the following baselines, CLIP [34] is a multimodal retriever pretrained from general-domain image-text pairs; GLoRIA [16] leverages attention-weighted image regions with contextual words to learn localized and global representations for radiology images and reports; MedCLIP [44] and CXR-CLIP [48] build on CLIP and utilize diagnostic labels as training signals for learning radiology image and text representations; BiomedCLIP [52] extends the radiology-specific dataset and pretrains on a larger magnitude of biomedical data to learn multimodal representations; Med-MARVEL utilizes universal encoder MARVEL [56] to conduct contrastive learning on each patient's self image-report pair without further training on factual image-report pairs.

We also compare our method with non-RAG approaches. "No Retriever" refers to directly fine-tuning the backbone to generate reports without retrieval augmentation; ORGan [14] first creates an observation plan, then feeds the plan and radiographs to generate the report through a tree reasoning mechanism. Upper-bound results using an oracle in training corpus with top-1 factual similarity to test query report are also presented.

Implementation Details. In our experiments, we use MARVEL [56] as our multimodal retriever backbone. MARVEL is a language model based on T5-ANCE [50], trained with modality-balanced hard negatives. We use LLaVA [27] as our multimodal foundation model backbone. Since each radiology study contains multiple image views for each patient, we select the frontal view. We also concatenate the finding and impression sections to form the X-ray report. To reduce training costs and address factual report pair imbalances, we rerank the retrieved reports by factual similarity and use the top 2 factual report pairs for each query to train our multimodal retriever. We leave more training details in Appendix A.

5 Evaluation Results

In this section, we present our experimental results. We first evaluate the overall performance between different retrievers under two settings in section 5.1. Next, we discuss the ablation studies in section 5.2. We then explore the fact-aware capability of our retriever in section 5.3 and section 5.4. Lastly, we show the superiority of our retriever through a case study in section 5.5.

5.1 Overall Performance

The results of our fact-aware RAG system are shown in Table 1. In MIMIC-CXR, FactMM-RAG outperforms state-of-the-art retrievers by a significant margin, up to 6.5% in F1CheXbert and 2%

Table 1: Overall performance of FactMM-RAG and baselines under the multimodal retrieval-augmentation setting. Models are evaluated by textual similarity and factual similarity between generated and reference reports. FactMM-RAG outperforms the best baseline with p-value < 0.05.

Model	MIMIC-CXR				CheXpert			
	Factual Similarity		Textual Similarity		Factual Similarity		Textual Similarity	
	F1CheXbert	F1RadGraph	ROUGE-L	BERTScore	F1CheXbert	F1RadGraph	ROUGE-L	BERTScore
No Retriever	0.496	0.234	0.294	0.549	0.371	0.173	0.231	0.469
ORGan [14]	0.541	0.240	0.308	0.552	0.431	0.181	0.232	0.470
CLIP [34]	0.507	0.241	0.300	0.552	0.381	0.172	0.231	0.468
GLoRIA [16]	0.476	0.232	0.294	0.543	0.397	0.173	0.231	0.468
MedCLIP [44]	0.517	0.238	0.298	0.549	0.408	0.182	0.238	0.471
CXR-CLIP [48]	0.501	0.243	0.302	0.553	0.406	0.183	0.241	0.471
BiomedCLIP [52]	0.502	0.233	0.293	0.546	0.380	0.173	0.232	0.469
Med-MARVEL [56]	0.537	0.237	0.306	0.549	0.454	0.185	0.243	0.472
FactMM-RAG	**0.602**	**0.257**	**0.307**	**0.561**	**0.475**	**0.185**	**0.236**	**0.475**
Oracle	0.972	0.523	0.495	0.677	0.951	0.384	0.350	0.548

Table 2: Ablation study of FactMM-RAG including multimodal retrieval and backbone variation.

