

# Medvision: Multimodal Retrieval-Augmented Generation System for Evidence based Medical Learning

Saketh Reddy Kommareddy

Revanth Mudavath

Sai Deepak Vempati

Oregon State University

{kommares, mudavatr, vempats}@oregonstate.edu

## Abstract

*This project presents a multimodal retrieval-augmented generation (MMed-RAG) system for automated radiology report generation using a subset of the IU-Xray dataset. Our architecture combines CLIP for vision-language embedding, FAISS for similar image retrieval, and FLAN-T5-xl as a decoder to generate radiology reports. The model was evaluated on the last 200 held-out test images using Top-k retrieval ( $k = 5$ ), achieving a BLEU score of 23.45, ROUGE-L of 20.91, and BERTScore F1 of 86.42. While these results are promising, they fall slightly below state-of-the-art systems reported in the literature—a likely consequence of training on only 1,000 images. Nevertheless, the results highlight the viability of integrating open-source encoders and retrievers for clinically relevant report generation in data-constrained settings. This work helps bridge the gap between advanced medical AI and real-world clinical deployment by prioritizing both performance and interpretability.*

## 1. Introduction

The field of medical artificial intelligence has seen rapid progress in leveraging deep learning to interpret radiological images and generate diagnostic reports. Despite impressive advancements, most existing systems are either purely image-based or purely text-based, lacking the ability to fully utilize the rich interplay between modalities such as visual scans and clinical narratives. This limitation has sparked interest in Multimodal Retrieval-Augmented Generation (RAG) frameworks, which aim to bridge this gap by combining vision and language inputs to generate more contextually accurate outputs [1].

Multimodal RAG refers to a class of systems that augment generative models by retrieving information from multiple modalities, such as images, textual records, or structured data, before producing an output. Unlike conventional large language models (LLMs), which rely solely on

text inputs and internal parameters for generation, Vision-Language Models (VLMs) are designed to process both visual and textual information [2]. By incorporating retrieval mechanisms, these systems are better equipped to ground their responses in external knowledge, thereby improving factual accuracy, relevance, and interpretability, especially in high-stakes fields like healthcare. In the context of radiology, accurate interpretation often requires synthesizing visual patterns from X-rays with historical and contextual data from previous reports. Recent works such as MedCLIP-Retireval [3], BioViL-T [4], and MMed-RAG [7] have demonstrated the potential of combining vision-language models with retrieval modules to improve performance on medical report generation, visual question answering (VQA), and captioning tasks.

However, these approaches often depend on large-scale pretraining with tens of thousands of annotated samples and intensive compute resources. This makes them difficult to reproduce or deploy in real-world, resource-constrained environments. In contrast, we explore a more efficient and interpretable alternative by proposing a compact yet effective Multimodal RAG system, trained on only 1,000 samples from the IU-Xray dataset [6].

Our system is designed not only to generate radiology reports but also to enhance factual reliability by explicitly grounding each output in retrieved, trusted sources. To this end, we present both the generated report and its retrieval context, a ranked set of semantically similar historical reports, which serves as a confidence signal and aids in clinical interpretability.

The architecture integrates:

1. A **CLIP-based encoder** to generate joint image-text embeddings;
2. A **FAISS-based retriever** to locate top-k similar prior reports;
3. A **FLAN-T5 decoder** to produce clinically informed summaries based on the input image and retrieval context.

Through this approach, we demonstrate that retrieval-augmented multimodal systems can remain robust and explainable, even when trained under limited supervision, paving the way for scalable and trustworthy deployment in medical AI settings.

## 2. Related Work

We review recent progress in medical multimodal AI, focusing on vision-language models designed for radiology report generation and retrieval-augmented generation. Our work builds upon these foundations while aiming for greater interpretability and lower data requirements.

