Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Towards Multimodal Retrieval-Augmented Generation for Medical Visual Question Answering

**Mai A. Shaaban**

mai.shaaban@alexu.edu.eg

Alexandria University

**Mohammad Reza Zarei**

Carleton University

**Adnan Khan**

Carleton University

**Abbas Akkasi**

Carleton University

**Mohammad Yaqub**

Mohamed bin Zayed University of Artificial Intelligence

**Majid Komeili**

Carleton University

**Article**

**Keywords:**

**Posted Date:** October 28th, 2025

**DOI:** https://doi.org/10.21203/rs.3.rs-7752202/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Towards Multimodal Retrieval-Augmented Generation for Medical Visual Question Answering

**Mai A. Shaaban[1,2,*], Mohammad Reza Zarei[3], Adnan Khan[3], Abbas Akkasi[3], Mohammad Yaqub[4], and Majid Komeili[3]**

[1]Faculty of Science, Alexandria University, Department of Mathematics and Computer Science, Alexandria, Egypt
[2]Mohamed bin Zayed University of Artificial Intelligence, Department of Machine Learning, Abu Dhabi, United Arab Emirates
[3]School of Computer Science, Carleton University, Ottawa, Canada
[4]Mohamed bin Zayed University of Artificial Intelligence, Department of Computer Vision, Abu Dhabi, United Arab Emirates
[*]mai.shaaban@alexu.edu.eg

## ABSTRACT

Medical visual question answering (MedVQA) is a critical AI healthcare task that combines medical image analysis with natural language understanding to assist clinicians in decision-making. While medical vision-language models have shown promise in this domain, they struggle with factual inaccuracies and hallucinations. Retrieval-augmented generation (RAG) improves the factual accuracy by grounding responses in external knowledge, yet text-only or image-only retrieval systems struggle with the inherently multimodal nature of medical data, leading to information loss. This paper proposes a novel multimodal RAG framework tailored for MedVQA, which leverages multimodal data, including medical images, reports, and generated captions, to provide more accurate clinical answers. We introduce a training paradigm that uses captions as auxiliary supervision, enhancing cross-modal alignment via contrastive learning. Comprehensive evaluations on MedVQA benchmarks demonstrate the framework's effectiveness, achieving a 7% average accuracy improvement over unimodal RAG baselines. Our study has the potential to better support clinicians in delivering accurate, timely, and trustworthy patient care by improving the reliability of MedVQA systems. The code is publicly available at https://github.com/AiMl-hub/MM-RAG-MedVQA.

## Introduction

Medical Visual Question Answering (MedVQA) is a pivotal field in AI healthcare, aiming to enhance clinical decision-making by enabling AI systems to interpret complex medical images and respond to natural language queries from clinicians[1]. Unlike general VQA, MedVQA demands a profound understanding of specialized medical imagery and complex domain knowledge, integrating computer vision for medical image analysis with natural language processing to understand questions and generate clinically relevant answers[2]. The clinical applications of MedVQA are substantial, promising to alleviate the burden on healthcare professionals, reduce diagnostic errors, and improve patient outcomes by automating image interpretation[2]. Use cases include assisting in the interpretation of various medical images such as X-rays, MRIs and CT scans[3,4]. Despite progress, MedVQA faces unique challenges, primarily the diversity between visual and textual modalities and the complexity of medical images[5,6].

Vision-Language Models (VLMs), have demonstrated remarkable capabilities in integrating visual and textual information, achieving strong performance in various tasks that involve natural images and text[7]. These models are designed for multimodal understanding, combining robust image feature extraction with advanced language comprehension and generation. Despite these advancements, the direct application of VLMs in the medical domain reveals significant limitations, particularly with respect to factual precision and the critical issue of hallucination[8]. Hallucinations, where models generate plausible but incorrect responses, can lead to erroneous recommendations and jeopardize patient care. Most VLMs lack specialized medical knowledge, relying primarily on generic information rather than the domain expertise required in healthcare, which often results in the missing of critical visual cues for diagnosis[9]. These limitations are further worsened by the scarcity and static nature of medical VLM datasets, which restrict models from acquiring the depth of knowledge necessary for accurate interpretation[2].

Retrieval-Augmented Generation (RAG) has emerged as a promising solution to address the limitations of static knowledge and factual inaccuracies in standalone generation (SG). RAG enhances the performance of large language models (LLMs) by grounding their responses in external, up-to-date knowledge sources, thereby improving factual accuracy and mitigating hallucinations[10]. RAG systems typically retrieve relevant information from curated knowledge bases (e.g., medical databases, literature) and condition the generative model with this context to formulate responses. This dynamic knowledge integration

is a significant advancement that allows models to access current information without costly retraining, which is particularly beneficial in rapidly evolving fields such as medicine[11]. RAG in medical applications has shown considerable improvements, achieving up to a 47.4% increase in factual accuracy compared to baseline methods[12]. Beyond accuracy, RAG improves transparency by enabling users to trace AI-generated responses back to their sources, fostering trust[13]. However, unimodal (text-only) RAG applications present distinct limitations when confronted with the inherently multimodal nature of medical data. The primary challenge stems from its reliance on textual knowledge bases, which are insufficient for tasks that require deep understanding and reasoning from medical images[14]. This creates a "unimodal bottleneck" where crucial visual information cannot be directly leveraged, leading to a loss of critical information and limiting the system's ability to answer questions that require deep visual interpretation[12,14,15].

