Reducing Hallucinations of Medical Multimodal Large Language Models with Visual Retrieval-Augmented Generation

Yun-Wei Chu[1,2‡], **Kai Zhang**[1,3‡], **Christopher Malon**[1], **Martin Renqiang Min**[1]

[1]NEC Laboratories America, [2]Purdue University, [3]Lehigh University
chu198@purdue.edu, kaz321@lehigh.edu, malon@nec-labs.com, renqiang@nec-labs.com

Abstract

Multimodal Large Language Models (MLLMs) have shown impressive performance in vision and text tasks. However, hallucination remains a major challenge, especially in fields like healthcare where details are critical. In this work, we show how MLLMs may be enhanced to support Visual RAG (V-RAG), a retrieval-augmented generation framework that incorporates both text and visual data from retrieved images. On the MIMIC-CXR chest X-ray report generation and Multicare medical image caption generation datasets, we show that Visual RAG improves the accuracy of entity probing, which asks whether a medical entities is grounded by an image. We show that the improvements extend both to frequent and rare entities, the latter of which may have less positive training data. Downstream, we apply V-RAG with entity probing to correct hallucinations and generate more clinically accurate X-ray reports, obtaining a higher RadGraph-F1 score.

1 Introduction

Recent advances in Multimodal Large Language Models (MLLMs) (OpenAI 2023; Liu et al. 2023a) have demonstrated impressive capabilities in complex vision-and-text tasks, showing significant potential in specialized domains. In healthcare, the development of Medical MLLMs (Med-MLLMs) (Li et al. 2023a; Wu et al. 2023) can support clinical decision-making processes, with the potential to enhance physician efficiency and improve patient health outcomes. However, numerous studies have demonstrated that MLLMs are prone to hallucination (Li et al. 2023b; Bai et al. 2024; Huang et al. 2024). The hallucination tendency of MLLM's has been demonstrated on Med-MLLM's as well (Wu, Kim, and Wu 2024). This is particularly concerning in the healthcare scenario, as depicted in Figure 1, where even a few wrong tokens in text can lead to significant misinterpretations, affecting medical diagnoses, treatment plans, and patient outcomes (Pal and Sankarasubbu 2024).

Retrieval-Augmented Generation (RAG) (Lewis et al. 2020) has become a prominent approach to mitigate the hallucination problem in Large Language Models (LLMs)

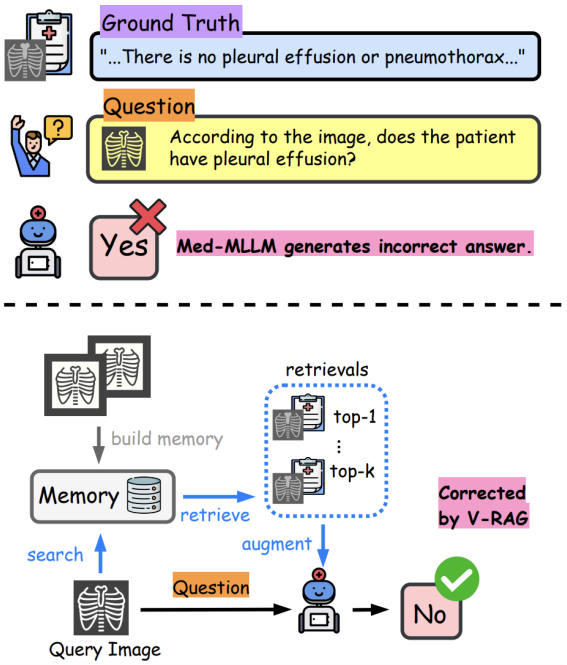

Figure 1: (Up) Hallucination issue of Med-MLLM. (Down) Framework of V-RAG to improve Med-MLLM.

by grounding text generation in retrieved knowledge relevant to a given query. Besides grounding, RAG potentially supplements the knowledge in a model's parameters with knowledge present in a corpus, enabling open book question answering to exceed closed book performance. Several prior works (Sarto et al. 2024; Liu et al. 2024; Zhou et al. 2024) have explored text-based RAG in MLLMs. This approach assumes that using text documents associated with images similar to the query image can effectively augment the model, treating the retrieved images as perfectly interchangeable with the query image. However, this assumption is not always accurate. In this work, we study **Visual-RAG (V-RAG)**, which considers not only the associated text from retrieved similar images but also the similar images themselves to provide more accurate responses to the given instruction. By incorporating both modalities, V-RAG allows the model to determine what is truly important from the retrieved content, enhancing its ability to deliver more contex-

‡Work done as an intern at NEC Laboratories America.