Recent advancements in medical AI have seen a surge of interest in combining visual and textual modalities to improve diagnostic understanding and report generation. Traditional approaches, such as those relying solely on CNN-based image encoders or sequence-to-sequence models for text generation, fall short in effectively capturing the interdependence between radiological scans and their clinical interpretations.

To address this limitation, several multimodal vision-language models have been proposed. MedCLIP [3] leverages contrastive learning between unpaired medical images and texts to build robust joint representations, which has proven effective for downstream tasks like retrieval and classification. Similarly, BioViL-T [4] employs dual-encoder architectures pre-trained on large medical corpora to enhance performance on both classification and captioning benchmarks. These models demonstrate strong capabilities but rely heavily on large-scale pretraining (upwards of 80K–200K samples) and specialized medical datasets, making them resource-intensive and difficult to reproduce in real-world settings.

More recently, Flamingo-Med [2] introduced a large-scale visual language model tailored for few-shot learning. While highly performant (BLEU 35.8, BERTScore F1 92.5), it depends on proprietary architectures and massive general-domain pretraining, limiting its accessibility and interpretability for clinical users.

MMed-RAG[7] is a recently proposed modular retrieval-augmented generation framework designed for explainable multimodal inference across diverse medical domains, including radiology, pathology, dermatology, and ophthalmology. It integrates BLIP-2 for visual encoding, FAISS for semantic retrieval, and LLaVA for language generation, aiming to improve factual grounding and interpretability in multimodal clinical tasks.

In contrast, our work adopts the core architectural principles of MMed-RAG—namely retrieval-guided generation using image and text inputs—but focuses specifically on radiology report generation using a more lightweight, resource-efficient setup. We reimplement the pipeline using open-source components such as CLIP, FAISS, and FLAN-

T5-xl, and train on a small-scale subset of the IU-Xray dataset. Our variant emphasizes data efficiency and interpretability, making it suitable for low-resource deployment without sacrificing semantic fidelity.

## 3. Methodology

### 3.1. Overview

Our architecture is adapted from the original MMed-RAG[7] framework. Unlike the original system which spans multiple medical domains, we focus specifically on radiology and reimplement the pipeline using lightweight, open-source components tailored for low-resource settings.

We follow a three-stage pipeline designed to improve factual accuracy and grounding in radiology report generation and visual question answering tasks. Given a chest X-ray image, we first extract a joint image-text embedding using a vision-language encoder based on CLIP. This embedding is used to retrieve semantically similar past cases using FAISS, a high-performance similarity search library.

The retrieved reports, along with a caption generated from the input image, are formatted into a prompt that is fed into a pre-trained FLAN-T5-xl language model. The model then generates a full radiology report or answers a clinical question, depending on the task.

Unlike traditional models that rely solely on internal representations, our system explicitly integrates external, domain-specific evidence into the generation process. This retrieval-augmented approach improves factual alignment and offers interpretability by surfacing the retrieved context alongside the output. While we leverage existing tools like CLIP, FAISS, and FLAN-T5-xl, our contribution lies in orchestrating them into a modular, interpretable pipeline tailored for low-resource radiology generation tasks.

### 3.2. System Architecture

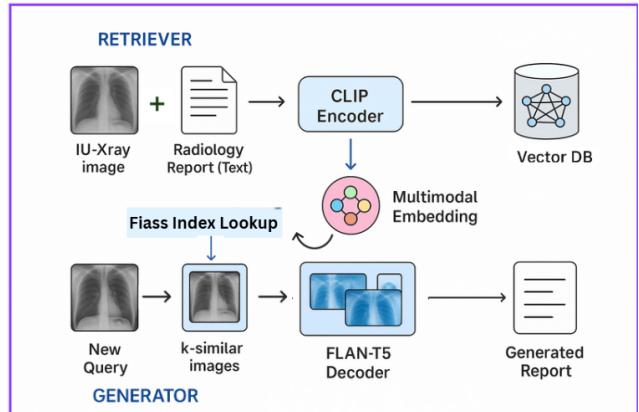


Figure 1. System architecture of MMedRAG

Our proposed system, MMed-RAG, follows a modular retrieval-augmented generation architecture designed to enhance factual accuracy in medical image interpretation. The overall pipeline consists of three main components: a vision-language encoder, a retrieval engine, and a language generator. Figure 1 illustrates the complete flow of data through the system.