Model	MIMIC-CXR				CheXpert			
	Factual Similarity		Textual Similarity		Factual Similarity		Textual Similarity	
	F1CheXbert	F1RadGraph	ROUGE-L	BERTScore	F1CheXbert	F1RadGraph	ROUGE-L	BERTScore
Setting: Multimodal Retrieval								
CLIP [34]	0.341	0.160	0.238	0.489	0.285	0.130	0.207	0.439
GLoRIA [16]	0.346	0.137	0.211	0.453	0.359	0.135	0.216	0.447
MedCLIP [44]	0.539	0.198	0.261	0.508	0.478	0.161	0.225	0.454
CXR-CLIP [48]	0.516	0.215	0.277	0.524	0.444	0.167	0.230	0.458
BiomedCLIP [52]	0.502	0.233	0.293	0.546	0.386	0.142	0.216	0.441
Med-MARVEL [56]	0.550	0.212	0.279	0.525	0.479	0.160	0.222	0.454
FactMM-RAG	**0.605**	**0.249**	**0.297**	**0.547**	**0.491**	**0.174**	**0.237**	**0.467**
Oracle	0.992	0.429	0.399	0.612	0.999	0.438	0.362	0.554
Setting: Multimodal Retrieval Augmented Generation								
ClueWeb-LLaVA$_{1.5}$	**0.602**	0.257	0.307	0.561	**0.495**	0.180	**0.239**	0.473
WebQA-LLaVA$_{1.5}$	0.572	**0.262**	0.304	0.562	0.456	0.184	0.237	**0.474**
Med-MARVEL-LLaVA$_{1.5}$	0.581	0.260	**0.311**	**0.563**	0.475	**0.185**	0.236	**0.474**
ClueWeb-LLaVA$_{1.6}$	0.601	0.252	0.303	0.558	0.492	0.178	0.237	0.471

in F1RadGraph. In CheXpert zero-shot evaluation, FactMM-RAG outperforms state-of-the-art retrievers by 2% and 1.2% in these two metrics, indicating our retriever's generalization capability compared to other models.

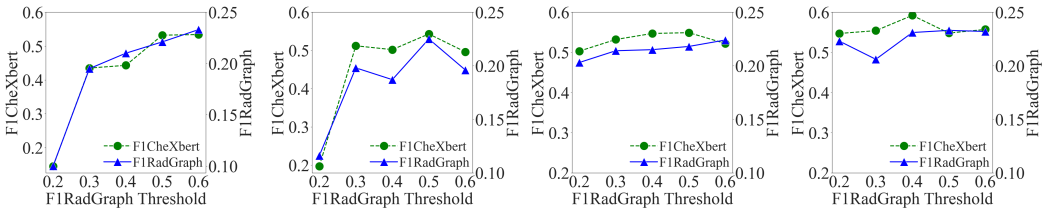
To establish the effectiveness of our RAG approach, we also show that FactMM-RAG significantly outperforms the fine-tuned backbone without retrieval augmentation by 10% and achieves competitive improvement over the SOTA non-RAG ORGan baseline.

Besides, we can observe that adopting the baseline retrievers on top of multimodal foundation models only yields marginal gains compared to the finetuning of foundation model generation without retrieval-augmentation. This shows that reports retrieved by baseline retrievers are factually-inferior to those from our retriever, potentially passing misleading information that prevents the foundation model from generating factual reports.

Specifically, compared to the retriever Med-MARVEL, we also observe factual-correctness performance gain based on two clinical metrics. Both use the same universal encoder backbone, but FactMM-RAG benefits from the injected factual medical knowledge, allowing it to search for the most similar and factually correct reports, thereby assisting the multimodal foundation model in generating more accurate reports.

5.2 Ablation Study

Multimodal Retrieval. Instead of relying on the multimodal foundation model to generate reports, we also evaluate the performance of the multimodal retrievers by directly encoding radiology images from the testing corpus and searching for the closest report from the training corpus for comparison with ground-truth reports. Table 2 shows that our retriever also achieves the best factual retrieval



(a) Threshold: 0 (b) Threshold: 0.4 (c) Threshold: 0.8 (d) Threshold: 1.0

Figure 2: Factual performance of FactMM-RAG controlled by different F1CheXbert and F1RadGraph thresholds. We vary the F1RadGraph thresholds under one fixed F1CheXbert threshold selected from {0, 0.4, 0.6, 0.8}.

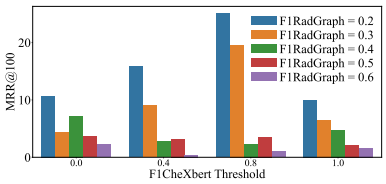
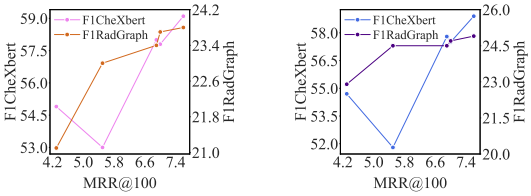


Figure 3: Retrieval evaluation of FactMM-RAG with different F1CheXbert and F1RadGraph thresholds. MRR calculates mean reciprocal rank at which the first relevant report that meets two factual similarity thresholds with query report is retrieved.



