RULE: Reliable Multimodal RAG for Factuality in Medical Vision Language Models

Peng Xia[1]*, **Kangyu Zhu**[2]*, **Haoran Li**[3], **Hongtu Zhu**[1], **Yun Li**[1], **Gang Li**[1], **Linjun Zhang**[4], **Huaxiu Yao**[1]
[1]UNC-Chapel Hill, [2]Brown University, [3]PolyU, [4] Rutgers University
{pxia,huaxiu}@cs.unc.edu

Abstract

The recent emergence of Medical Large Vision Language Models (Med-LVLMs) has enhanced medical diagnosis. However, current Med-LVLMs frequently encounter factual issues, often generating responses that do not align with established medical facts. Retrieval-Augmented Generation (RAG), which utilizes external knowledge, can improve the factual accuracy of these models but introduces two major challenges. First, limited retrieved contexts might not cover all necessary information, while excessive retrieval can introduce irrelevant and inaccurate references, interfering with the model's generation. Second, in cases where the model originally responds correctly, applying RAG can lead to an over-reliance on retrieved contexts, resulting in incorrect answers. To address these issues, we propose RULE, which consists of two components. First, we introduce a provably effective strategy for controlling factuality risk through the calibrated selection of the number of retrieved contexts. Second, based on samples where over-reliance on retrieved contexts led to errors, we curate a preference dataset to fine-tune the model, balancing its dependence on inherent knowledge and retrieved contexts for generation. We demonstrate the effectiveness of RULE on medical VQA and report generation tasks across three datasets, achieving an average improvement of 47.4% in factual accuracy. We publicly release our benchmark and code in https://github.com/richard-peng-xia/RULE.

1 Introduction

Artificial Intelligence (AI) has showcased its potential in medical diagnosis, including disease identification, treatment planning, and recommendations (Tăuţan et al., 2021; Wang et al., 2019; Ye et al., 2021; Xia et al., 2024b; Hu et al., 2024b,a). In particular, the recent development of Medical Large Vision Language Models (Med-LVLMs) has

*Equal Contribution.

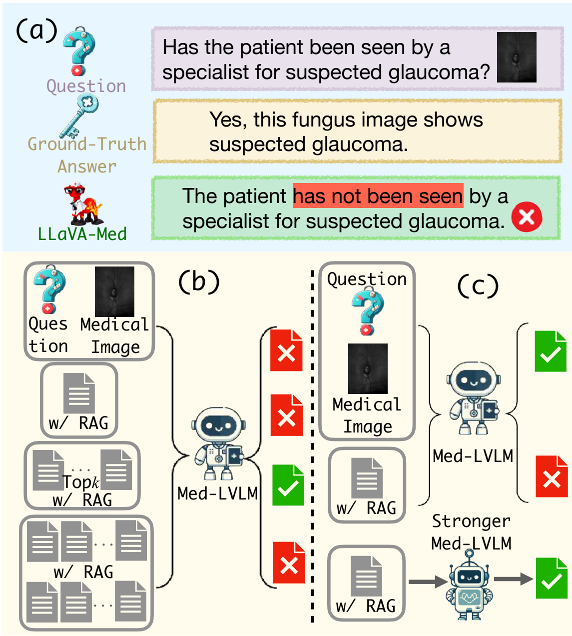

Figure 1: (a) An example of factuality issue in Med-LVLM. (b) Utilizing either too few or too many retrieved contexts as references may not provide effective guidance for the model's generation. Calibrating the number of retrieved contexts can effectively control the risk of factual inaccuracies. (c) Med-LVLMs often overly rely on retrieved contexts, leading to incorrect responses even when the original answers are correct without RAG. A stronger fine-tuned model can effectively balance its own knowledge with the retrieved contexts.

introduced more accurate and customized solutions to clinical applications (Li et al., 2023; Moor et al., 2023; Zhang et al., 2023; Wu et al., 2023). While Med-LVLMs have demonstrated promising performance, they remain prone to generating responses that deviate from factual information, potentially resulting in inaccurate medical diagnoses. This susceptibility to hallucination underscores the need for enhanced mechanisms to ensure factual alignment in critical medical applications (see an example in Figure 1(a)) (Royer et al., 2024; Xia et al., 2024a)). Such errors pose a significant risk to clinical decision-making processes and can lead to adverse outcomes.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 1081–1093
November 12-16, 2024 ©2024 Association for Computational Linguistics