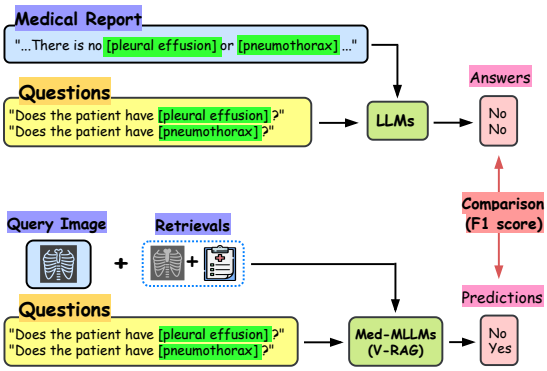

Figure 2: Entity probing asks entity-based questions to an MLLM and compares predictions against answers grounded in an LLM's interpretation of a reference caption.

tually relevant answers, as illustrated in Figure 1.

With certain multi-image-trained Med-MLLMs, we see that V-RAG improves a detailed understanding of an image beyond what is possible with text-based RAG techniques. We demonstrate this through **entity probing**. Entity probing presents an image to an MLLM and asks yes/no questions about disease entities, and compares predictions against answers grounded in an LLM's interpretation of a reference report or caption (Figure 2). Entity probing gives us a clinical perspective on text generations across medical domains which is not captured by natural language generation metrics such as ROUGE, while avoiding sensitivity to entity phrasing. We show that V-RAG, as an inference technique applied to carefully selected Med-MLLMs trained on multi-image datasets, enhances understanding more effectively than original Med-MLLMs and previous text-based RAG systems.

To improve the model's multimodal understanding when presented with rich retrievals, we design a general fine-tuning technique to boost Med-MLLM capabilities in V-RAG. This approach strengthens image-text comprehension and enables effective learning from similar resources retrieved during multimodal queries. It benefits not only Med-MLLMs trained on multi-image dataset but also single-image-trained models to leverage multi-image inputs in V-RAG, thereby improving performance. This frees researchers from relying on specific pre-trained models that may not be aligned with their task in order to use V-RAG, allowing V-RAG to be applied to any model and dataset of interest. Our key contributions are summarized as follows:

- We analyze hallucinations in MLLMs on chest X-ray report generation and medical image captioning datasets through entity probing, showing that V-RAG mitigates hallucinations more effectively than baseline RAG techniques. These benefits extend to both frequent and rare entities.
- To enhance Med-MLLMs' multimodal comprehension with V-RAG, we introduce general image-text fine-tuning tasks to boost model performance and improve their understanding when multimodal retrievals are presented. These tasks enable an MLLM originally trained with single images to become capable of V-RAG using multiple retrieved images.
- We show that entity probing with V-RAG can be used to revise chest X-ray reports to contain fewer hallucinations and have better detailed accuracy, as measured by RadGraph-F1 score.

2 Related Work

Medical Multimodal Large Language Models. Substantial advancements have been made in adapting MLLMs to medical imaging (Zhang et al. 2023b; Wu et al. 2023; Moor et al. 2023; Lee et al. 2023). The primary focus has been on training these models for radiology tasks using medical images (like X-rays, MRIs, and CT scans) along with their textual descriptions/reports. Li et al. (2023a) used GPT-4 to generate instruction-following data for fine-tuning, improving MLLMs' conversational ability for open-ended biomedical image inquiries. Chen et al. (2024) developed a foundation model for chest X-Ray interpretation with an image-text bridger to align modalities. However, we found that these medical multimodal foundation models still suffer from hallucinations. We aim to mitigate this issue in Med-MLLMs through a visual-based Retrieval-Augmented Generation (RAG) approach, enabling these models to generate factually accurate answers.

Retrieval-Augmented Generation (RAG). RAG (Lewis et al. 2020) mitigates hallucination in LLMs by retrieving and integrating domain-specific knowledge from external databases, enhancing text generation with accurate, aligned information and effectively addressing this challenge (Guu et al. 2020; Siriwardhana et al. 2022; Shahul et al. 2023). Despite RAG's popularity, very few studies have applied RAG to MLLMs. Prior studies primarily enhance image captioning by reranking labels of retrieved images (Liu et al. 2024; Qu et al. 2024a) or directly incorporating texts from these images into prompts to improve generation (Liu et al. 2023c; Sarto et al. 2024; Zhou et al. 2024). In healthcare, researchers have developed domain-specific retrieval pipelines (Sun et al. 2024) and explored the optimal number of retrievals (Xia et al. 2024) to ensure the factuality of Med-MLLMs. All these previous works retrieve similar images based on the query image but consider only the text/label associated with the retrieved images. Thus these methods assume that the retrieved images are perfectly interchangeable with the query image, which is not always the case.

A more effective approach might involve comparing the query image with retrieved images and their reports, allowing the model to identify what is truly relevant for generation. This is the "V-RAG" method of our paper. Qu et al. (2024b) attempted a similar approach with "Coarse (I+T)," though it performed worse than using only associated texts ("Coarse (T)" in their Table 6), which they noted was likely due to limited multi-image reasoning in the MLLMs they considered. We address this by analyzing MLLMs trained for multi-image reasoning, and also by introducing an architecture and fine-tuning method to make single-image-trained MLLMs "V-RAG-capable," enabling them to benefit from this approach.