(a) Multimodal Retrieval (b) Multimodal RAG

Figure 4: Analysis of fact-aware capability propagation. The x-axis MRR measures the retriever's performance on retrieving factually relevant reports.

performance compared to other baselines under this setting across two datasets. This demonstrates that training the multimodal retriever with mined factually-informed report pairs can enhance its radiology image understanding capabilities and directly align it with precise reports.

Backbone Variation. We also investigate the impact of different retriever and foundation model backbones on radiology report generation in Table 2. We initialize our retriever model from two checkpoints: WebQA and ClueWeb in [56]. We observe that the ClueWeb checkpoint provides a marginal gain compared to the WebQA checkpoint. This can be attributed to the larger scale of the ClueWeb dataset used for pretraining. We also utilize Med-MARVEL as our retriever backbone, which exhibits similar performance to other backbones after training. This implies that even if our retriever is initialized with a backbone from a general domain, our factually-informed training strategy enables it to fully leverage medical knowledge and quickly adapt to the radiology-specific domain without degrading performance.

5.3 Fact-aware Capability Control

The factual similarity threshold in Equation 1 plays a critical role in controlling the fact-awareness of our multimodal retriever. We examine the performance of FactMM-RAG under different thresholds, as shown in Figure 2. Not only utilizing F1RadGraph thresholds, we also employ F1CheXbert to curate additional thresholds from the report's diagnostic labels to mine report pairs.

Under the same F1CheXbert threshold for mining report pairs, we observe that an increase in the F1RadGraph threshold correlates with an improvement in factual performance. However, adopting stricter thresholds for identifying report pairs does not yield further improvements and reaches saturation. After calculating the average number of report pairs per query, we find that high thresholds can exclude many relevant report pairs, as shown in Figure 3. This exclusion results in the potential loss of factually useful pairs, thereby hindering the training of our multimodal retriever driven by additional factual medical knowledge.

Rather than relying on diagnostic labels from CheXbert to identify high-quality report pairs, Figure 2a demonstrates that the F1RadGraph threshold alone can also effectively mine factual report pairs

Table 3: Case study on generated reports from MIMIC-CXR. Cyan text indicates radiological consistency with the ground-truth report. Orange text highlights extra accurate details provided by FactMM-RAG compared to Med-MARVEL. Red text denotes observations missing in Med-MARVEL.

Radiology Image	Med-Marvel	FactMM-RAG	Reference
	Single portable view of the chest. *There are bilateral pleural effusions,* moderate on the left and small on the right. *There is also pulmonary vascular redistribution and hazy alveolar infiltrate. cardiac silhouette is enlarged* but unchanged. Median sternotomy wires and mediastinal clips are again noted.	A left-sided pacemaker is in place with leads terminating in the right atrium and right ventricle. *The patient is status post median sternotomy and CABG.* The heart is moderately enlarged. *There is mild pulmonary edema. A small left pleural effusion is present.* ***There is atelectasis at the left lung base.*** *No pneumothorax is seen.*	*The patient is status post median sternotomy and CABG.* Left-sided dual-chamber pacemaker is noted with leads terminating in right atrium and right ventricle, unchanged. Cardiomegaly is similar. *There is continued mild to moderate pulmonary edema*, slightly improved compared to the prior exam. *Small layering bilateral pleural effusions* also may be slightly decreased in the interval. ***Bibasilar airspace opacities likely reflect atelectasis.*** *There is no pneumothorax.* No acute osseous abnormalities are visualized.
F1RadGraph	0.218	0.413	
CheXbert Observations	Cardiomegaly, Edema, Pleural Effusion	Cardiomegaly, Edema, Atelectasis, Pleural Effusion	Cardiomegaly, Edema, Atelectasis, Pleural Effusion
	The heart is mildly enlarged. The aorta is mildly tortuous. The mediastinal and hilar contours appear unchanged. *There is no pleural effusion or pneumothorax.* Streaky left basilar opacity suggests minor atelectasis. *There is no definite pleural effusion or pneumothorax.* The bones appear demineralized. There is mild-to-moderate rightward convex curvature centered along the mid thoracic spine.	*Heart size is mildly enlarged. The aorta is tortuous.* Mediastinal and hilar contours are otherwise unremarkable. *Pulmonary vasculature is normal.* Linear opacities in the left lower lobe are compatible with subsegmental atelectasis. *No focal consolidation, pleural effusion or pneumothorax is present. There are no acute osseous abnormalities.*	*Moderate enlargement of the cardiac silhouette* with a left ventricular predominance is unchanged. *The aorta remains tortuous*, and the hilar contours are stable. *Pulmonary vascularity is not engorged.* There is minimal atelectasis within the lung bases, but *no focal consolidation is present. No pleural effusion or pneumothorax is identified. There are no acute osseous abnormalities.*
F1RadGraph	0.333	0.526	
CheXbert Observations	Cardiomegaly, Atelectasis	Cardiomegaly, Atelectasis	Cardiomegaly, Atelectasis