Recently, Retrieval-Augmented Generation (RAG) (Gao et al., 2023; Qu et al., 2024a,b) has emerged as a promising method for enhancing the factual accuracy of responses from Med-LVLMs. By integrating external, reliable data sources, RAG guides the model in producing factual medical responses, enriching its knowledge base with supplementary information. For example, RAG has been used in tasks such as visual question answering (VQA) (Yuan et al., 2023) and report generation (Kumar and Marttinen, 2024; Tao et al., 2024). However, as illustrated in Figure 1(b) and Figure 1(c), directly applying RAG strategy to Med-LVLMs presents *two significant challenges*: (1) A small number of retrieved contexts may not cover the reference knowledge required for the question, thus limiting the model's factual accuracy. Conversely, a large number of retrieved contexts may include low-relevance and inaccurate references, which can interfere with the model's generation; (2) Med-LVLMs may overly rely on the retrieved information. In this situation, the model might correctly answer on its own, but incorporating the retrieved contexts could lead to incorrect responses.

To tackle these challenges, we propose the **R**eliable m**U**ltimoda**L** RAG called RULE for **ME**d-LVLMs. First, RULE introduces a provable strategy for factuality risk control through calibrated selection of the number of retrieved contexts k, ensuring that Med-LVLMs provably achieve high accuracy without the need for additional training (Angelopoulos et al., 2021). Specifically, this strategy modifies the Med-LVLM through a post-processing step that performs hypothesis testing for each k to determine whether the risk can be maintained above an acceptable threshold. This process begins by calculating the p-value for each k. Fixed sequence testing is then used to determine which k values can be accepted. Second, to mitigate over-reliance on retrieved knowledge, we introduce a knowledge balanced preference fine-tuning strategy. This strategy harmonizes the model's internal knowledge with retrieved contexts during medical response generation. Here, we identify samples where the model initially responds correctly but gives incorrect answers after incorporating retrieved contexts as dispreferred samples, indicating retrieval over-dependence. Conversely, ground-truth responses are considered as preferred samples. The curated preference data is then utilized for fine-tuning the preferences in Med-LVLMs.

Our primary contributions of this paper is RULE, which introduces an innovative approach to enhance retrieval-based Med-LVLMs. RULE not only controls factual risk by calibrating the selection of reference contexts but also balances the model's knowledge and retrieved contexts through preference fine-tuning using a curated preference dataset. Across three medical Visual Question Answering (VQA) and report generation benchmarks, including radiology and ophthalmology, our empirical results demonstrate that RULE effectively improves the factual accuracy of Med-LVLMs, achieving a 14.46% improvement over the best prior methods for mitigating hallucination. In addition, empirically verify the effectiveness of the proposed components and demonstrate the compatibility of RULE.

2 Preliminaries

In this section, we will provide a brief overview of Med-LVLMs and preference optimization.

Medical Large Vision Language Models. Med-LVLMs connects the LLMs and medical visual modules, enabling the model to use medical images x_v and clinical queries x_t as inputs x. This allows the model to autoregressively predict the probability distribution of the next token. The text output of Med-LVLMs is denoted as y.

Preference Optimization. Preference optimization has achieved remarkable results in efficiently fine-tuning LLMs, significantly aligning their behavior with the goals. Typically, give an input x, a language model policy π_θ can produce a conditional distribution $\pi_\theta(y \mid x)$ with y as the output text response. The recently popular DPO (Rafailov et al., 2023) utilizes preference data achieve objective alignment in LLMs. The preference data is defined as $\mathcal{D} = \{x^{(i)}, y_w^{(i)}, y_l^{(i)}\}_{i=1}^N$, where $y_w^{(i)}$ and $y_l^{(i)}$ represent preferred and dispreferred responses given an input prompt x. The probably of obtaining each preference pair is $p(y_w \succ y_l) = \sigma(r(x, y_w) - r(x, y_l))$, where $\sigma(\cdot)$ is the sigmoid function. In DPO, the optimization can be formulated as classification loss over the preference data as:

$$\mathcal{L}_{DPO}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}} \left[\log \sigma \left(\alpha \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \alpha \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right]. \quad (1)$$

where π_θ represents the reference policy, which is the LLM fine-tuned through supervised learning.