Multimodal Retrieval-Augmented Generation (MM-RAG) offers a powerful solution to overcome the limitations of unimodal RAG by directly integrating diverse data types, such as medical images and corresponding medical reports, into retrieval processes[16,17]. This comprehensive approach is crucial for developing accurate, contextually relevant, and reliable AI systems in healthcare. MM-RAG systems are designed to combine visual data from medical images (e.g., X-rays, MRIs, CT scans) with textual data (e.g., clinical notes, medical literature) to generate informed and precise responses. Building MM-RAG pipelines involves addressing the inherent challenges of managing and integrating information across different modalities, with architectural approaches that include embedding all modalities in the same vector space or using separate stores with a multimodal reranker[12,13,15]. Although RAG and MM-RAG aim to reduce generative hallucinations by grounding responses in external knowledge bases, they can inadvertently create retrieval-based hallucinations[10,18,19]. Two major forms include "selection-hallucination," where insufficient or excessive retrieval introduces irrelevant information, and "context-generation-hallucination," where VLMs misinterpret retrieved images into incorrect textual descriptions[12,13,20]. This means that a poorly designed retrieval-based system can inject new errors. Furthermore, an over-reliance on retrieved data can paradoxically lead to factual errors, sometimes degrading performance rather than consistently enhancing it. Thus, to address this issue, developing a retrieval model with improved multimodal alignment, such as through enhanced loss functions or domain-specific constraints, could reduce factual inconsistencies[21].

Motivated by the limitations of unimodal retrieval and the need to enhance multimodal understanding, this study proposes **MM-RAG-MedVQA**, a novel **M**ultimodal **R**etrieval-**A**ugmented **G**eneration system for **M**edical **V**isual **Q**uestion **A**nswering. This framework is the first to integrate end-to-end generated captions for fine-tuning to effectively address complex, knowledge-intensive clinical queries. The MM-RAG-MedVQA methodology is based on three core components: image caption generation, clinically relevant context retrieval from a large-scale multimodal medical knowledge database, and precise response generation using a medical VLM.

A key innovation of our approach is its novel training paradigm for improving the alignment between visual and textual features, which leverages image captions as auxiliary supervision signals in addition to medical images and reports. These captions serve as proxies for region-specific details (e.g., "right lower lobe consolidation"), enabling precise retrieval of medical reports with spatially aligned findings. This strategy improves the feature alignment between visual and textual embeddings through a carefully designed contrastive loss function, enhancing the model's ability to understand and relate multimodal information. Specifically, our framework employs an MM-RAG retrieval mechanism, where joint embeddings of the query image and its generated caption are compared against a vector database of image-report pairs. The most relevant medical reports are then retrieved and concatenated with the query image, its caption, and the user's question. This augmented input is fed into a frozen medical VLM, which synthesizes this rich multimodal context to generate clinically accurate and evidence-based responses. The effectiveness of the proposed approach is validated through comprehensive automated and human evaluations on two MedVQA benchmarks, demonstrating superior performance in clinically relevant scenarios.

The main contributions of this study are summarized as follows:

1. Developing MM-RAG-MedVQA, a cross-modal RAG system, which integrates medical images, reports, and end-to-end generated captions to address complex, knowledge-intensive clinical queries.

2. Introducing a novel training paradigm to ensure cross-modal feature alignment that leverages generated captions as auxiliary supervision signals alongside medical images and reports. This paradigm utilizes contrastive learning to align visual and textual embeddings.

3. Conducting automated and human evaluations on two MedVQA benchmarks, demonstrating superior performance in accuracy, faithfulness, and completeness. Our evaluation protocol offers a structured and reliable approach for validating such complex tasks.

## Results

This section outlines the structure and composition of the external knowledge database, describes the evaluation datasets used, specifies the hyperparameter configurations, and details the evaluation protocol employed to assess model performance. The section also presents a comprehensive analysis of the experimental results obtained using our proposed MM-RAG-MedVQA framework, comparing it against SG and unimodal RAG methods. Additionally, we assess the faithfulness and completeness of the generated answers using automated and human evaluations.

### Datasets

We use three datasets to build the external knowledge base and evaluate our framework. MIMIC-CXR-JPG[22,23], a comprehensive chest X-ray dataset containing paired images and radiology reports. For each of the 14 diagnostic classes, we sample 100 image-report pairs from the training split to initialize the vector database and fine-tune the model. We replace the original reports with preprocessed versions from the CXR-PRO[24] dataset, removing references to prior reports or images that are irrelevant for single-image analysis.

For evaluation, we use the test splits of two publicly available MedVQA datasets: MIMIC-CXR-VQA[25] and Medical-Diff-VQA[26]. Both datasets follow the standard VQA format, providing diverse image-question-answer triplets. Examples of questions are provided in Table 1.

| Dataset | Question Type | Example |
|---|---|---|
| MIMIC-CXR-VQA | Abnormality | Is the <object> abnormal? |
| | Presence | Are there any <category> in the <object>? |
| | Attribute | Which <category> is related, <attribute_1> or <attribute_2>? |
| | Anatomy | List all common anatomical locations related to both <attribute_1> and <attribute_2>. |
| Medical-Diff-VQA | Abnormality | What abnormalities are seen in the <location>? |
| | Presence | Is there any evidence of <abnormality>? |
| | Location | Where in the image is the <abnormality> located? |
| | Level | What level is the <abnormality>? |
| | Type | What type is the <abnormality>? |

**Table 1.** Question types and examples from MIMIC-CXR-VQA and Medical-Diff-VQA datasets. <category> and <abnormality> tags refer to disease names; <object> can be anatomical structures (e.g., "left lung" or "right lung"); <attribute> can include anatomical findings, diseases, devices, tubes/lines, and technical assessment.