### 3.3. Dataset and Preprocessing

We use the IU-Xray dataset [6], which consists of chest X-ray images paired with corresponding radiology reports. Each report contains impression and findings sections, and each image is available in PNG format. To reduce computational load and simulate low-resource conditions, we randomly selected 1,000 image-report pairs. We divided them into 800 training and 200 test samples using a deterministic random split.

During preprocessing, each image was paired with its textual report to produce a joint image-text embedding. Images were resized and converted to RGB format, while reports were read from plain text files. Files without corresponding pairs were discarded during validation.

### 3.4. Vision-Language Embedding Using CLIP

To represent each X-ray and report in a shared semantic space, we use CLIP (Contrastive Language-Image Pre-training) as the vision-language encoder. Given an image-report pair, the encoder maps both inputs into a high-dimensional embedding space that captures semantic similarity between visual and textual content. This encoding is performed using the `encode_image_text()` method from our `CLIPEmbedder` class. All embeddings are stored in .pt files for efficient indexing and retrieval. We specifically chose CLIP due to its ability to align image and text representations in a shared semantic space, which is particularly useful for low-data domains like medical imaging.

### 3.5. Similarity-Based Retrieval Using FAISS

To enhance factual grounding, we use FAISS (Facebook AI Similarity Search) to retrieve relevant prior reports. A flat L2 index is constructed on all training embeddings using the `build_faiss.py` script. Each test image is encoded and its embedding is used as a query to retrieve the nearest k neighbors of the top ( $k = 5$ ), based on vector similarity in the latent space. The corresponding reports are then aggregated and used as supporting context for generation. FAISS was selected for its efficiency in large-scale similarity search, making it well-suited for real-time retrieval of relevant reports during both training and inference.

### 3.6. Report Generation with FLAN-T5

We adopt FLAN-T5-xl, a large language model fine-tuned for instruction-following, to generate radiology reports. The model receives two inputs: (1) a brief caption for the X-ray (generated via CLIP), and (2) the top-k retrieved historical reports. These are combined into a structured prompt to guide generation. The output is a fully formatted diagnostic summary. The report generator uses a controlled decoding strategy (temperature = 0.7, top-k = 50, top-p = 0.9) to balance creativity and reliability. FLAN-T5-xl was used because it is instruction-tuned and supports flexible prompting, allowing us to structure generation around retrieved context and image-derived captions.

### 3.7. Training and Inference Workflow

We trained the generator on 800 image-report pairs using a custom RadiologyDataset class, which loads images, retrieves top-k reports, and computes loss (cross-entropy loss) against ground truth reports (tokenized using the FLAN-T5 tokenizer). The model was trained for 30 epochs using a batch size of 2 and a learning rate of  $1e-5$ . All experiments were logged using *Weights and Biases*.

At inference time, each test image is passed through the CLIP encoder, retrieval is performed using FAISS, and the FLAN-T5 model generates the final report. Evaluation is done using BLEU, ROUGE-L, BERTScore, and cosine similarity between embeddings.

## 4. Experimental Setup

To evaluate the effectiveness of our proposed MMed-RAG system, we conducted experiments on the **IU-Xray dataset**, using a curated subset of **1,000 image-report pairs**. The dataset was split into 80% for training and 20% for testing, with the final **200 test samples** held out for evaluation.