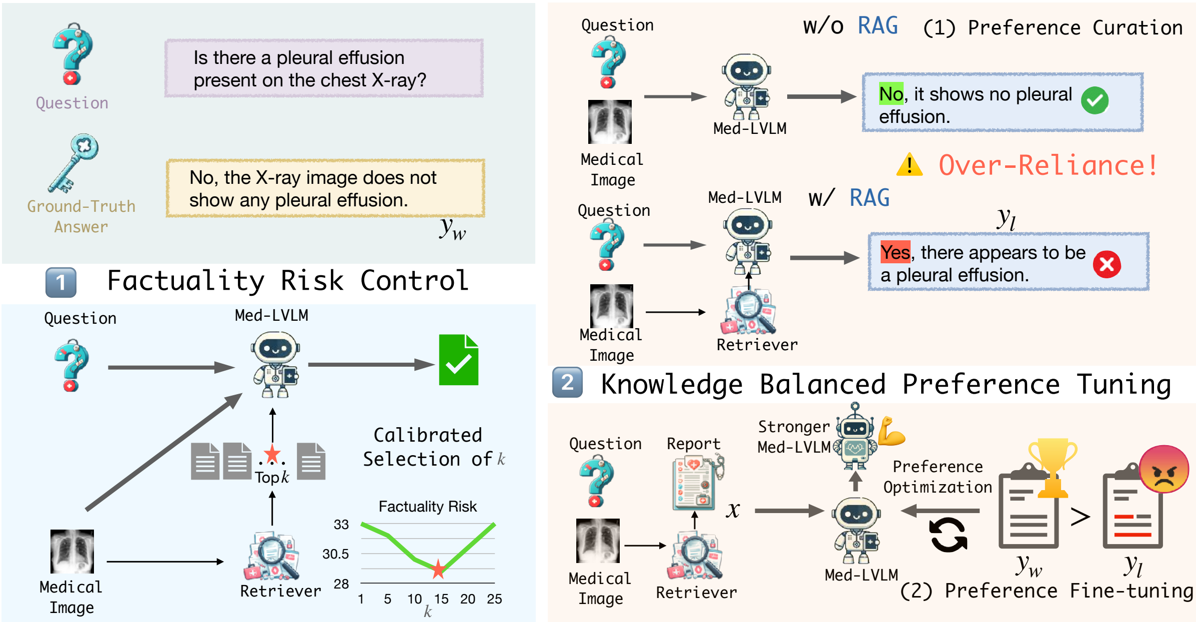

Figure 2: The framework of RULE comprises two main components: (1) a factuality risk control strategy through the calibrated selection of k; (2) knowledge-retrieval balance tuning. During the tuning phase, we initially construct a preference dataset from samples where the model errs due to excessive reliance on retrieved contexts. We subsequently fine-tune the Med-LVLM using this dataset by employing preference optimization.

3 Methodology

In this section, as illustrated in Figure 2, we will introduce RULE as an efficient solution for improving factuality of Med-LVLMs. Specifically, our approach consists of three main modules that work together to optimize the model's performance. First, we apply the retrieval strategy to Med-LVLMs, enhancing the model's ability to leverage retrieved information. Second, we implement a statistical method to control the factuality risk through calibrated selection of retrieved contexts. Third, we develop a preference optimization method to balance the model's reliance on its own knowledge and the retrieved contexts. Next, we will detail these three key modules in detail as follows:

3.1 Context Retrieval for Reference

Med-LVLMs often generate non-factual responses when dealing with complex medical images. RAG can provide the model with external knowledge as a reference, thereby effectively enhancing the factual accuracy. In the multimodal knowledge retrieval stage, RULE retrieves textual descriptions/reports that are most similar to the features of the target medical images. These references contain a wealth of image-based medical facts and serve to guide the generation of responses for the medical image.