### Evaluation

We employ a hybrid approach that integrates automated and manual assessments. This strategy tackles the limitations of traditional reference-based metrics (e.g., BERT Score), which often fail to assess open-ended medical questions[27]. Such questions require understanding and semantic comprehension, qualities that traditional metrics struggle to measure effectively.

The automated evaluation uses natural language inference (NLI) to determine if the generated answers fully align with or contradict the reference ground truth. This approach improves confidence scoring and identifies cases in which the responses are factually correct but do not address all aspects of the questions[28–30]. Additionally, we use G-Eval[27], an LLM-as-a-judge framework that uses chain-of-thought (CoT) reasoning to assess outputs against tailored evaluation criteria. The process consists of three main steps:

1. Step Formulation: The LLM evaluator turns the natural language descriptions of evaluation rules into clear step-by-step instructions,
2. Assessment Phase: The LLM evaluator applies these predefined steps to systematically review system outputs,
3. Score Computation: The model's evaluations are weighted by their log-probability values to generate an aggregated G-Eval score.

We specifically evaluate two metrics using G-Eval: the *faithfulness* and *completeness* of system responses. Faithfulness measures whether the output of a model is factually consistent with the given input or the source context, free of hallucinations or contradictions, while completeness assesses whether the response covers all relevant aspects of the input. These metrics form a robust framework for evaluating AI-generated outputs against custom criteria. The evaluation steps are presented in Table 2.

| Evaluation Metric | CoT Prompt |
|---|---|
| Faithfulness | 1. Extract medical claims or diagnoses from the actual output<br>2. Verify each medical claim against the retrieved contextual information<br>3. Identify any contradictions or unsupported medical claims that could lead to misdiagnosis<br>4. Heavily penalize hallucinations, especially those that could result in incorrect medical advice |
| Completeness | 1. Determine if the response answers every part of the input question<br>2. Identify any missing elements, skipped sub-questions, or incomplete reasoning<br>3. Check whether the output provides sufficient detail for each aspect mentioned<br>4. Do not penalize for brevity if the coverage is complete and accurate |

**Table 2.** Chain-of-thought prompts for the evaluator LLM.

To validate the reliability of LLM-based evaluation metrics, we perform manual verification on a randomly selected representative subset of 100 model-generated responses. For each instance in the subset, we compare the generated answer with the ground truth to determine whether it was factually aligned (correct) or contradictory (incorrect). We then assess whether the LLM evaluator accurately classifies these cases. This step ensures that the evaluator's judgments are consistent with human verification, thereby strengthening the credibility of the automated assessment methodology.

## Implementation Details

For multimodal encoding, we employed BioMedCLIP[31] with a maximum token length of 256 and a vector dimension of 512. The image encoder processes images of size $224 \times 224$ pixels, divided into $16 \times 16$ pixel patches. We used a batch size of 32 over 50 epochs with a learning rate of $1 \times 10^{-5}$ on a single NVIDIA A100-40GB GPU. Optimization was performed using AdamW with a decoupled weight decay of 0.05. Table 3 summarizes training configurations.

For retrieval, we used Qdrant[32], a high-performance vector database, to store and index multimodal embeddings. The hyperparameter *top-k* was set at values ranging from 1 to 10. For answer generation, we used LLaVA-Med-v1.5-7B[33], an open-source, state-of-the-art medical VLM. For caption generation, we used MAIRA-2[34]. For NLI-based evaluation, we used SciFive[35], while for G-Eval we used GPT-4o mini[36].

| Component | Parameter | Value |
|---|---|---|
| Encoder | Max Token Length | 256 |
| | Vector Dimensionality | 512 |
| | Image Size | $224 \times 224$ |
| | Patch Size | $16 \times 16$ |
| Training | Epochs | 50 |
| | Batch Size | 32 |
| | Learning Rate | $1 \times 10^{-5}$ |
| | Weight Decay | 0.05 |
| | Optimizer | AdamW |
| | Loss Temperature ($\tau$) | 0.1 |

**Table 3.** Model architecture and training hyperparameters.

## Quantitative Performance

The quantitative results are summarized in Table 4. The SG method achieved the lowest accuracy among all methods. The RAG approach showed better performance, with the text-only retrieval variant outperforming SG by 2.76% in MIMIC-CXR-VQA and 3.32% in Medical-Diff-VQA. Performance was further improved with image-based retrieval; the $I \rightarrow I$ retrieval variant outperformed SG by +3.34% and +10.52% on the respective datasets.

Our proposed MM-RAG-MedVQA framework achieved the highest accuracy scores of 53.48% on MIMIC-CXR-VQA and 67.98% on Medical-Diff-VQA. These results correspond to an average improvement of approximately 7%, including gains of +5.64% (MIMIC-CXR-VQA) and +15.6% (Medical-Diff-VQA) over the SG baseline, as well as +2.3% (MIMIC-CXR-VQA) and +5.08% (Medical-Diff-VQA) compared with the best retrieval-augmented baseline.