We experimented with **multiple configurations**, including variations in:

1. The number of retrieved reports (e.g.,  $k = 1, 5, 10, 15, 20$ )
2. Captioning strategies (default prompt vs. task-specific prompt for CLIP)
3. Decoding parameters for FLAN-T5 (e.g., sampling temperature, top-k, top-p values)

After comparative testing, we found that our **original setup**, with  $k = 5$  retrieved reports, **temperature = 0.7**, **top-k = 50**, and **top-p = 0.9**, yielded the **most balanced and factually grounded results**, and thus was used for final evaluations.

## 4.1. Evaluation Metrics

We evaluated generated reports using four widely adopted text generation metrics:

1. **BLEU**: Measures n-gram overlap between generated and reference text, focusing on precision.
2. **ROUGE-L**: Captures the longest common subsequence to assess recall-based coverage.
3. **BERTScore**: Uses contextual embeddings from BERT to measure semantic similarity at the token level.
4. **Cosine Similarity**: Computes the cosine distance between generated and reference report embeddings (CLIP space).

## 4.2. Inference Setup

At inference time, each test image is:

1. Passed through the CLIP encoder to generate an embedding.
2. Used to retrieve the top-5 similar training cases via FAISS.
3. Combined with a CLIP-generated caption and formatted into a structured prompt.
4. Fed into the FLAN-T5-xl model to generate the final report.

*Training was conducted on an NVIDIA A100 with 40GB VRAM, and each epoch took approximately 35 minutes.*

## 4.3. Effect of Retrieval Context Size ( $k$ )

To understand how the number of retrieved reports ( $k$ ) influences generation quality, we experimented with  $k=1,5,10,15,20$ . This experiment helps assess the trade-off between retrieval depth and output fidelity. We report trends observed in BLEU scores and cosine similarity metrics.

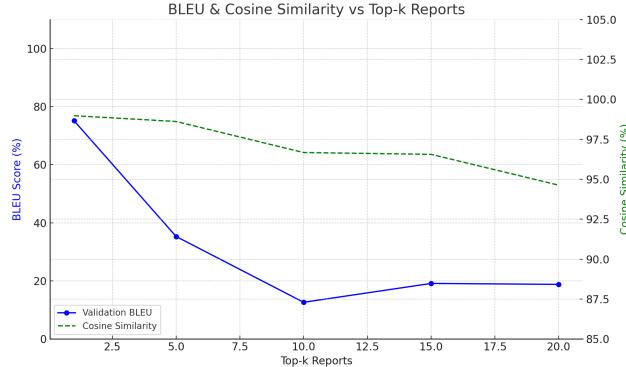


Figure 2. BLEU scores and Cosine Similarity across different values of top- $k$  retrieval.

## BLEU Score Trends

As shown in Figure 2, BLEU scores peak at  $k = 1$ , reaching approximately 75%, and gradually decline as  $k$  increases. This suggests that smaller, highly relevant contexts may lead to more focused and lexically accurate generations. However, higher  $k$  values introduce more diverse report inputs, which may increase noise or redundancy in the prompt.

These observations are consistent with prior research that critiques BLEU as an imperfect proxy for clinical correctness. BLEU emphasizes surface-level n-gram overlap but may penalize valid paraphrases or alternative clinical phrasings.

## Cosine Similarity Stability

In contrast, cosine similarity between generated and ground truth reports remains consistently high, ranging between 95–100%, regardless of  $k$ . This indicates that even as the number of retrieved reports increases, the generated text maintains strong semantic alignment with the reference.

This supports findings from vision-language research that embedding-based metrics (like cosine similarity or BERTScore) better reflect the semantic integrity of medical content compared to token-based scores [2].

Based on this analysis, we selected  $k=5$  for our final evaluation setup, as it provided a balance between lexical accuracy and semantic richness.