Following the design of CLIP (Radford et al., 2021), the retriever will first encode each image and the corresponding reports into embeddings using a vision encoder and a text encoder, respectively. Specifically, all medical images X_{img} are encoded into image representations $V_{img} \in \mathbb{R}^{N\times P}$ by a vision encoder $\mathcal{E}_{img}$ (i.e., $V_{img} = \mathcal{E}_{img}(X_{img})$), where N is the number of medical images that need to be retrieved, and P is the dimension of the embedding. Similarly, we generate text embeddings $V_{txt} \in \mathbb{R}^{N\times P}$ for all corresponding medical reports X_{txt} by applying a text encoder $\mathcal{E}_{txt}$, i.e., $V_{txt} = \mathcal{E}_{txt}(X_{txt})$. Subsequently, to adapt the general vision and text encoders to the medical domain, we fine-tune the encoders using the training data with a contrastive learning loss, defined as:

$$\mathcal{L} = \frac{\mathcal{L}_{img} + \mathcal{L}_{text}}{2},$$
$$\text{where } \mathcal{L}_{img} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(S_{i,i})}{\sum_{j=1}^{N}\exp(S_{i,j})}, \quad (2)$$
$$\mathcal{L}_{text} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(S_{i,i})}{\sum_{j=1}^{N}\exp(S_{j,i})},$$

where $S \in \mathbb{R}^{N\times N}$ represents the similarity matrix between image and text modalities, calculated as: $S = \frac{V_{img}}{|V_{img}|} \cdot (\frac{V_{txt}}{|V_{txt}|})^T$, where each element $S_{i,j}$ represents the similarity between the image representation of example i and the text representation of example j. Equation (2) aims to learn the representations by maximizing the similarity of text and image modalities representing the same example, while minimizing the similarity of text and image modalities representing different examples.

After fine-tuning the image and text encoders, during inference, when faced with a target medical image x_t requiring the generation of its medical report, we extract the top-K similar medical reports $\text{TopK}_{j \in \{1...N\}} S_{t,j}$. We then use the retrieved medical report to guide the generation of the medical report for the target medical image. with the following prompt guidance: `"You are provided with a medical image, a image-related question and a reference report. Please answer the question based on the image and report. [Question] [Reference Report] [Image]"`.

3.2 Factuality Risk Control Through Calibrated Retrieved Context Selection

For the RAG strategy, the top-3/5 result is typically used as a reference (Gao et al., 2023). However, it sometimes fails to encompass all relevant retrieved contexts, especially when facing the fine-grained features of medical images. Additionally, an excessive amount of retrieved contexts may introduce low-relevance and inaccurate references, which can interfere with the model's generation. Thus, an algorithm that can automatically determine the optimal number of retrieved contexts, based on the risk of factual errors, is particularly crucial.

In this section, motivated by (Angelopoulos et al., 2021), we propose the following strategy to choose a subset $\hat{\Lambda}$ for the number of retrievals k from a candidate set $C_K \subseteq \mathbb{N}$ such that the factuality risk $FR(k)$ can be provably controlled for any $k \in \hat{\Lambda}$. Specifically, first, for each $k \in C_K$, the strategy first calculates the factuality risk $FR(k)$, computed as $1 - \text{ACC}(\mathcal{M}(x, (q, T_k)))$, where x denotes the target medical image, q denotes the question, T_k means the selected top-K retrieved contexts, and $\text{ACC}(\cdot)$ measures the ratio of correct answers provided by the Med-LVLM $\mathcal{M}$ to the total number of answers. Next, two probabilities p_{k1} and p_{k2} are computed as:

$$\begin{aligned} p_{k1} &= \exp(-nh_1(FR(k) \wedge \alpha, \alpha)), \\ p_{k2} &= e \cdot \mathbb{P}(Bin(n, \alpha) \leq \lceil nFR(k) \rceil), \end{aligned} \tag{3}$$

where $h_1(a, b) := a \log(a/b) + (1 - a) \log((1 - a)/(1 - b))$ is the Kullback-Leibler divergence between two Bernoulli distributions and α denotes risk upper bound. p_{k2} representing the probability that, in a binomial distribution with parameters n and α, denoted by $Bin(n, \alpha)$, the observed value is less than or equal to $\lceil nFR(k) \rceil$. Then, the minimum of these two probabilities $p_k = \min(p_{k1}, p_{k2})$ is taken. Finally, we use any family-wise error rat (FWER)-controlling procedure, such as Bonferroni correction (Van der Vaart, 2000) or sequential graphical testing (Bretz et al., 2009), to choose $\hat{\Lambda}$. For example, for Bonferroni correction, if p_k is less than or equal to $\delta/|C_K|$, where δ denotes tolerance level, then k is added to the set $\hat{\Lambda}$. The proposed strategy calculates the model's factuality risk under different k values, computes the corresponding probabilities using two approaches, and selects those k values that meet the risk tolerance to control the overall factuality risk.