| Dataset | Method | Retrieval Modality | Embedding Model | Accuracy (%) ↑ |
|---|---|---|---|---|
| MIMIC-CXR-VQA | SG | – | – | 47.84 |
| | Unimodal RAG | $C \rightarrow M$ | BioMedCLIP (BERT) | 50.60 |
| | | $I \rightarrow I$ | BioMedCLIP (ViT) | 51.18 |
| | MM-RAG-MedVQA (w/o FT) | $I+C \rightarrow I+M$ | BioMedCLIP (ViT+BERT) | 52.24 |
| | MM-RAG-MedVQA (w/o $\mathscr{L}_{I \rightarrow C}$) | $I+C \rightarrow I+M$ | FT BioMedCLIP | 53.22 |
| | MM-RAG-MedVQA ($\mathscr{L}_{ours}$) | $I+C \rightarrow I+M$ | FT BioMedCLIP | **53.48** |
| Medical-Diff-VQA | SG | – | – | 52.38 |
| | Unimodal RAG | $C \rightarrow M$ | BioMedCLIP (BERT) | 55.70 |
| | | $I \rightarrow I$ | BioMedCLIP (ViT) | 62.90 |
| | MM-RAG-MedVQA (w/o FT) | $I+C \rightarrow I+M$ | BioMedCLIP (ViT+BERT) | 61.82 |
| | MM-RAG-MedVQA (w/o $\mathscr{L}_{I \rightarrow C}$) | $I+C \rightarrow I+M$ | FT BioMedCLIP | 63.50 |
| | MM-RAG-MedVQA ($\mathscr{L}_{ours}$) | $I+C \rightarrow I+M$ | FT BioMedCLIP | **67.98** |

**Table 4.** Accuracy comparison on MedVQA benchmarks. Abbreviations: $I$ (Image), $C$ (Caption), $M$ (Medical report), FT (Fine-tuned), $\mathscr{L}_{I \rightarrow C}$ (Image-Caption Loss), $\mathscr{L}_{ours}$ (our composite loss).

In terms of answer quality, the system achieved a *faithfulness* score of 61.90% and a *completeness* score of 86.54%. Manual verification revealed a strong agreement (above 98%) between the automated LLM evaluator and human judgments.

An ablation study on encoder freezing (Table 5) revealed that for MIMIC-CXR-VQA, performance showed minimal variation across configurations. For Medical-Diff-VQA, significant gains were observed under multimodal tuning. Further analysis revealed high performance when using 3-10 retrieved medical reports. Performance consistently declined across all methods when larger retrieval sets were used.

| Training Configuration | MIMIC-CXR-VQA | Medical-Diff-VQA |
|---|---|---|
| Unfrozen visual, Frozen text | 53.20 | 64.60 |
| Frozen visual, Unfrozen text | 52.58 | 61.40 |
| Both encoders unfrozen | 53.48 | 67.98 |

**Table 5.** Comparison of accuracy across different encoder configurations.

### Qualitative Performance

Figure 1 presents two example queries that compare unimodal RAG systems with the proposed MM-RAG-MedVQA system. In Query A, the text-only retrieval method concluded "No pneumothorax," the image-only method failed to detect the condition, and MM-RAG-MedVQA correctly identified a "small left apical pneumothorax".

In Query B, both unimodal methods missed a lung opacity, whereas MM-RAG-MedVQA correctly localized it to the "left lung." The mention of "left pleural effusion" in the retrieved medical report was consistent with this finding.

## Discussion

Our results demonstrate that the MM-RAG-MedVQA framework enhances performance in Medical Visual Question Answering. The consistent gains across both datasets, particularly the +15.6% improvement over the SG baseline on the challenging