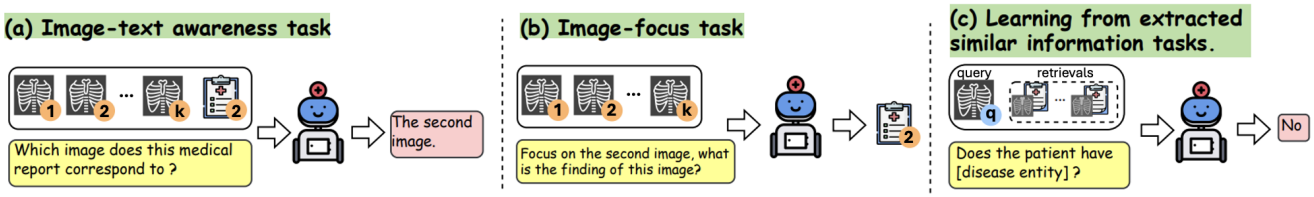

Figure 3: Fine-tuning tasks to make Med-MLLM V-RAG-capable by (a) improving image-and-text association abilities, (b) focusing on specific images, and (c) making decisions using extracted similar data.

3 V-RAG with Existing multi-image-trained Med-MLLMs

Figure 1 illustrates the V-RAG framework. This section details each component and explains how we enhance model performance during V-RAG.

3.1 Multimodal Retrieval

We aim to retrieve images and corresponding textual descriptions that match the features of target medical images. These references, rich in visual and textual medical details, guide response generation for the medical image. To extract embeddings, we employ BiomedCLIP (Zhang et al. 2023a), which provides robust representations across a diverse range of biomedical image types. For a given medical image X_{img}, we extract its image embedding $\mathcal{E}_{img} \in \mathbb{R}^d$, with d representing the dimension (i.e., 512 for BiomedCLIP), and store it in memory $\mathcal{M}$ for retrieval.

To facilitate efficient search operations during the inference phase, we construct the memory $\mathcal{M}$ using FAISS (Douze et al. 2024), a vector storage and retrieval system that utilizes GPU computation. Instead of exact kNN search, we employ an approximate kNN search using the Hierarchical Navigable Small World (HNSW) algorithm (Malkov and Yashunin 2016) to identify the top-k nearest neighbors, effectively retrieving the images in $\mathcal{M}$ most similar to a given query image.

3.2 Inference with V-RAG

In the inference stage, we first encode the query image X_q to obtain its corresponding image embedding. We then retrieve the top-k images in $\mathcal{M}$; the retrieved set of similar images and their reports are represented as $(I_1, ...I_k)$ and $(R_1, ..., R_k)$. We then use the retrievals to guide the generation of Med-MLLM for the query image by appending each reference before the question, following this prompt guidance: "`...This is the i-th similar image and its report for your reference.` $[\text{Reference}]_i$`... Answer the question with only the word yes or no. Do not provide explanations. According to the last query image and the reference images and reports,` $[\text{Question}]$ $[\text{Query Image}]$", where $[\text{References}]_i$ is structured as $[(I_i,\ R_i)]$.

3.3 Enhancing Med-MLLMs for V-RAG

Some MLLMs may lack the training to distinguish information from multiple images. To address this, we introduce three fine-tuning tasks to enhance image-text association in the V-RAG process. Given a dataset of images paired with captions or reports, we define the original dataset as $S = (\text{img}_i, \text{P}_i, \text{A}_i)|_{i=1}^N$, where img_i denotes the i-th image, P_i and A_i represent the prompt and the answer, respectively, and N is the total number of samples. We then construct fine-tuning tasks on this dataset with our designed objectives as follows.

Image-text awareness task. We aim to enhance Med-MLLM's image-and-text association ability by training the model to identify the relevant image corresponding to provided text from multiple images. To achieve this, we construct a multi-image dataset, $M_{position}$, from dataset S, to ask the model to identify the position of the image related to the given text, as depicted in Figure 3(a). First, we randomly select K images (where K ranges from 1 to 5 in our case) and form the image collection $(\text{img}_{i_1}, ..., \text{img}_{i_K})$. Next, we choose an integer j from $[1, K]$ and retrieve the textual document R_{i_j}, corresponding to img_{i_j}. We then collect $M_{position}$ using $\{(\text{img}_{i_1}, \text{img}_{i_2}, ..., \text{img}_{i_K}, \text{P}'_{i_j}, \text{A}'_{i_j})\}$. Here, P'_{i_j} is a newly formulated prompt designed to ask a position-based question in addition to the original question P_{i_j}, associating A_{i_j} with the provided images. For example, "`What image from 1 to` K `does this` A_{i_j} `correspond to?` P_{i_j}". A'_{i_j} is the answer indicating the position of img_{i_j} among the provided images, for example, "`The` j`-th image.`"