We have the following result that ensures with probability at least $1 - \delta$, the factuality risk produced is controlled by α.

Proposition 1 *Let $\alpha, \delta \in (0, 1)$. If the training dataset $\mathcal{D}_{Med} = \{x_i, y_i, q_i\}_{i=1}^N$ is i.i.d. and the output of the above algorithm $\hat{\Lambda} \neq \emptyset$, then*

$$\mathbb{P}_{\mathcal{D}_{Med}}(\sup_{k \in \hat{\Lambda}} FR(k) \leq \alpha) \geq 1 - \delta.$$

In practice, we calibrate the selection of k on the validation sets of each dataset to minimize factuality risk. Consequently, the optimal k calibrated by this algorithm can be directly used on the test sets.

3.3 Knowledge Balanced Preference Tuning

In addition to selecting the optimal number k of retrieved contexts, it is likely that these contents often fail to fully capture the details of every lesion or normal area in medical images. Therefore, when the retrieved contexts is inaccurate, a reliable Med-LVLM is expected to remain unaffected by the unreliable information and independently use its own knowledge to answer medical questions. However, empirically, as illustrated in Table 1, approximately half of all incorrect responses by the retrieval-augmented Med-LVLM are due to an over-reliance on retrieved contexts. This significantly affects the application of the retrieval augmented generation strategy to Med-LVLMs.

Table 1: Over-Reliance Ratio (%) of Med-LVLM with retrieval, which is the proportion of errors due to over-reliance on retrieved contexts relative to the total number of incorrect answers.

IU-Xray	FairVLMed	MIMIC-CXR
47.42	47.44	58.69

To address this issue, we propose a Knowledge-Balanced Preference Tuning (KBPT) strategy to mitigate over-reliance on retrieved contexts and enhance factuality in medical content generation. Specifically, we select samples $\mathcal{D} = \{x^{(i)}, y^{(i)}, q^{(i)}\}_{i=1}^{N}$ from the a separate set with samples are not used to fine-tune the retriever in Section 3.1, where x, y, q denotes input medical image, ground-truth answer and question, respectively. We identify responses $a_b = \mathcal{M}(x, q)$ where the model originally answers (i.e., $a_b = y$) correctly but gives incorrect answers $a_f = \mathcal{M}(x, (q, t))$ after incorporating retrieved contexts as dispreferred responses, as they indicate over-dependence on the retrieval. Conversely, ground-truth answers y are considered preferred responses. We denote the preference dataset as $\mathcal{D}_o = \{x^{(i)}, y_{w,o}^{(i)}, y_{l,o}^{(i)}\}_{i=1}^{N}$, where $y_{w,o}^{(i)}$, $y_{l,o}^{(i)}$ are represented as preferred and dispreferred responses, respectively.

Based on the curated preference data, we fine-tune the Med-LVLM using direct preference optimization. Following Eqn. (1), the loss is calculated as follows:

$$\mathcal{L}_{kbpt} = -\mathbb{E}_{(x,y_{w,o},y_{l,o})\sim\mathcal{D}} \left[\log \sigma \left(\alpha \log \frac{\pi_\theta(y_{w,o}|x)}{\pi_o(y_{w,o}|x)} - \alpha \log \frac{\pi_\theta(y_{l,o}|x)}{\pi_o(y_{l,o}|x)}\right)\right]. \quad (4)$$

Algorithm 1: Reliable Multimodal RAG for Factuality (**RULE**)

Input: $\mathcal{D} = \{x^{(i)}, y^{(i)}, q^{(i)}\}_{i=1}^{N}$: Dataset; π_θ: Parameters of the Med-LVLM; $\mathcal{D}_o$: Preference dataset; Med-LVLM: $\mathcal{M}(\cdot,\cdot)$; Retriever: $\mathcal{R}(\cdot)$; $\mathcal{D}_o$: Preference dataset.