for training our multimodal retriever. As the F1RadGraph threshold increases, FactMM-RAG even matches the performance under high threshold settings in Figure 2d. This signifies that employing our training strategy with curated factual query-report pairs still imposes useful supervision signals without relying on explicit diagnostic label guidance.

5.4 Fact-aware Capability Propagation

To further understand the benefits of our retriever for the foundation model, we explore the effective propagation of fact-aware capabilities from the retriever to the foundation model. To demonstrate this behavior, we use the mined factual report pairs as reference reports for the query report. We then use the retrieval metric Mean Reciprocal Rank (MRR) as an intermediate evaluation, shown in Figure 4. From the plot, we observe that as training progresses, the retrieval metric increases alongside two clinical metrics. This factually-oriented upward trend in our retriever's performance in Figure 4a is also reflected in the foundation model's performance in Figure 4b. This indicates that employing a factually-informed reference report selection strategy to train our multimodal retriever can also enhance the foundation model's ability to generate factually accurate radiology reports.

5.5 Case Study

In this section, we present two examples from MIMIC-CXR to qualitatively analyze our retriever's fact-aware capability, as illustrated in Table 3. In the first example, we observe that FactMM-RAG provides symptom observations consistent with the ground-truth report and generates more accurate factual details compared to Med-MARVEL, e.g., "post median sternotomy, atelectasis, not pneumothorax"; In the second example, we further observe that although both retrievers generate reports with diagnostic labels matching the ground-truth report, FactMM-RAG provides additional details compared to Med-MARVEL, such as "pulmonary vasculature is normal, no acute osseous abnormalities". These characteristics confirm that adopting our fact-aware retriever can assist multimodal foundation models in generating more accurate radiology reports. We also show the retrieved reports from two samples in the Appendix A.

6 Conclusion

In this paper, we aim at improving radiology report generation by introducing a fact-informed medical multimodal retriever for retrieval-augmented generation. In particular, we utilize RadGraph to annotate chest radiograph reports and mine clinically-relevant pairs. We integrate factual information into a universal multimodal retriever, presenting FactMM-RAG, a fact-aware multimodal retrieval-augmented radiology report generation pipeline. FactMM-RAG outperforms all state-of-the-art retrievers evaluated by factual correctness and textual coherence for final report generation in MIMIC-CXR and CheXpert datasets. We further confirm the benefit of our multimodal retriever from the analysis of fact-aware capability control and propagation. Given the pervasive applications of machine learning in clinical diagnoses using chest X-rays, we hope our factual-informed approach inspires further work in multimodal generative artificial intelligence in healthcare contexts.

7 Limitations

Despite the strong performance of our FactMM-RAG pipeline, we acknowledge potential limitations of our proposed method. In particular, our work only emphasizes chest radiology domains. It is also worth exploring our retrieval-augmented factual report generation pipeline in broader medical domains, such as brain scan or histology datasets.

Another concern lies in the chosen evaluation metrics, F1RadGraph and F1CheXbert. F1CheXbert reflects high-level observational accuracy, while F1RadGraph assesses the correctness of radiology entities and clinical relationships. However, other radiologically-specific metrics, such as report conciseness and clarity, should also be considered [41, 49]. Ideally, we should incorporate methods of evaluation directly aligned with human evaluations or involve domain expertise itself in our pair-mining and final evaluation procedure. Moreover, it is worth exploring a long-tail evaluation by leveraging more fine-grained ground-truth label annotations [13].

8 Societal Impact

Automated radiology report generation offers significant societal benefits, including improved diagnostic accuracy, increased efficiency, and enhanced access to care in underserved areas. These systems can help radiologists focus on complex cases, reduce human error, and provide valuable insights into public health trends, ultimately leading to better patient outcomes and more informed healthcare policies.