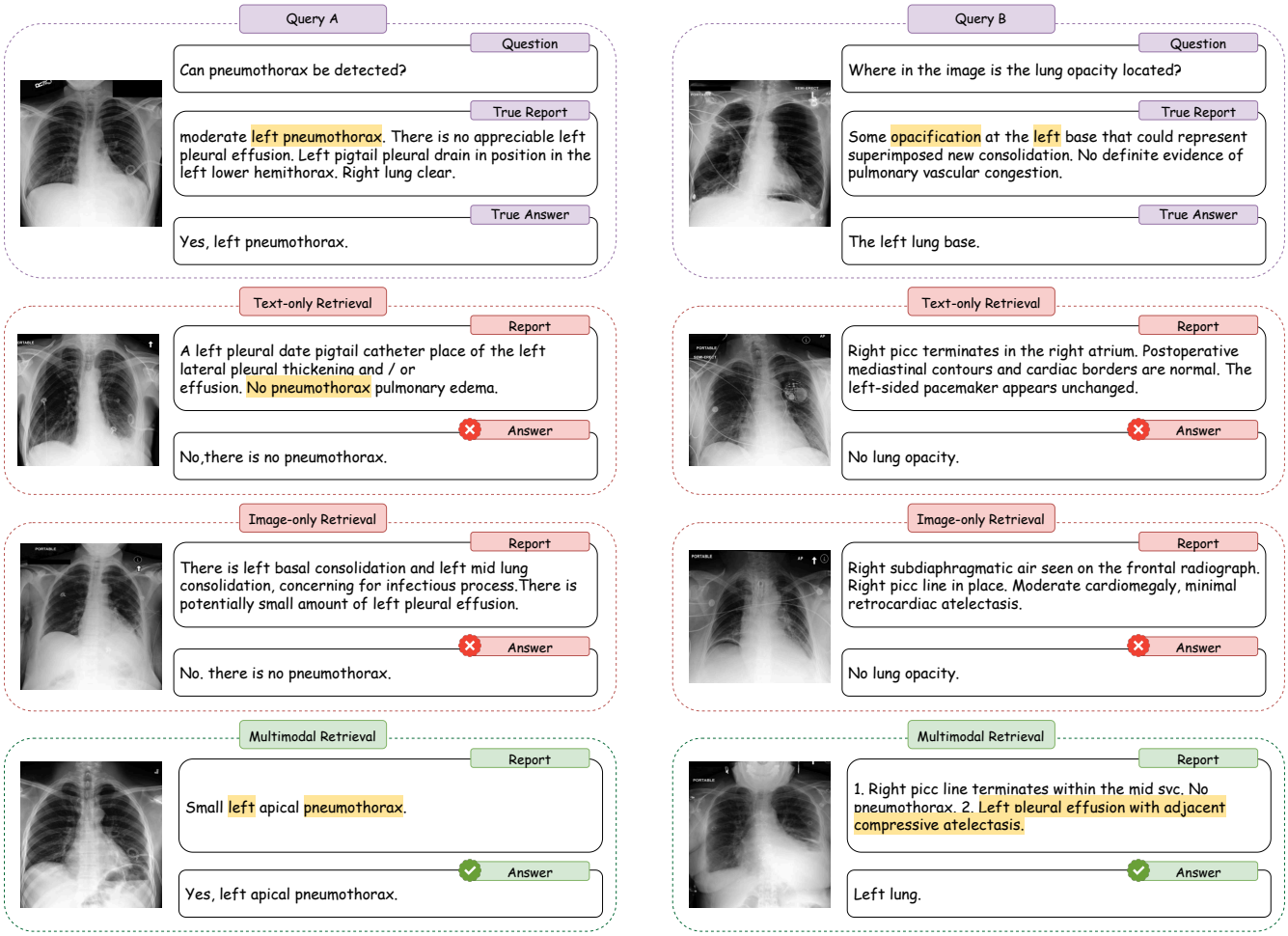


**Figure 1.** Examples of queries and the generated answers comparing unimodal RAG against the proposed MM-RAG-MedVQA approach.

Medical-Diff-VQA, underscore the effectiveness of our multimodal retrieval approach. The superior performance suggests that combining image and caption embeddings provides a more clinically relevant context than unimodal alternatives, effectively grounding responses in domain-specific knowledge.

The wide performance gap between image-based retrieval and SG, especially in Medical-Diff-VQA (+10.52% over SG), indicates that visual similarity retrieval captures complementary signals to text-only methods. This is likely because Medical-Diff-VQA's focus on fine-grained visual distinctions (e.g., subtle lesions or anatomical variants) allows image embeddings to provide more discriminative retrieval cues than image captions alone. These findings highlight a clear need for multimodal retrieval systems capable of fusing cross-modal signals in clinical settings, where neither images nor text alone suffice to mitigate hallucinations or outdated knowledge.

The gains we observed can be attributed to our composite loss function, which jointly optimizes alignment across images, generated captions, and radiology reports. The ablation study confirms that the combination of $\mathscr{L}_{I\to C}$ and $\mathscr{L}_{I\to M}$ losses outperforms using $\mathscr{L}_{I\to M}$ alone, indicating that the captions provide crucial semantic grounding that enhances the overall performance of the system.

Qualitative analysis reinforces the quantitative findings. MM-RAG-MedVQA demonstrated superior accuracy in the detection and localization of conditions such as pneumothorax and lung opacity, reducing false negatives that plagued unimodal methods. This illustrates the system's ability to integrate and interpret both textual and visual data, leading to improved diagnostic precision. Furthermore, the system's ability to make clinically consistent inferences, such as linking "left pleural effusion" to "opacity," suggests that combining textual context with visual cues enables a more comprehensive spatial understanding.

The notably higher *completeness* score (86.54%) demonstrates the framework's ability to provide comprehensive answers, although there is room for improvement in reducing minor factual inaccuracies. The strong agreement between automated and

human evaluation validates our methodology, while observed disagreements highlight the need for more granular evaluation criteria in medical contexts.

The finding that performance degrades with larger retrieval sets (>10 reports) emphasizes the critical importance of retrieval precision in clinical applications, where irrelevant or contradictory information could mislead the answer generation process. This also highlights a limitation of current cosine similarity metrics and points to a fruitful future direction for incorporating more sophisticated semantic matching techniques.

Our MM-RAG-MedVQA framework demonstrates how retrieval-augmented multimodal systems can enhance the reliability of medical AI assistants. These systems could support clinical workflows by: (1) reducing diagnostic time through automated preliminary findings, (2) minimizing hallucinations that plague pure generative approaches, and (3) providing transparent evidence trails through retrieved medical reports. However, the persistent gap in faithfulness indicates that these systems should currently serve as adjuncts rather than replacements for clinical judgment.

Future directions include: (1) integrating electronic health records for patient-specific context, (2) developing clinician-in-the-loop refinement protocols, and (3) exploring advanced semantic retrieval metrics. The framework's modular design permits these incremental improvements while maintaining core retrieval-augmentation benefits.

In conclusion, this paper introduced MM-RAG-MedVQA, a novel multimodal retrieval-augmented framework that improves the clinical accuracy of the answers generated in the MedVQA system. Our approach addressed key limitations of existing methods by jointly aligning medical images, reports, and generated captions, leading to fewer factual hallucinations. The success of this framework highlights the critical importance of multimodal evidence integration for building reliable clinical decision support tools that can responsibly augment expert judgment.

## Methods

This section presents our MM-RAG-MedVQA framework and unimodal retrieval approaches. The methodology consists of three core components: (1) image caption generation, (2) context retrieval from medical knowledge sources, and (3) answer generation using a frozen VLM. Figure 2 illustrates the overall system architecture.