Output: π_{ref}: Parameters of the reference model.

▷ *Training Stage*
Initialize $\mathcal{D}_o$ with an empty set
foreach $(x, y, q) \in \mathcal{D}$ **do**
  Generate retrieved contexts $t \leftarrow \mathcal{R}(x)$
  Get the predictions of the model w/o retrieval $a_b \leftarrow \mathcal{M}(x, q)$
  Get the predictions of the model w/ retrieval $a_f \leftarrow \mathcal{M}(x, (q, t))$
  if $a_b = y$ and $a_f \neq y$ **then**
    Select the preferred response $y_{w,o} \leftarrow y$
    Select the dispreferred response $y_{l,o} \leftarrow a_f$
    Put $\{x, y_{w,o}, y_{l,o}\}$ into $\mathcal{D}_o$;
foreach $(x, y_{w,o}, y_{l,o}) \in \mathcal{D}_o$ **do**
  Compute the losses $\mathcal{L}_o$ following Eqn. (4)
  Update π_{ref} by minimizing $\mathcal{L}_o$
▷ *Inference Stage*
foreach test sample (x, q) **do**
  Select top-k retrieved contexts of calibrated algorithm $T_k \leftarrow \mathcal{R}(x)$
  Get the predictions of the model w/ KBPT and retrieval $a \leftarrow \mathcal{M}(x, (q, T_k))$

4 Experiment

In this section, we evaluate the performance of RULE, aiming to answer the following questions: (1) Can RULE effectively improve the factuality of Med-LVLMs compared to other baselines and open-sourced Med-LVLMs? (2) Do all proposed components boost the performance? (3) How does RULE change attention weights of retrieved contexts to balance model knowledge and retrieved contexts? (4) How do different types of data or models influence DPO fine-tuning?

4.1 Experimental Setups

Implementation Details. We utilize LLaVA-Med-1.5 7B (Li et al., 2023) as the backbone model. During the preference optimization process, we adapt LoRA fine-tuning (Hu et al., 2021). For the training of retriever, the vision encoder is a ResNet-50 (He et al., 2016), and the text encoder is a bio-BioClinicalBERT (Alsentzer et al., 2019). We use the AdamW optimizer with a learning rate of 10^{-3}, weight decay of 10^{-2} and a batch size of 32. The model is trained for 360 epochs. For more detailed information on training hyperparameters and training data, please see Appendix A and C.

Baselines. We compare RULE with LVLM hallucination mitigation methods that have already shown promising results in natural images, including Greedy Decoding, Beam Search (Sutskever et al., 2014), DoLa (Chuang et al., 2023), OPERA (Huang et al., 2023), VCD (Leng et al., 2023). These methods manipulate the logits of the model's output tokens to enhance factual accuracy. Furthermore, we compare the performance with other open-source Med-LVLMs, including Med-Flamingo (Moor et al., 2023), MedVInT (Zhang et al., 2023), RadFM (Wu et al., 2023).

Evaluation Datasets. To ensure that the retrieved report content is relevant to the visual question content and to facilitate experimentation, we utilize three medical vision-language datasets, i.e., MIMIC-CXR (Johnson et al., 2019), IU-Xray (Demner-Fushman et al., 2016), and Harvard-FairVLMed (Luo et al., 2024), encompassing radiology and ophthalmology. The training set is split into two parts: one part is used to train the retriever (Section 3.1), and the other part is used to construct the preference dataset for KBPT (Section 3.3).

Additionally, we construct VQA pairs for KBPT and evaluation. Specifically, the reports from training set for preference dataset and reports from orig-

Table 2: Factuality performance (%) of Med-LVLMs on the three VQA datasets. Notably, we report the accuracy, precision, recall, and F1 score. The best results and second best results are **bold** and <u>underlined</u>, respectively.

