

MR-RAG: Multimodal Relevance-Aware Retrieval-Augmented Generation for Medical Visual Question Answering

Supplementary Material

A. Report Generation Performance

To further assess the generative capability of different methods, we conduct a report generation task where models are required to produce complete medical reports conditioned on a given image. To better adapt our approach—originally designed for the VQA setting—to this task where questions are not provided, we made the following adjustments: (1) the Multimodal Cooperative Retrieval (MCR) module computes document relevance using only image-image and image-text similarity; (2) in the Importance-aware Flow Augmentation (IFA) module, the attention path previously directed from **Reports**→**Question** is redirected to **Reports**→**Instruction**, where the instruction is phrased as: “You need to provide a medical report containing findings, based on the image and reference report(s).”

As shown in Table 5, we compare our method against two RAG-based baselines—FactMM-RAG and MMed-RAG—built on the LLaVA-Med-1.5 backbone, and evaluate using ROUGE-L and METEOR metrics across two datasets. Our method achieves competitive performance across both datasets and metrics, indicating its strength in generating more fluent and semantically relevant reports.

Table 5. Report generation performance compared against RAG-based baselines (FactMM-RAG and MMed-RAG), evaluated using ROUGE-L and METEOR.

Method	FairVLMed		IU-Xray	
	RG-L	MTR	RG-L	MTR
LLaVA-Med-1.5	0.098	0.179	0.177	0.278
+FactMM-RAG	<u>0.104</u>	0.170	0.201	<u>0.293</u>
+MMed-RAG	0.099	<u>0.180</u>	0.161	0.267
MR-RAG (Ours)	0.132	0.184	<u>0.200</u>	0.314

B. Efficiency Analysis

To assess the computational efficiency of our proposed MR-RAG, we report both inference latency and GPU memory usage compared with the LLaVA-Med-1.5. As shown in Table 6, MR-RAG introduces only marginal overhead. Specifically, the Multimodal Cooperative Retrieval (MCR) adds negligible cost to the retrieval stage, while the Importance-Aware Information Flow Augmentation (IFA) increases inference latency by merely **3.4%** (from 1.067 to 1.103 s/iteration) and memory usage by **4.9%** (23.9 vs. 25.1

Table 6. Efficiency comparison between the baseline RAG and our proposed MR-RAG.

Model	Latency (s/it)	Memory (GB)
LLaVA-Med-1.5	1.067	23.9
MR-RAG (Ours)	1.103(+3.4%)	25.1(+4.9%)

GB). These results demonstrate that our improvements in accuracy and robustness come with minimal computational trade-offs.

C. Instruction Prompt Templates

To unify the input format for the VQA task, we set a structured instruction prompt that explicitly defines the available information and clarifies the expected answer format.

Visual Question Answering Task

USER: You are provided with a chest X-ray image, a image-related question and $\{\text{top}k\}$ reference reports:
Reports: <formatted reports>
Image: <image>

It should be noted that the diagnostic information in the reference reports cannot be directly used as the basis for diagnosis, but should only be used for reference and comparison. Please answer the question based on the image and reports and start with a single word from two options: [yes, no]. Question: <question>

Medical Report Generation Task

USER: You are a helpful medical assistant. Your task is report generation. You are provided with a chest X-ray image and $\{\text{top}k\}$ reference reports:
Reports: <formatted reports>
Image: <image>
You need to provide a medical report containing findings, based on the image and reference report(s).

D. Datasets

Harvard-FairVLMed is a multimodal benchmark that focuses on fairness evaluation in medical vision-language tasks. It contains paired image and text data collected from diverse sources, primarily centered around fundus photography. Following the experimental protocol in RULE and MMed-RAG, we use 4,285 cases for testing.

Table 7. Full performance comparison of MR-RAG (Ours) against Med-Flamingo and MiniGPT-Med.

Model	Harvard-FairVLMed(%)				IU-Xray(%)				MIMIC-CXR(%)			
	Acc	Prec	Recall	F1	Acc	Prec	Recall	F1	Acc	Prec	Recall	F1
Med-Flamingo	<u>78.34</u>	<u>63.16</u>	67.78	<u>65.38</u>	63.98	56.77	59.27	57.99	<u>78.02</u>	<u>75.23</u>	74.43	<u>74.82</u>
MiniGPT-Med	65.80	57.23	59.13	58.16	<u>72.57</u>	<u>61.79</u>	<u>64.22</u>	<u>62.98</u>	72.21	71.92	<u>75.78</u>	73.79
MR-RAG (Ours)	84.47	67.61	<u>66.05</u>	66.78	88.13	84.50	86.01	85.19	79.35	79.80	77.56	78.61

Table 8. Full ablation study results showing the contributions of MCR and IFA module

Model Variant	Harvard-FairVLMed(%)				IU-Xray(%)				MIMIC-CXR(%)			
	Acc	Prec	Recall	F1	Acc	Prec	Recall	F1	Acc	Prec	Recall	F1
LLaVA-Med-1.5	77.35	59.50	62.56	60.34	85.57	81.42	85.94	83.00	71.92	71.02	72.01	71.16
+MCR	<u>83.19</u>	<u>65.79</u>	66.37	<u>66.06</u>	86.85	82.72	86.57	<u>84.21</u>	76.80	<u>77.05</u>	<u>73.27</u>	<u>74.29</u>
+MCR+IFA	84.47	67.61	<u>66.05</u>	66.78	88.12	84.49	<u>86.01</u>	85.19	79.35	79.80	77.56	78.61

IU-Xray is a publicly available dataset consisting of chest X-ray images paired with their corresponding diagnostic reports. It contains 7,470 image-report pairs. Following RULE and MMed-RAG, we use a split comprising 2,573 samples for testing.

MIMIC-CXR is one of the largest publicly available chest X-ray datasets, consisting of over 370,000 images and 227,000 associated radiology reports. Following RULE and MMed-RAG, we use 3,460 cases for testing.

E. Full Comparison with Open-source Medical LVLMS

To provide a more comprehensive comparison, Table 7 presents the complete evaluation results of our method (MR-RAG) and two open-source medical LVLMS, Med-Flamingo and MiniGPT-Med.

F. Full Module Ablation Results

We present the complete results of the module-level ablation study in Table 8. The inclusion of the MCR module substantially improves performance across all metrics, while integrating IFA yields further gains, confirming the complementary effectiveness of both components.

G. Generality of MR-RAG on Other Backbones

To verify the generality of MR-RAG, we apply it to MiniGPT-Med [1]. Thanks to its modular design, MR-RAG can adaptively modulate attention in key layers identified via a small-scale blocking experiment, allowing seamless integration with different LVLMS backbones.

Table 9 shows consistent improvements across three datasets. Accuracy increases from 69.63% to 74.48%, F1

Table 9. Effect of applying MR-RAG to MiniGPT-Med across three datasets. The improvements across all metrics demonstrate the generality and transferability of our method.

Method	Acc	Prec	Rec	F1
MiniGPT-Med	69.63	63.29	65.98	64.60
+ MR-RAG	74.48	68.39	69.45	68.91

from 64.60% to 68.91%, demonstrating that MR-RAG’s retrieval-augmented generation and relevance-aware attention are effective beyond the original backbone.

Ethical Statement

Our work focuses on improving medical vision-language models to assist clinicians in interpreting medical images. While these models have the potential to enhance clinical workflows and support medical decision-making, we acknowledge that they are not intended to replace professional medical judgment. We emphasize that the models should be used as assistive tools and that any outputs must be verified by qualified healthcare professionals before clinical use. We encourage further research into model interpretability and bias reduction to improve the reliability and fairness of medical LVLMS systems.