However, there are also potential downsides. Overreliance on AI could diminish critical thinking among medical professionals, and biased algorithms may exacerbate healthcare inequalities. Additionally, concerns about data privacy, job displacement, and legal accountability must be addressed to ensure that the benefits of these systems are realized without compromising patient trust or care quality.

References

[1] Muhammad Aurangzeb Ahmad, Ilker Yaramis, and Taposh Dutta Roy. Creating trustworthy llms: Dealing with hallucinations in healthcare ai, 2023.

[2] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. Improving language models by retrieving from trillions of tokens, 2022.

[3] Pierre Chambon, Christian Bluethgen, Jean-Benoit Delbrouck, Rogier Van der Sluijs, Małgorzata Połacin, Juan Manuel Zambrano Chaves, Tanishq Mathew Abraham, Shivanshu Purohit, Curtis P. Langlotz, and Akshay Chaudhari. Roentgen: Vision-language foundation model for chest x-ray generation, 2022.

[4] Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W. Cohen. Murag: Multimodal retrieval-augmented generator for open question answering over images and text, 2022.

[5] Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Cross-modal memory networks for radiology report generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5904–5914, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.459. URL https://aclanthology.org/2021.acl-long.459.

[6] 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, 2024.

[7] Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. Improving the Factual Correctness of Radiology Report Generation with Semantic Rewards. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4348–4360, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.319. URL https://aclanthology.org/2022.findings-emnlp.319.

[8] Jean-Benoit Delbrouck, Maya Varma, Pierre Chambon, and Curtis Langlotz. Overview of the RadSum23 Shared Task on Multi-modal and Multi-anatomical Radiology Report Summarization. In Dina Demner-fushman, Sophia Ananiadou, and Kevin Cohen, editors, *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 478–482, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.bionlp-1.45. URL https://aclanthology.org/2023.bionlp-1.45.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[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 Subhrajit Roy, Stephen Pfohl, Emma Rocheteau, Girmaw Abebe Tadesse, Luis Oala, Fabian Falck, Yuyin Zhou, Liyue Shen, Ghada Zamzmi, Purity Mugambi, Ayah Zirikly, Matthew B. A. McDermott, and Emily Alsentzer, editors, *Proceedings of Machine Learning for Health*, volume 158 of *Proceedings of Machine Learning Research*, pages 209–219. PMLR, 04 Dec 2021. URL https://proceedings.mlr.press/v158/endo21a.html.

[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, 2024.

[12] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: Retrieval-augmented language model pre-training, 2020.

[13] Gregory Holste, Song Wang, Ajay Jaiswal, Yuzhe Yang, Mingquan Lin, Yifan Peng, and Atlas Wang. CXR-LT: Multi-Label Long-Tailed Classification on Chest X-Rays (version 1.1.0), 2023. URL https://doi.org/10.13026/c4tr-kr83.

[14] Wenjun Hou, Kaishuai Xu, Yi Cheng, Wenjie Li, and Jiang Liu. Organ: Observation-guided radiology report generation via tree reasoning, 2023. URL https://arxiv.org/abs/2306.06466.

[15] Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A. Ross, and Alireza Fathi. Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory, 2023.

[16] Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. GLoRIA: A Multimodal Global-Local Representation Learning Framework for Label-Efficient Medical Image Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3942–3951, 2021.

[17] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison, 2019.

[18] Lucky Iyeke, Rebecca Moss, Rachel Hall, Jun Wang, Lovleen Sandhu, Benjamin Appold, Ella Kalontar, Dimitra Menoudakos, Mahesh Ramnarine, Samuel P. LaVine, Seungwoo Ahn, and Michelle Richman. Reducing unnecessary 'admission' chest x-rays: An initiative to minimize low-value care. *Cureus*, 14(10):e29817, Oct 2022. doi: 10.7759/cureus.29817.

[19] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning with retrieval augmented language models, 2022.

[20] Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. RadGraph: Extracting Clinical Entities and Relations from Radiology Reports, 2021.

[21] Alistair E. W. 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), December 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0322-0. URL `http://dx.doi.org/10.1038/s41597-019-0322-0`.

[22] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. Dense passage retrieval for open-domain question answering, 2020.

[23] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, 2021.

[24] Christy Y. Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. Hybrid retrieval-generation reinforced agent for medical image report generation, 2018.

[25] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*, 2023.

[26] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*, 2004. URL `https://api.semanticscholar.org/CorpusID:964287`.

[27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning, 2023.