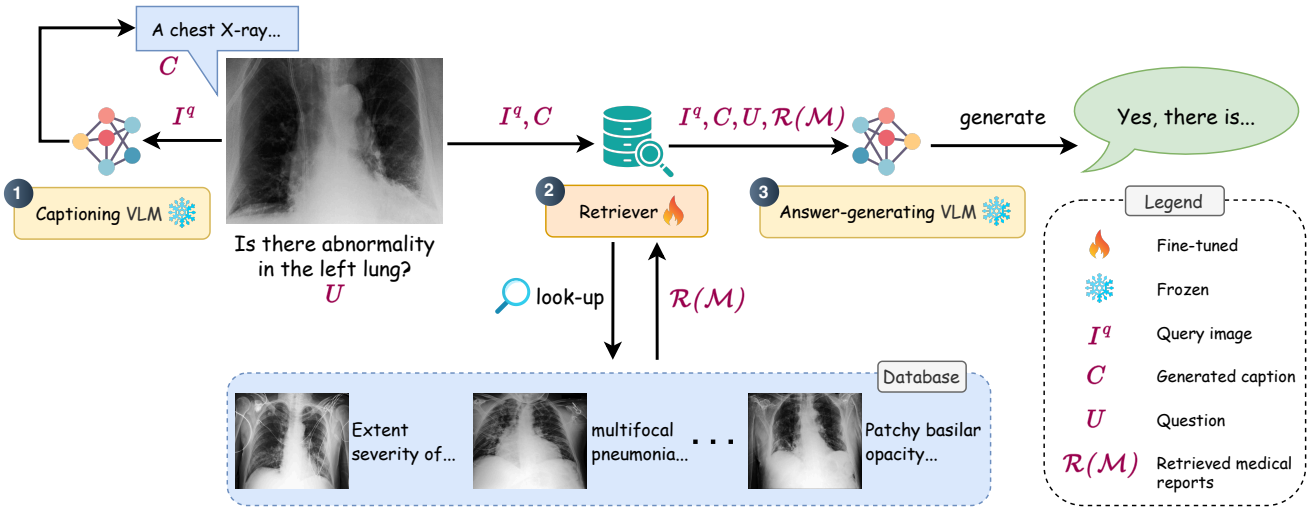


**Figure 2.** Architecture overview of the proposed MM-RAG-MedVQA framework: (1) An end-to-end module generates descriptive captions for query images; (2) The query image and its corresponding caption are processed by the retriever to identify the most relevant image-report pairs from the vector database; (3) The retrieved medical reports along with the query that consists of a question, image, and generated caption are fed into a VLM, which synthesizes the multimodal information to produce clinically accurate responses.

### Caption Generation

We develop an end-to-end caption generation module that produces detailed descriptions of medical images using a frozen medical VLM. Given an input image $I$, the VLM processes $I$ to generate a textual caption $C$ that identifies anatomical structures, abnormalities, and clinically relevant features. These generated captions are then utilized during the retrieval process to maximize the clinical relevance of the retrieved results by providing rich semantic information that complements visual features.

## Retrieval

### *Unimodal Retrieval*

The answer generation process in our retrieval-augmented approach depends both on the retrieved medical reports and the input image. Therefore, we employ unimodal and multimodal retrieval to identify and rank the most relevant reports. For unimodal retrieval, we implement two strategies:

- **C → M Retrieval**: Uses generated image captions solely to retrieve relevant medical reports from a vector database. The similarity is calculated in the text embedding space (1):

$$\text{sim}(C,M) = \frac{E_t(C)^\top E_t(M)}{\|E_t(C)\|\|E_t(M)\|}, \tag{1}$$

  where $E_t$ denotes the text encoder, $C$ represents the query caption, and $M$ corresponds to a medical report in the database.

- **I → I Retrieval**: Directly compares query images with database images using the similarity of visual features (2):

$$\text{sim}(I^q, I^d) = \frac{E_v(I^q)^\top E_v(I^d)}{\|E_v(I^q)\|\|E_v(I^d)\|}, \tag{2}$$

  where $E_v$ is the visual encoder, $I^q$ denotes the query image and $I^d$ represents the database images.

### *Multimodal Retrieval (MM-RAG-MedVQA)*

Our framework extends unimodal approaches through cross-modal retrieval using BioMedCLIP[31], a medical domain adaptation of CLIP[7]. Given a batch of $N$ paired image-text medical samples, BioMedCLIP trains an image encoder and a text encoder jointly to align their embeddings in a shared multimodal space. The objective is to maximize cosine similarity between embeddings of matched pairs while minimizing similarity for the remaining incorrect $N^2 - N$ pairings. This is achieved by optimizing the symmetric InfoNCE loss[37], which applies the cross-entropy loss to the normalized similarity scores (3):

$$\mathscr{L}_{I\to T} = -\frac{1}{2N}\left(\sum_{i=1}^{N}\log\frac{e^{\text{sim}(E_v(I_i),E_t(T_i))/\tau}}{\sum_{j=1}^{N}e^{\text{sim}(E_v(I_i),E_t(T_j))/\tau}} + \sum_{i=1}^{N}\log\frac{e^{\text{sim}(E_v(I_i),E_t(T_i))/\tau}}{\sum_{j=1}^{N}e^{\text{sim}(E_v(I_j),E_t(T_i))/\tau}}\right), \tag{3}$$

where:

- $E_v(I_i)$ and $E_t(T_i)$ denote the image and text (either caption or medical report) embeddings for the $i$-th pair,
- $\text{sim}(E_v(I_i), E_t(T_i))$ computes cosine similarity,
- $\tau$ represents the temperature parameter.