Image-focus task. In this task (Figure 3(b)), we aim to direct Med-MLLM to focus on one specific image from a set of multiple images and subsequently perform text generation based on that image, thereby improving performance by minimizing distractions from other visual inputs. To achieve this, we create another dataset, M_{focus}, also from image dataset S. We start by randomly selecting K images from S to form the collection $(\text{img}_{i_1}, ..., \text{img}_{i_K})$, and then choose an integer j from $[1, K]$. We then collect $\{(\text{img}_{i_1}, ..., \text{img}_{i_K}, \text{P}''_{i_j}, \text{A}_{i_j})\}$ to form M_{focus}, where P''_{i_j} is a new prompt designed to help the model focus on our specified image, img_{i_j}, and pose the original question P_{i_j} for that image. For example in Figure 3(b), the new prompt P''_{i_j} is "`Focus on the` j`-th image,` P_{i_j}`.`", where P_{i_j} is the original prompt that asks for a finding/report to be generated from a given image.

Strategies to make easier learning tasks. Various conditions may be applied to the random selection of images for

both image-text awareness and image-focus tasks. For example, when the image dataset S consists of images $\texttt{img}_i$ with radiology reports $\texttt{A}_i$, we require that the selected report $\texttt{A}_{i_j}$ for the focus image contains at least one CheXpert (Irvin et al. 2019) label that is distinct from those in the other reports $\{A_{i_m}\}|_{m=1,m\neq j}^{K}$. This strategy simplifies the learning task by ensuring that there are no alternative images to which the report could apply equally well. For easier and more diverse datasets, such a strategy may not be necessary.

Learning from extracted similar information task. We aim to assist Med-MLLM in decision-making by using extracted similar information during V-RAG. To do so, we simulate the V-RAG scenario and construct a multi-image dataset, M_{vrag}. Given a query image $\texttt{img}_q$ in the validation set, we search for the top-K similar images ($\texttt{img}_{q_1}$, ... , $\texttt{img}_{q_K}$) from memory $\mathcal{M}$, pairing them with their corresponding documents ($\texttt{A}_{q_1}$, ..., $\texttt{A}_{q_K}$). We then conduct M_{vrag} using $\{(\texttt{img}_{q_1}, \texttt{A}_{q_1}, ..., \texttt{img}_{q_K}, \texttt{A}_{q_K}), \texttt{img}_q, \texttt{P}_q^{'''}, \texttt{A}_q\}$. Here, $\texttt{A}_q$ is the answer for query image and $\texttt{P}_q^{'''}$ is a new prompt designed to supply related information alongside the original question $\texttt{P}_q$. Taking disease entity probing as example (in Figure 3(c)), $\texttt{P}_q^{'''}$ can be "`Based on the query image, and the similar images and their reports:` $(\texttt{img}_{q_1}, \texttt{A}_{q_1}, ..., \texttt{img}_{q_K}, \texttt{A}_{q_K})$, $\texttt{P}_q$," and $\texttt{P}_q$ is "`Does the patient have [disease entity]?`"

4 Experiment

4.1 Experimental Setups

We selected RadFM (Wu et al. 2023), an existing multi-image-trained Med-MLLM, as our base model to evaluate the effectiveness of V-RAG and our proposed fine-tuning tasks on multi-image-trained models. To assess the capability of making single-image-trained MLLMs V-RAG capable, we utilized LLaVA (Liu et al. 2023b) as the backbone model. We employed LoRA (Hu et al. 2021) to fine-tune both LLaVA and RadFM on our designed tasks, applying a learning rate of 5e-5 for all fine-tuning tasks.

4.2 Baselines

We compare our method with the original Med-MLLM, RadFM, which does not include retrievals, with other baselines that do. RAT (Sarto et al. 2024) and Img2Loc (Zhou et al. 2024) are identical methods which incorporate text associated with retrieved similar images into the prompt. RAR (Liu et al. 2024) also incorporates the text associated with retrieved similar images, but it re-ranks those texts using the MLLM before generation. We set $k = 5$ as the number of retrievals for every RAG-based method.

4.3 Datasets and Evaluation Metrics

Entity Probing We utilize two medical vision-language datasets: MIMIC-CXR (Johnson et al. 2019), containing chest X-ray images for radiology, and MultiCaRe (Nievas Offidani and Delrieux 2024), offering a variety of images across medical specialties. We follow the official data split for MIMIC-CXR and randomly split MultiCaRe into train, validation, and test sets with a ratio of 8:1:1. To construct VQA pairs for disease entity probing, we employ a biomedical named entity recognition (NER) model [1] (Zhang et al. 2021) to extract all disease entities from the dataset's reports. We input these reports into LLMs (in our case, Llama-2 7B) to create closed-ended QA data with yes or no answers. For example, we ask "`Does the patient have [disease entity] based on the report: [Report]?`", with answers formatted as `Yes/No`, simplifying error analysis. The use of LLMs allows for interpreting complex semantic structures within the text to accurately deduce potential answers. For instance, given the `[Report]`: "An upper GI series on post-operative day 5 showing the duodenum ruling out stenosis." and `[disease entity]`: "`stenosis`", the LLM correctly answers "`No.`" By sampling segments from a medical report, we generate a sequence of concise, closed-ended questions paired with LLM-generated answers. The VQA dataset is then formed by associating these disease probing QA pairs with the original medical images.