[28] Kang Liu, Zhuoqi Ma, Mengmeng Liu, Zhicheng Jiao, Xiaolu Kang, Qiguang Miao, and Kun Xie. Factual serialization enhancement: A key innovation for chest x-ray report generation, 2024.

[29] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

[30] Yasuhide Miura, Yuhao Zhang, Emily Bao Tsai, Curtis P. Langlotz, and Dan Jurafsky. Improving factual completeness and consistency of image-to-text radiology report generation, 2021.

[31] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In Stefan Hegselmann, Antonio Parziale, Divya Shanmugam, Shengpu Tang, Mercy Nyamewaa Asiedu, Serina Chang, Tom Hartvigsen, and Harvineet Singh, editors, *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pages 353–367. PMLR, 10 Dec 2023. URL `https://proceedings.mlr.press/v225/moor23a.html`.

[32] Ankit Pal and Malaikannan Sankarasubbu. Gemini goes to med school: Exploring the capabilities of multimodal large language models on medical challenge problems & hallucinations, 2024.

[33] Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Med-halt: Medical domain hallucination test for large language models, 2023.

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, 2021.

[35] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023.

[36] Vignav Ramesh, Nathan Andrew Chi, and Pranav Rajpurkar. Improving radiology report generation systems by removing hallucinated references to non-existent priors, 2022.

[37] Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. Replug: Retrieval-augmented black-box language models, 2023.

[38] 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, 2020.

[39] Annemiek M. Speets, Yolanda van der Graaf, Arno W. Hoes, Sjouke Kalmijn, Arnold P. Sachs, Matthijs J. Rutten, Jan W. Gratama, Annemiek D. Montauban van Swijndregt, and Willem P. Mali. Chest radiography in general practice: Indications, diagnostic yield, and consequences for patient management. *The British Journal of General Practice: The Journal of the Royal College of General Practitioners*, 56(529):574–578, 2006.

[40] Liwen Sun, Abhineet Agarwal, Aaron Kornblith, Bin Yu, and Chenyan Xiong. Ed-copilot: Reduce emergency department wait time with language model diagnostic assistance, 2024.

[41] Binit Sureka et al. Seven c's of effective radiology reporting. *The Journal of National Accreditation Board for Hospitals & Healthcare Providers*, 1(1):17, January-June 2014. URL `https://link.gale.com/apps/doc/A416342585/AONE?u=anon~8d1a1cee&sid=googleScholar&xid=2b3f1ebd`. Accessed 14 June 2024.

[42] Omkar Thawkar, Abdelrahman Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Shahbaz Khan. Xraygpt: Chest radiographs summarization using large medical vision-language models. *arXiv: 2306.07971*, 2023.

[43] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, Basil Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David Fleet, Philip Mansfield, Sushant Prakash, Renee Wong, Sunny Virmani, Christopher Semturs, S Sara Mahdavi, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Karan Singhal, Pete Florence, Alan Karthikesalingam, and Vivek Natarajan. Towards generalist biomedical ai, 2023.

[44] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text, 2022.

[45] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data, 2023.

[46] Qianqian Xie, Jiayu Zhou, Yifan Peng, and Fei Wang. Factreranker: Fact-guided reranker for faithful radiology report summarization, 2023.

[47] Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. Retrieval-augmented multimodal language modeling, 2023.

[48] Kihyun You, Jawook Gu, Jiyeon Ham, Beomhee Park, Jiho Kim, Eun K. Hong, Woonhyuk Baek, and Byungseok Roh. Cxr-clip: Toward large scale chest x-ray language-image pre-training. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 101–111. Springer Nature Switzerland, 2023. doi: 10.1007/978-3-031-43895-0_10. URL https://doi.org/10.1007%2F978-3-031-43895-0_10.

[49] 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, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. Evaluating progress in automatic chest x-ray radiology report generation. *medRxiv*, 2022. doi: 10.1101/2022.08.30.22279318. URL https://www.medrxiv.org/content/early/2022/08/31/2022.08.30.22279318.

[50] Shi Yu, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. Openmatch-v2: An all-in-one multi-modality plm-based information retrieval toolkit. pages 3160–3164, 07 2023. doi: 10.1145/3539618.3591813.

[51] Zichun Yu, Chenyan Xiong, Shi Yu, and Zhiyuan Liu. Augmentation-adapted retriever improves generalization of language models as generic plug-in, 2023.

[52] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs, 2024.

[53] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020.