We further improve multimodal retrieval capability through additional fine-tuning with our proposed training strategy, which jointly optimizes alignment between medical images and both medical reports and generated captions. The input space comprises three modalities: images $\mathscr{I} = \{I_1,\ldots,I_N\}$, medical reports $\mathscr{M} = \{M_1,\ldots,M_N\}$, and generated captions $\mathscr{C} = \{C_1,\ldots,C_N\}$, where $N$ is the size of the training set. Each sample $X_n = (I_n, C_n, M_n)$ represents an aligned image-caption-report triplet.

We implement selective parameter optimization, unfreezing only the final transformer blocks of visual and text encoders while keeping earlier layers frozen (see Table 6). The composite loss function combines alignment objectives for both text modalities (4):

$$\mathscr{L}_{ours} = \frac{1}{2}\left(\mathscr{L}_{I\to C} + \mathscr{L}_{I\to M}\right), \tag{4}$$

where each term follows the formulation in (3).

Our framework employs an *I+C→I+M* retrieval setup, where '+' indicates that both image and text modalities are used, and joint embeddings are formed by concatenation $[\cdot;\cdot]$. Given a query $q$ consisting of an image $I^q$ and its caption $C$, we calculate its joint embedding as $E(q) = [E_v(I^q); E_t(C)] \in \mathbb{R}^{2d}$. Each entry in the database $d$ consists of an image $I^d$ and a medical report $M$, we calculate its joint embedding as $E(d) = [E_v(I^d); E_t(M)] \in \mathbb{R}^{2d}$. We then rank database entries by cosine similarity:

$$\text{sim}(q,d) = \frac{E(q)^\top E(d)}{\|E(q)\|\,\|E(d)\|}. \tag{5}$$

| Component | Layer Description |
|---|---|
| Global Parameters | logit_scale |
| Vision Token/Position | visual.trunk.cls_token<br>visual.trunk.pos_embed |
| Vision Transformer (Block 11) | Attention: norm1, qkv, proj<br>MLP: norm2, fc1, fc2 |
| Vision Normalization | visual.trunk.norm.weight/bias |
| Vision Projection | visual.head.proj.weight |
| Text Transformer (Layer 11) | Attention: query/key/value, output<br>Intermediate dense layer<br>Output dense layer + LayerNorm |
| Text Projection | text.proj.0.weight (first linear)<br>text.proj.2.weight (second linear) |

**Table 6.** Trainable layers in our proposed method.

***Vector Database***

The retrieval system operates on a vector database $\mathscr{D}$ containing pre-computed embeddings of image-report pairs. Each entry stores either: (1) visual embeddings for $I \rightarrow I$ retrieval, (2) textual embeddings for $C \rightarrow M$ retrieval, or (3) joint multimodal embeddings for $I$+$C \rightarrow I$+$M$ retrieval. The database supports efficient $k$-nearest-neighbor search, returning the most relevant items $\mathscr{R} = \{(I_1^d, M_1), \ldots, (I_k^d, M_k)\}$ for answer generation, where $k < N$ and $\mathscr{R} \subset \mathscr{D}$.

## Answer Generation

For a given query element $(I^q, C, U)$, where $U$ is the question, we augment top-$k$ retrieved medical reports $\mathscr{R}(\mathscr{M}) = \{M_1, \ldots, M_k\}$ to a text prompt. Hence, the final input to a frozen medical VLM consists of $I^q$, $C$, $U$ and $\mathscr{R}(\mathscr{M})$ to generate the answer. Algorithm 1 summarizes the complete workflow.

**Algorithm 1** MM-RAG-MedVQA

**Input:** Query Image $I^q$, Question $U$, Vector Database $\mathscr{D}$
**Output:** Answer *Ans* from VLM

**Caption Generation**
  Generate a caption $C$ for $I^q$
**Retrieval**
  Calculate embeddings of the query element $q(I^q, C)$
  Retrieve top-$k$ similar elements $\mathscr{R}$ from $\mathscr{D}$ w.r.t. $q$
  Concatenate the retrieved medical reports $\mathscr{R}(\mathscr{M})$
**Answer Generation**
  Feed $((I^q, C, U), \mathscr{R}(\mathscr{M}))$ as input to VLM to generate answer *Ans*

Our framework incorporates a carefully designed prompt template that explicitly structures the interaction between retrieved evidence and generation. The prompt enforces two key principles: (1) cross-modal verification between the input image, generated caption, and retrieved reports, with preference given to consensus observations, and (2) output standardization based on question type (closed-ended vs. open-ended). This structured prompting approach reduces hallucination risks while maintaining the VLM's generative flexibility, as evidenced by the high scores in our evaluations. The prompt design also includes safeguards against over-reliance on either visual or textual evidence alone, requiring agreement between both modalities when possible. The template for the prompt is presented in Figure 3.

# Data availability

All datasets are publicly available through PhysioNet (http://mimic.physionet.org).

**Prompt Template**

You are a medical vision-language model that answers questions based on images and relevant medical reports.
**Instructions:**

- Compare reports and image description to the image; focus on details that agree.
- Ignore conflicting content unless multiple independent reports strongly agree against the image.
- For closed-ended questions: answer 'Yes' or 'No' only. For multiple-choice: pick only from the given options. For open-ended questions: give a short, factual answer.