For example, in MIMIC-CXR, we exclude entities in the "INDICATION" section of the report, as these reflect patient history or the reason for conducting the evaluation rather than X-ray findings. Across both datasets, we found that less frequent entities are often already covered by more frequent ones (e.g., "`right lower lobe atelectasis`" as a particular kind of "`atelectasis`"). Therefore, we map each entity to its shortest terminal subphrase occurring as an entity in the training set, to reduce redundancy and clarify entity frequency. For each test set of MIMIC-CXR and MultiCaRe, we parse 9,411 and 21,653 VQA pairs, respectively, with 385 and 10,434 distinct entities. We use Precision, Recall, and F1 Score as the primary metrics to evaluate answer correctness in disease entity probing.

Report generation We apply disease entity probing with V-RAG to mitigate hallucinations in generated text through a rewrite strategy. After a Med-MLLM generates an initial report of findings for an X-ray, the NER model extracts all disease entities from the generated report and from the reports of the k most similar images. For each entity, the query image is probed using the Med-MLLM with V-RAG.

The originally generated report and entity probing results are input to a text-only LLM (Llama 3.1 70B chat), with the prompt: `Consider the following chest X-ray report from a junior radiologist: -----begin report----- [REPORT] -----end report----- A senior radiologist has inspected the X-ray image and answered the following questions: -----begin questions---- [QUESTIONS AND ANSWERS] -----end questions----- Please rewrite the junior radiologist's report to reflect the senior radiologist's answers.` We measure RadGraph-F1 scores (Delbrouck et al. 2024) of the findings of the original and revised reports.

[1] Stanza i2b2: https://stanfordnlp.github.io/stanza/biomed.html

Method	MIMIC-CXR			MultiCaRe		
	Precision	Recall	F1	Precision	Recall	F1
RadFM	**0.921**	0.206	0.381	**0.972**	0.290	0.432
+ RAR	0.871	0.397	0.535	0.962	0.536	0.664
+ RAT / Img2Loc	0.760	**0.943**	0.711	0.961	0.915	0.901
+ V-RAG	0.770	0.920	0.721	0.960	0.952	0.920
+ V-RAG (fine-tuned)	0.790	0.921	**0.751**	0.961	**0.999**	**0.940**

Table 1: Overall entity probing performance for different methods across two datasets. V-RAG's superiority shows the value of using complete retrieval information, both text and images. The improved performance of our fine-tuned V-RAG demonstrates enhanced image-text association abilities in Med-MLLM during V-RAG.

Retrieval Modality		MIMIC-CXR	MultiCaRe
Image	Text		
		0.381	0.432
✓		0.705	0.735
	✓	0.711	0.901
✓	✓	**0.721**	**0.920**

Table 2: Ablation study on RAG with different retrieval modalities. Improved F1 scores across both datasets shows the importance of integrating both image and text from retrievals to make informed decisions in V-RAG.

Fine-tuning Tasks			MIMIC-CXR	MultiCaRe
Position	Focus	V-RAG		
			0.721	0.920
		✓	0.729	0.933
	✓	✓	0.741	0.935
✓		✓	0.748	0.937
✓	✓	✓	**0.751**	**0.940**

Table 3: Ablation study of V-RAG using RadFM trained on various fine-tuning tasks. The F1 gains achieved through our three proposed tasks show improved image-text association abilities for existing multi-image-trained Med-MLLMs.

5 Evaluation Results

5.1 Overall performance for existing multi-image-trained Med-MLLMs

We first evaluate V-RAG's performance for existing Med-MLLM that originally trained on multi-image datasets. Table 1 shows entity probing results comparing our method to baselines. Across both datasets, V-RAG outperforms text-only RAG baselines in F1 scores. This improves the model's ability to extract relevant information for decision-making. Furthermore, with our proposed fine-tuning tasks, V-RAG (fine-tuned) achieves superior F1 scores over both baselines and the un-fine-tuned version. This shows that we have significantly enhanced Med-MLLM's capabilities by equipping it with robust image-text association skills.

5.2 Ablation study

We now conduct ablation studies to better understand our proposed method across various configurations.

Multimodal retrieval. Table 2 shows the F1 scores of RAG (top-5) across different retrieval modalities. We observe that providing only similar images without text makes it challenging for Med-MLLM to extract entity information from visuals, though it offers marginal improvements over Med-MLLM without RAG. Adding text for similar images significantly enhances performance, highlighting the rich information provided by texts in entity probing. By integrating both modalities, V-RAG effectively links retrieved texts and images, enabling more comprehensive decision-making and achieving the best performance. This underscores the importance of multimodal retrieval in V-RAG, rather than relying solely on text as most existing MLLM baselines.