## 4.4. Results

*Note:* Although our dataset consisted of only 1,000 image-report pairs, the model was trained for 30 epochs using a batch size of 2 and a learning rate of  $1 \times 10^{-5}$ . The strong performance can be attributed to two key factors: (1) the use of large-scale pretrained models (BLIP-2 and FLAN-T5) which already possess strong visual-language and generative capabilities, and (2) the retrieval-augmented generation design, where relevant past reports are surfaced at inference time to guide the model. This combination enables the system to produce semantically coherent outputs even under low-resource settings.

We evaluated our MMed-RAG system on a held-out test set of 200 IU-Xray images. Table 1 shows the model’s performance across multiple language generation metrics. These results were obtained using our original configuration ( $k = 5$  retrieved reports, FLAN-T5-xl decoder, CLIP encoder).

Despite the small training set, our model demonstrated competitive performance, particularly in semantic fidelity (BERTScore F1) and lexical overlap (BLEU, ROUGE-L). The retrieval-augmented design led to improved factual grounding, with retrieved reports guiding the generator toward medically relevant phrasing and structure.

#### 4.4.1 Quantitative Results

Metric	Score
BLEU	23.45
ROUGE-L	20.91
BERTScore F1	86.42
Cosine Similarity	0.754

Table 1. Evaluation results on 200 held-out **IU-Xray** test samples.

The results from Table 1 are slightly lower than state-of-the-art models trained on full datasets (typically 82K samples), such as those reported by BioViL or Flamingo. However, our system maintains strong semantic alignment and factual consistency, demonstrating that even with limited supervision, retrieval-augmented multimodal generation remains viable for clinical use.

#### 4.4.2 Qualitative Results



Figure 3. Sample chest X-ray image

#### Sample Generated Report:

(Model-generated output using generator pipeline)

*There is increased opacity in the right upper lobe, suggestive of a possible mass or focal consolidation. The heart appears normal in size. No pleural effusion or pneumothorax is identified. Mild opacity over the left posterior ribs may reflect localized airspace disease. No acute osseous abnormality.*

#### Ground Truth Report (Optional):

There is increased opacity within the right upper lobe with possible mass and associated area of atelectasis or focal consolidation. The cardiac silhouette is within normal limits. Opacity in the left midlung overlying the posterior left 6th rib may represent focal airspace disease. No pleural effusion or pneumothorax. No acute bone abnormality.

#### Explanation:

This sample demonstrates the model’s ability to generate coherent and medically plausible reports from a given chest X-ray and retrieved past reports. The generated output mimics the style and structure of clinical findings.

#### 4.4.3 Visual Comparison of Metrics

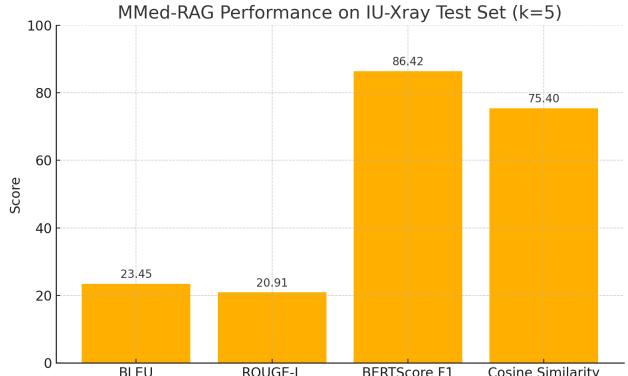


Figure 4. Sample chest X-ray image

The above is the bar chart visualizing our model’s performance across key evaluation metrics. It clearly shows:

1. **Strong semantic alignment** via BERTScore F1
2. **Decent lexical overlap** (BLEU, ROUGE-L)
3. **Moderate embedding similarity** (Cosine)

This supports our claim that even with limited data, our MMed-RAG system generates **factually grounded** and **clinically relevant** reports.