**Retrieved Reports:** {input context}
**Image Description:** {input caption}
**Question:** {input question}

**Figure 3.** The prompt used for answer generation.

## References


1. Lin, Z. *et al.* Medical visual question answering: A survey. *Artif. Intell. Medicine* **143**, 102611 (2023).
2. Zhang, X. *et al.* Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415* (2023).
3. Lau, J. J., Gayen, S., Ben Abacha, A. & Demner-Fushman, D. A dataset of clinically generated visual questions and answers about radiology images. *Sci. data* **5**, 1–10 (2018).
4. He, X., Zhang, Y., Mou, L., Xing, E. & Xie, P. Pathvqa: 30000+ questions for medical visual question answering (2020). 2003.10286.
5. Dong, W. *et al.* Generative models in medical visual question answering: A survey. *Appl. Sci.* **15**, 2983 (2025).
6. Sharma, D., Purushotham, S. & Reddy, C. K. Medfusenet: An attention-based multimodal deep learning model for visual question answering in the medical domain. *Sci. Reports* **11**, 19826 (2021).
7. Radford, A. *et al.* Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763 (PmLR, 2021).
8. Moor, M. *et al.* Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, 353–367 (PMLR, 2023).
9. Tu, T. *et al.* Towards generalist biomedical ai. *Nejm Ai* **1**, AIoa2300138 (2024).
10. Lewis, P. *et al.* Retrieval-augmented generation for knowledge-intensive nlp tasks. *Adv. Neural Inf. Process. Syst.* **33**, 9459–9474 (2020).
11. Sun, L., Zhao, J. J., Han, W. & Xiong, C. Fact-aware multimodal retrieval augmentation for accurate medical radiology report generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 643–655 (2025).
12. Xia, P. *et al.* Rule: Reliable multimodal rag for factuality in medical vision language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 1081–1093 (2024).
13. Shaaban, M. A., Saleem, T. J., Papineni, V. R. & Yaqub, M. Motor: Multimodal optimal transport via grounded retrieval in medical visual question answering. *arXiv preprint arXiv:2506.22900* (2025).
14. Xia, P. *et al.* Mmed-rag: Versatile multimodal rag system for medical vision language models. *arXiv preprint arXiv:2410.13085* (2024).
15. Yuan, Z. *et al.* Ramm: Retrieval-augmented biomedical visual question answering with multi-modal pre-training. MM '23, 547–556, DOI: 10.1145/3581783.3611830 (Association for Computing Machinery, New York, NY, USA, 2023).

16. Chen, W., Hu, H., Chen, X., Verga, P. & Cohen, W. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 5558–5570 (2022).

17. Chen, Z. *et al.* Heterorag: A heterogeneous retrieval-augmented generation framework for medical vision language tasks. *arXiv preprint arXiv:2508.12778* (2025).

18. Gao, Y. *et al.* Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997* (2023).

19. Liu, H. *et al.* A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253* (2024).

20. Wang, J., Ashraf, T., Han, Z., Laaksonen, J. & Anwer, R. M. Mira: A novel framework for fusing modalities in medical rag. *arXiv preprint arXiv:2507.07902* (2025).

21. Zhang, W. & Zhang, J. Hallucination mitigation for retrieval-augmented large language models: a review. *Mathematics* **13**, 856 (2025).

22. Johnson, A. E. W. *et al.* Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Sci. data.* **6** (2019-12-12).

23. Johnson, A. E. *et al.* Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042* (2019).

24. Ramesh, V., Chi, N. A. & Rajpurkar, P. Improving radiology report generation systems by removing hallucinated references to non-existent priors. In *Machine Learning for Health*, 456–473 (PMLR, 2022).

25. Bae, S. *et al.* Ehrxqa: A multi-modal question answering dataset for electronic health records with chest x-ray images. *Adv. Neural Inf. Process. Syst.* **36** (2024).

26. Hu, X. *et al.* Medical-diff-vqa: a large-scale medical dataset for difference visual question answering on chest x-ray images. *PhysioNet* **12**, 13 (2023).

27. Liu, Y. *et al.* G-eval: NLG evaluation using gpt-4 with better human alignment. In Bouamor, H., Pino, J. & Bali, K. (eds.) *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2511–2522, DOI: 10.18653/v1/2023.emnlp-main.153 (Association for Computational Linguistics, Singapore, 2023).

28. Zheng, L. *et al.* Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23 (Curran Associates Inc., Red Hook, NY, USA, 2023).

29. Chiang, C.-H. *et al.* Can large language models be an alternative to human evaluations? In *Annual Meeting of the Association for Computational Linguistics* (2023).

30. Chen, J. *et al.* Can NLI models verify QA systems' predictions? In Moens, M.-F., Huang, X., Specia, L. & Yih, S. W.-t. (eds.) *Findings of the Association for Computational Linguistics: EMNLP 2021*, 3841–3854, DOI: 10.18653/v1/2021.findings-emnlp.324 (Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021).

31. Zhang, S. *et al.* Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915* (2023).

32. Qdrant - vector database. Accessed in 2024.

33. Li, C. *et al.* Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Adv. Neural Inf. Process. Syst.* **36** (2024).

34. Bannur, S. *et al.* Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449* (2024).

35. Phan, L. N. *et al.* Scifive: a text-to-text transformer model for biomedical literature. *arXiv preprint arXiv:2106.03598* (2021).

36. Menick, J. *et al.* Gpt-4o mini: advancing cost-efficient intelligence. *Open AI: San Francisco, CA, USA* (2024).

37. Oord, A. v. d., Li, Y. & Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).


## Author contributions statement

M.S. conceived the initial idea. M.Z., A.K., A.A. and M.K. conceived the experiments. M.S. conducted the experiments. A.A., M.K. and M.Y. provided continuous supervision of the study. M.S., A.K. and M.Y. contributed to the final writing of the paper. All authors reviewed the manuscript.

# Additional information

## Competing interests

The authors declare no competing interests.


## Funding declaration

This research did not receive funding.