Fine-tuning tasks for V-RAG. To enhance V-RAG's performance, we proposed three fine-tuning tasks for Med-MLLM, each with 6,000 instances. In Table 3, we examine how different combinations of these tasks impact performance. Initially, using only the M_{vrag} dataset, we enable Med-MLLM to learn from extracted similar information, yielding performance gains that enhance the model's understanding of downstream V-RAG tasks. Adding the image-text association tasks $M_{position}$ and M_{focus} provides further gains, with $M_{position}$ offering more benefits due to the complexity of M_{focus}, which involves generating a full medical report and is more challenging to learn with limited data.

5.3 Analysis of entities across frequency levels

In addition to analyzing the overall entities in the test set, we conducted an analysis to see how they differ in appearance. We categorized the entities from the test set into the most frequent 50 and the less frequent ones, analyzing their performance separately. Rare entities were almost exclusively found in positive contexts, which created a label imbalance. To address this, we balanced the test sets for rare entities by adding additional negative probing questions for each entity until the number of positive examples equaled the number of negative examples. Negative examples were paired with a randomly chosen image, and we verified using Llama-2 that the associated report did not suggest the presence of the entity. We tested 1,000 samples for both frequent and rare entities across two datasets. Figure 4 shows the F1 scores for each test set. Our V-RAG method outperforms both the original method and the RAG baselines in both settings. The improvement of V-RAG over other methods in the rare entity setting demonstrates the practical utility of our approach, emphasizing its effectiveness in utilizing information from multiple modalities to answer queries that neither the original model nor text-based RAG methods could address.

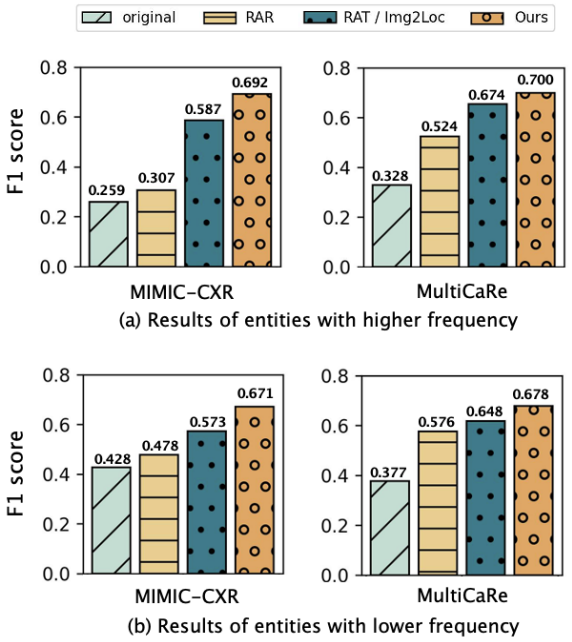

Figure 4: F1 performance of each method on disease entities with different frequencies. The superior performance of our method, particularly in probing rare entities, demonstrates its effectiveness and applicability in real-world scenarios.

Model	MIMIC-CXR		
	Precision	Recall	F1
LLaVA$_S$	**0.953**	0.475	0.604
LLaVA$_{\{M_{vrag}\}}$	0.914	0.867	0.852
LLaVA$_{\{M_{focus}+M_{vrag}\}}$	0.908	0.903	0.859
LLaVA$_{\{M_{position}+M_{vrag}\}}$	0.908	0.910	0.862
LLaVA$_{\{M_{position}+M_{focus}+M_{vrag}\}}$	0.897	**0.944**	**0.870**

Table 4: Entity probing results for single-image-trained MLLM and MLLM enhanced with our proposed fine-tuning tasks. The superiority indicates that our tasks effectively make a MLLM V-RAG-capable.

5.4 Can we make a single-image-trained MLLM V-RAG-capable?

After observing the performance gains of V-RAG and our fine-tuning methods on multi-image pre-trained Med-MLLMs, we now explore whether single-image-trained MLLMs can also be enabled to perform V-RAG. We extract all single image-text pairs from the MIMIC-CXR training set to create the single-image dataset S, resulting in 100,098 samples. We then fine-tune LLaVA-v1.5-7B with Vicuna backbone (Liu et al. 2023a) on S using LoRA for one epoch, resulting in a single-image Med-MLLM denoted as LLaVA$_S$. From the single-image dataset S, we extract 10k samples for each fine-tuning task in Section 3, creating the multi-image datasets $M_{position}$, M_{focus}, and M_{vrag}. We then fine-tune LLaVA$_S$ on these tasks, producing the model LLaVA$_{\{task\}}$.

To evaluate our idea, we conducted entity probing on the MIMIC-CXR test set. For the single-image model LLaVA$_S$, we input a single test image to probe for a disease entity.

Reports	RadGraph F1		
	Simple	Partial	Complete
Original	.163	.145	.102
Revised	.194	.172	.118

Table 5: Revising reports with V-RAG entity probing results.