[54] Yuhao Zhang, Derek Merck, Emily Bao Tsai, Christopher D. Manning, and Curtis P. Langlotz. Optimizing the factual correctness of a summary: A study of summarizing radiology reports, 2020.

[55] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. Contrastive Learning of Medical Visual Representations from Paired Images and Text, 2022.

[56] Tianshuo Zhou, Sen Mei, Xinze Li, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, Yu Gu, and Ge Yu. MARVEL: Unlocking the Multi-Modal Capability of Dense Retrieval via Visual Module Plugin, 2024.

[57] Erdi Çallı, Ecem Sogancioglu, Bram van Ginneken, Kicky G. van Leeuwen, and Keelin Murphy. Deep learning for chest x-ray analysis: A survey. *Medical Image Analysis*, 72: 102125, August 2021. ISSN 1361-8415. doi: 10.1016/j.media.2021.102125. URL http://dx.doi.org/10.1016/j.media.2021.102125.

A Appendix

A.1 Retriever Training Procedure

To training our fact-aware multimodal retriever, we not only use mined factual report pairs as positive reports to the query image, but also incorporate the query image's corresponding report. Following [50, 56], we also adopt modality-balanced hard negatives to train the retriever after in-batch negative

Visual Question Answering:

> Generate a radiology report from this image: `<image>`

Retrieval Augmented Generation:

> Here is a report of a related patient: "`<document>`"
> Generate a radiology report from this image: `<image>`

Figure 5: Prompt templates for Visual Question Answering and Retrieval Augmented Generation

Table 4: Case study on retrieved reports from MIMIC-CXR. Cyan text indicates radiological consistency with the ground-truth report. Orange text highlights extra accurate details provided by FactMM-RAG compared to Med-MARVEL. Red text denotes observations missing in Med-MARVEL.

Radiology Image	Med-Marvel	FactMM-RAG	Reference
	A single portable chest radiograph was obtained. *Bilateral pleural effusions and mild atelectasis have increased.* Cardiomegaly is unchanged. There is no consolidation or pneumothorax. Pacing leads, sternotomy wires, vascular clips, and abdominal surgical clips are unchanged.	No focal consolidation is identified. There is unchanged appearance of opacifications in the left lung base, *likely due to a combination of atelectasis and pleural effusion. There is a small right pleural effusion. Mild pulmonary edema persists.* The heart is moderately enlarged, but stable. Left sided pacemaker is seen with transvenous leads in the right atrium, right ventricle, and left ventricle.	The patient is status post median sternotomy and CABG. Left-sided dual-chamber pacemaker is noted with leads terminating in right atrium and right ventricle, unchanged. Cardiomegaly is similar. *There is continued mild to moderate pulmonary edema,* slightly improved compared to the prior exam. *Small layering bilateral pleural effusions also may be slightly decreased in the interval. Bibasilar airspace opacities likely reflect atelectasis.*There is no pneumothorax. No acute osseous abnormalities are visualized.
F1RadGraph	0.274	0.345	
CheXbert Observations	Cardiomegaly, Atelectasis, Pleural Effusion	Cardiomegaly, Edema, Atelectasis, Pleural Effusion	Cardiomegaly, Edema, Atelectasis, Pleural Effusion
	The heart is mildly enlarged with a left ventricular configuration. There is mild-to-moderate unfolding of the thoracic aorta. The arch is partly calcified. The mediastinal and hilar contours appear unchanged. *There are streaky left basilar opacities suggesting minor atelectasis.* A small eventration is noted along the anterior right hemidiaphragm. There is an air-fluid level in the stomach. Air-fluid levels are seen in the epigastric region. There is no evidence for free air. Cholecystectomy clips project over the right upper quadrant. Moderate degenerative changes are similar along the mid thoracic spine.	*Moderate enlargement of the cardiac silhouette is unchanged. The aorta remains tortuous.* The mediastinal and *hilar contours are normal.* Pulmonary vasculature is normal. *Streaky atelectasis is noted* in the left lower lobe. The right lung is clear. *No focal consolidation, pleural effusion or pneumothorax is present.* Multiple clips are noted within the left upper abdomen.	*Moderate enlargement of the cardiac silhouette* with a left ventricular predominance is unchanged. *The aorta remains tortuous, and the hilar contours are stable.* Pulmonary vascularity is not engorged. There is *minimal atelectasis within the lung bases*, but *no focal consolidation is present. No pleural effusion or pneumothorax is identified.* There are no acute osseous abnormalities.
F1RadGraph	0.197	0.621	
CheXbert Observations	Cardiomegaly, Atelectasis	Cardiomegaly, Atelectasis	Cardiomegaly, Atelectasis

training from the multimodal dense retrieval stage. We use AdamW [29] as our optimizer and training epochs = 15, early stopping epoch = 5, batch size = 32, learning rate = 5e-6, and the temperature hyperparameter τ = 0.01. For our MARVEL backbone, we use T5-ANCE [50] as the text encoder and vision transformer [9] as the vision encoder. Models are trained using 1 NVIDIA RTX A6000 for 10 hours.