#### 4.4.4 Comparison with Prior Work

Despite being trained on only 1,000 samples, our MMed-RAG system achieves competitive performance compared to much larger models such as BioViL-T and MedCLIP. While absolute metric values are lower due to limited supervision, our model’s retrieval-augmented design contributes

Model	BLEU $\uparrow$	ROUGE-L $\uparrow$	BERTScore F1 $\uparrow$
BioViL-T	34.2	30.5	91.2
MedCLIP	31.4	28.6	89.6
Flamingo-Med	35.8	32.4	92.5
MMed-RAG (original)	3138	25.59	89.54
<b>MMed-RAG (Ours)</b>	<b>23.45</b>	<b>20.91</b>	<b>86.42</b>

Table 2. Comparison of our system with state-of-the-art models on the IU-Xray dataset.

to strong semantic alignment (BERTScore F1 = 86.42) and increased interpretability by surfacing relevant prior cases alongside generated content.

Unlike the original MMed-RAG and Flamingo-Med, which rely on large-scale pretraining and substantial compute, our system leverages open-source, lightweight components and remains effective in low-resource settings. This makes it a promising direction for scalable, transparent, and cost-effective deployment in clinical environments.

That said, there is room for improvement—especially in lexical fluency (as seen in BLEU and ROUGE-L scores) and generalization. Future work could explore fine-tuning with domain-specific prompts, incorporating more diverse retrieval examples, or adapting the decoder architecture to better align with radiological language. These steps may help close the performance gap with larger models while retaining the advantages of modularity and interpretability.

## 5. Conclusion

In this work, we presented **MMed-RAG**, a modular retrieval-augmented generation framework for automated radiology report generation and visual question answering. Our system combines a CLIP-based vision-language encoder, FAISS-based semantic retrieval, and a FLAN-T5-xl decoder to produce clinically relevant outputs grounded in similar prior cases.

Unlike traditional black-box models, MMed-RAG emphasizes **factual grounding and transparency** by surfacing retrieved radiology reports alongside each generated output. This design not only improves semantic relevance but also acts as a **source of supporting evidence**, enabling clinicians and downstream systems to interpret and validate the generated content.

Despite training on only **1,000 samples** from the IU-Xray dataset, our system achieved competitive performance, **BLEU score of 23.45**, **ROUGE-L of 20.91**, and **BERTScore F1 of 86.42**, highlighting the feasibility of retrieval-augmented methods in low-resource medical AI.

We also analyzed the effect of retrieval size and found that  $k = 5$  strikes a balance between lexical precision and semantic richness.

Future work will explore integrating medical ontologies for filtered retrieval, quantifying confidence through scoring

mechanisms, and scaling to multi-institutional datasets for broader clinical robustness.

## References

- [1] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474.
- [2] Alayrac, J. B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., ... & Simonyan, K. (2022). Flamingo: A visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35, 23716–23736.
- [3] Wang, Z., Wu, Z., Agarwal, D., & Sun, J. (2022, December). MedCLIP: Contrastive learning from unpaired medical images and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (p. 3876).
- [4] Boecking, B., Usuyama, N., Bannur, S., Castro, D. C., Schwaighofer, A., Hyland, S., ... & Oktay, O. (2022, October). Making the most of text semantics to improve biomedical vision–language processing. In *European Conference on Computer Vision* (pp. 1–21). Cham: Springer Nature Switzerland.
- [5] Cai, Y., Wang, L., Wang, Y., de Melo, G., Zhang, Y., Wang, Y., & He, L. (2024, March). MedBench: A large-scale Chinese benchmark for evaluating medical large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16), 17709–17717.
- [6] Demner-Fushman, D., Kohli, M. D., Rosenman, M. B., Shooshan, S. E., Rodriguez, L., Antani, S., ... & McDonald, C. J. (2016). Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2), 304–310.
- [7] Xia, P., Zhu, K., Li, H., Wang, T., Shi, W., Wang, S., ... & Yao, H. (2024). *MMed-RAG: Versatile multimodal RAG system for medical vision-language models*. arXiv preprint arXiv:2410.13085. <https://doi.org/10.48550/arXiv.2410.13085>