For the multi-image model LLaVA$_{\{task\}}$, we implemented V-RAG to assess its performance and determine if it can be effectively V-RAG capable with our designed tasks. We set the context length of LLaVA to be 4096 and consider the top-3 retrievals for LLaVA$_{\{task\}}$ when performing V-RAG.

Table 4 shows the entity probing performance of single-image-trained MLLM and MLLM with multi-image capabilities resulting from our proposed fine-tuning tasks. For the single-image model LLaVA$_S$, we input a single test image and tasked the model with probing for a disease entity based on the given image. For the multi-image model LLaVA$_{\{task\}}$, we perform V-RAG to assess its performance. We set the context length of LLaVA to be 4096 and consider the top-3 retrievals for LLaVA$_{\{task\}}$ when performing V-RAG. Results demonstrate that, with the support of our designed fine-tuning tasks, we enable the single-image-trained MLLM to effectively perform V-RAG.

5.5 Improving generated reports

We have shown that disease entity probing provides a valuable clinical perspective on model outputs. However, since entity probing is typically not the final task for an MLLM, it is essential to demonstrate the utility of V-RAG in report generation. We find that our strategy using Llama 3.1 70B Chat to rewrite the generated reports using the V-RAG entity probing results yields 19% relative improvements in the simple and partial RadGraph-F1, compared to the original findings, as shown in Table 5. These results highlight the practical benefits of V-RAG-enhanced entity probing, demonstrating its value not only in probing accuracy but also in improving the accuracy of generated medical reports.

6 Conclusion

When faced with a long report generation task, Medical Multimodal Large Language Models may exhibit biases and hallucinate details. We have introduced an entity probing method to examine these details, and shown that V-RAG improves entity probing accuracy for both frequent and rare entities. The lack of multi-image support in mainstream models has been a barrier to the adoption of V-RAG, leading almost all prior work to work only with the text corresponding to similar images. Our special image-and-text fine-tuning tasks pave the way for multi-image-trained and single-image-trained models to become capable or more powerful at V-RAG, and we have shown that the use of both retrieved text and retrieved images benefits entity probing performance. Downstream, revision using entity probing with V-RAG can increase a report's accuracy on clinical details, improving the RadGraph-F1 score of a generated report. Our research contributes towards more medically trustworthy MLLMs for healthcare applications.

References

Bai, Z.; Wang, P.; Xiao, T.; He, T.; Han, Z.; Zhang, Z.; and Shou, M. Z. 2024. Hallucination of Multimodal Large Language Models: A Survey. *ArXiv*, abs/2404.18930.

Chen, Z.; Varma, M.; Delbrouck, J.-B.; Paschali, M.; Blankemeier, L.; Veen, D. V.; Valanarasu, J. M. J.; Youssef, A.; Cohen, J. P.; Reis, E. P.; Tsai, E. B.; Johnston, A.; Olsen, C.; Abraham, T. M.; Gatidis, S.; Chaudhari, A. S.; and Langlotz, C. P. 2024. CheXagent: Towards a Foundation Model for Chest X-Ray Interpretation. *ArXiv*, abs/2401.12208.

Delbrouck, J.-B.; Chambon, P.; Chen, Z.; Varma, M.; Johnston, A.; Blankemeier, L.; Van Veen, D.; Bui, T.; Truong, S.; and Langlotz, C. 2024. RadGraph-XL: A Large-Scale Expert-Annotated Dataset for Entity and Relation Extraction from Radiology Reports. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics ACL 2024*, 12902–12915. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics.

Douze, M.; Guzhva, A.; Deng, C.; Johnson, J.; Szilvasy, G.; Mazaré, P.-E.; Lomeli, M.; Hosseini, L.; and Jégou, H. 2024. The Faiss library.

Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; and Chang, M.-W. 2020. REALM: Retrieval-Augmented Language Model Pre-Training. *ArXiv*, abs/2002.08909.

Hu, J. E.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *ArXiv*, abs/2106.09685.

Huang, W.; Liu, H.; Guo, M.; and Gong, N. Z. 2024. Visual Hallucinations of Multi-modal Large Language Models. *ArXiv*, abs/2402.14683.

Irvin, J. A.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R. L.; Shpanskaya, K. S.; Seekins, J.; Mong, D. A.; Halabi, S. S.; Sandberg, J. K.; Jones, R.; Larson, D. B.; Langlotz, C.; Patel, B. N.; Lungren, M. P.; and Ng, A. 2019. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. In *AAAI Conference on Artificial Intelligence*.

Johnson, A. E. W.; Pollard, T. J.; Berkowitz, S. J.; Greenbaum, N. R.; Lungren, M. P.; ying Deng, C.; Mark, R. G.; and Horng, S. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6.

Lee, S.; Kim, W. J.; Chang, J.; and Ye, J.-C. 2023. LLM-CXR: Instruction-Finetuned LLM for CXR Image Understanding and Generation.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Kuttler, H.; Lewis, M.; tau Yih, W.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *ArXiv*, abs/2005.11401.

Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; and Gao, J. 2023a. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. *ArXiv*, abs/2306.00890.

Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W. X.; and rong Wen, J. 2023b. Evaluating Object Hallucination in Large Vision-Language Models. In *Conference on Empirical Methods in Natural Language Processing*.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2023a. Improved Baselines with Visual Instruction Tuning. *ArXiv*, abs/2310.03744.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023b. Visual Instruction Tuning. *ArXiv*, abs/2304.08485.

Liu, H.; Son, K.; Yang, J.; Liu, C.; Gao, J.; Lee, Y. J.; and Li, C. 2023c. Learning Customized Visual Models with Retrieval-Augmented Knowledge. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15148–15158.

Liu, Z.; Sun, Z.; Zang, Y.; Li, W.; Zhang, P.; wen Dong, X.; Xiong, Y.; Lin, D.; and Wang, J. 2024. RAR: Retrieving And Ranking Augmented MLLMs for Visual Recognition. *ArXiv*, abs/2403.13805.

Malkov, Y.; and Yashunin, D. A. 2016. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42: 824–836.

Moor, M.; Huang, Q.; Wu, S.; Yasunaga, M.; Zakka, C.; Dalmia, Y.; Reis, E. P.; Rajpurkar, P.; and Leskovec, J. 2023. Med-Flamingo: a Multimodal Medical Few-shot Learner. *ArXiv*, abs/2307.15189.

Nievas Offidani, M. A.; and Delrieux, C. A. 2024. Dataset of clinical cases, images, image labels and captions from open access case reports from PubMed Central (1990–2023). *Data in Brief*, 52.

OpenAI. 2023. GPT-4 Technical Report.

Pal, A.; and Sankarasubbu, M. 2024. Gemini Goes to Med School: Exploring the Capabilities of Multimodal Large Language Models on Medical Challenge Problems & Hallucinations. *ArXiv*, abs/2402.07023.

Qu, X.; Chen, Q.; Wei, W.; Sun, J.; and Dong, J. 2024a. Alleviating Hallucination in Large Vision-Language Models with Active Retrieval Augmentation.

Qu, X.; Chen, Q.; Wei, W.; Sun, J.; and Dong, J. 2024b. Alleviating Hallucination in Large Vision-Language Models with Active Retrieval Augmentation. arXiv:2408.00555.

Sarto, S.; Cornia, M.; Baraldi, L.; Nicolosi, A.; and Cucchiara, R. 2024. Towards Retrieval-Augmented Architectures for Image Captioning. *ACM Transactions on Multimedia Computing, Communications and Applications*.

Shahul, E.; James, J.; Anke, L. E.; and Schockaert, S. 2023. RAGAs: Automated Evaluation of Retrieval Augmented Generation. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Siriwardhana, S.; Weerasekera, R.; Wen, E.; Kaluarachchi, T.; Rana, R. K.; and Nanayakkara, S. 2022. Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering. *Transactions of the Association for Computational Linguistics*, 11: 1–17.

Sun, L.; Zhao, J.; Han, M.; and Xiong, C. 2024. Fact-Aware Multimodal Retrieval Augmentation for Accurate Medical Radiology Report Generation.

Wu, C.; Zhang, X.; Zhang, Y.; Wang, Y.; and Xie, W. 2023. Towards Generalist Foundation Model for Radiology. *ArXiv*, abs/2308.02463.

Wu, J.; Kim, Y.; and Wu, H. 2024. Hallucination Benchmark in Medical Visual Question Answering. arXiv:2401.05827.

Xia, P.; Zhu, K.; Li, H.; Zhu, H.; Li, Y.; Li, G.; Zhang, L.; and Yao, H. 2024. RULE: Reliable Multimodal RAG for Factuality in Medical Vision Language Models.

Zhang, S.; Xu, Y.; Usuyama, N.; Bagga, J. K.; Tinn, R.; Preston, S.; Rao, R. N.; Wei, M.-H.; Valluri, N.; Wong, C.; Lungren, M. P.; Naumann, T.; and Poon, H. 2023a. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs.

Zhang, X.; Wu, C.; Zhao, Z.; Lin, W.; Zhang, Y.; Wang, Y.; and Xie, W. 2023b. PMC-VQA: Visual Instruction Tuning for Medical Visual Question Answering. *ArXiv*, abs/2305.10415.

Zhang, Y.; Zhang, Y.; Qi, P.; Manning, C. D.; and Langlotz, C. P. 2021. Biomedical and clinical English model packages for the Stanza Python NLP library. *Journal of the American Medical Informatics Association*.

Zhou, Z.; Zhang, J.; Guan, Z.; Hu, M.; Lao, N.; Mu, L.; Li, S.; and Mai, G. 2024. Img2Loc: Revisiting Image Geolocalization using Multi-modality Foundation Models and Image-based Retrieval-Augmented Generation. *ArXiv*, abs/2403.19584.