A.2 RAG Finetuning Procedure

To create a RAG dataset for fine-tuning LLaVA, we search the nearest-neighbor document d^*_{txt} for a query image q_{img} using a retriever's embeddings. We filter out any results that involve retrieving a patient's own report, the same patient's other studies, or malformed reports in the training dataset (specified by being less than 5 characters). We apply the prompt templates in Figure 5, and fine-tune LLaVA-1.5 for one epoch. Models are trained using 8x NVIDIA RTX A6000 for 4 hours, with epochs=1, learning rate=2e-5, global batch size=128, from vicuna-7b-v1.5 checkpoint. We save the checkpoint after one full pass of the training dataset for final evaluation.

A.3 Evaluation Details

Here, we provide implementation details regarding the evaluation methodology.

F1-RadGraph. For F1-RadGraph score computation, we follow previous work (MIMIC-CXR-RRS) [2] in employing RG_{ER} as F1-Radgraph score computation on an instance level. Using the `radgraph` library implementation, this equates to utilizing `reward_level="partial"`.

F1-CheXbert. F1-CheXbert score computation consists of the micro-averaged F1-score between 5 selected classes from the CheXbert labeler. Naturally, F1-CheXbert scores are only computable over entire datasets. For instance-level CheXbert scores (used for pair mining), we employ the proportion of equivalent predicted classes between a reference and predicted text sample. These instance-level F1-CheXbert scores can be computed using `np.sum(ref == hyp) / 5`, and take on values $\in \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$.

CheXpert Hidden Test Set. We use the 1000 hidden test reports from MIMIC-CXR-RRS and download the CheXpert images from Stanford AIMI Shared Datasets [3].

Oracle Retrieval. Oracle Retrieval is performed via ground-truth access to a reference document's generated report. For training queries, this is always known, and an oracle retriever would obtain documents as $Oracle(q_i) \doteq \underset{j \in \text{corpus}, j \neq i}{\arg\max}\ s(q_i, d_j)$, where $s(q, d)$ is the sum of the F1-RadGraph and F1-CheXbert instance-wise scores. In practice, this results in retrieving samples with F1-CheXbert=1.0 and the largest F1-RadGraph score within the partition. Test-time retrieval performs the same operation, without the restriction of $j \neq i$ as self-retrieval is not possible due to the corpus being the training dataset.

Oracle RAG. Oracle-LLaVA is obtained by fine-tuning LLaVA under identical conditions, utilizing Oracle Retrieval for retrieving documents in the training and test set.

A.4 Case study on Retrieved Reports

We now conduct case study on the retrieved reports shown in Table 4. We show that our FactMM-RAG captures most of the factual details in retrieved reports compared to the ground-truth reports. Thus, the factual correctness of our retriever can be propagated to the multimodal foundation models effectively. However, the reports retrieved from Med-MARVEL contain erroneous information, which negatively impacts the report generation by multimodal foundation models.

[2] https://vilmedic.app/papers/acl2023/

[3] https://stanfordaimi.azurewebsites.net/datasets/8cbd9ed4-2eb9-4565-affc-111cf4f7ebe2

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction accurately reflect the paper's contributions and scope

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have created a separate "Limitations" section in our paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: We do not include the theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We describe our novel architecture fully and report training details such as hyperparameters and prompts in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Since MIMIC dataset is highly sensitive and may contain ethics issues related to patient privacy, we plan to release our code upon paper acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We have discussed the experimental details in main body and appendix

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report our major results with statistical significance.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer:[Yes]

   Justification: We report it in the appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We conduct our research in conformity with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discussed in section 8.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer:[NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have cited the original paper that produced the code package or dataset.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Our model/code is documented well and we plan to release the codebase upon acceptance.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing. We passed the MIMIC data usage training related to human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing. We passed the MIMIC data usage training related to human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.