Models	IU-Xray				Harvard-FairVLMed				MIMIC-CXR			
	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
LLaVA-Med-1.5	75.47	53.17	80.49	64.04	63.03	92.13	61.46	74.11	75.79	81.01	79.38	80.49
+ Greedy	76.88	54.41	82.53	65.59	78.32	91.59	82.38	86.75	<u>82.54</u>	82.68	81.73	85.98
+ Beam Search	76.91	54.37	<u>84.13</u>	66.06	<u>80.93</u>	<u>93.01</u>	<u>82.78</u>	<u>88.08</u>	81.56	<u>83.04</u>	**84.76**	<u>86.36</u>
+ DoLa	<u>78.00</u>	<u>55.96</u>	82.69	<u>66.75</u>	76.87	92.69	79.40	85.53	81.35	80.94	81.07	85.73
+ OPEAR	70.59	44.44	**100.0**	61.54	71.41	92.72	72.49	81.37	69.34	72.04	79.19	76.66
+ VCD	68.99	44.77	69.14	54.35	65.88	90.93	67.07	77.20	70.89	78.06	73.23	75.57
RULE (Ours)	**87.84**	**75.41**	80.79	**78.00**	**87.12**	**93.57**	**96.69**	**92.89**	**83.92**	**87.01**	<u>82.89</u>	**87.49**

Table 3: Factuality performance (%) of Med-LVLMs on the three report generation datasets. Notably, we report the average BLEU, ROUGE-L, METEOR.

Models	IU-Xray			MIMIC-CXR			Harvard-FairVLMed		
	BLEU	ROUGE-L	METEOR	BLEU	ROUGE-L	METEOR	BLEU	ROUGE-L	METEOR
LLaVA-Med-1.5	9.64	12.26	8.21	12.11	13.05	11.16	18.11	11.36	10.75
+ Greedy	11.47	15.38	12.69	16.63	14.26	14.19	17.98	11.49	13.77
+ Beam Search	<u>12.10</u>	<u>16.21</u>	<u>13.17</u>	16.97	14.74	14.43	<u>18.37</u>	<u>12.62</u>	14.50
+ DoLa	11.79	15.82	12.72	<u>17.11</u>	<u>14.89</u>	<u>14.81</u>	18.26	12.51	<u>14.51</u>
+ OPERA	10.66	14.70	12.01	15.40	12.52	13.72	16.59	11.47	13.63
+ VCD	10.42	14.14	11.59	15.18	12.30	13.38	16.73	11.38	13.89
+ RULE (Ours)	**27.53**	**23.16**	**27.99**	**18.61**	**15.96**	**17.42**	**22.35**	**14.93**	**17.74**

inal test set are input into GPT-4 (OpenAI, 2023) to create closed-ended VQA data with *yes* or *no* answers, *e.g.*, *"Is there any pulmonary nodule?"*. By sampling segments from a medical report, we can generate a sequence of concise, closed-ended questions posed to the model, each with accurate answers. The questions are in *yes/no* format, making it easier to analyze errors caused by over-reliance on retrieved contexts compared to open-ended questions. The detailed construction process and dataset statistics are provided in the Appendix A.

Evaluation Metrics. For Med-VQA task, we use Accuracy as the primary metric and, for detailed comparisons, we also adopt Precision, Recall, and F1 Score. For report generation task, we use BLEU Score (Papineni et al., 2002), ROUGE-L (Lin, 2004) and METEOR (Banerjee and Lavie, 2005) as the metrics.

4.2 Results

In this section, we provide comprehensive comparison results with different baseline methods and other open-sourced Med-LVLMs.

Comparison with Baseline Methods. We present the results of a comparison between RULE and various hallucination reduction methods in Table 2. According to these results, RULE demonstrates the best overall performance, effectively and accurately diagnosing diseases with an average accuracy improvement of 47.4% on two tasks across all datasets. We also observe that RULE performs notably better on the IU-Xray and Harvard-FairVLMed compared to MIMIC-CXR. This difference is attributed to the excessive length of the reports available for retrieval in MIMIC-CXR, where overly long references tend to confuse the Med-LVLM. Even when dealing with the relatively niche ophthalmology data (i.e., Harvard-FairVLMed), RULE demonstrates superior results, significantly enhancing the factual accuracy of the Med-LVLM. In contrast, the performance of decoding methods is quite unstable, showing significant rates of missed or incorrect diagnoses across different datasets, as indicated by the precision and recall values.

Comparison with Other Med-LVLMs. In Table 4, we present the comparison with different open-sourced Med-LVLMs. RULE demonstrates state-of-the-art (SOTA) performance across all datasets. Although the second-best model, Med-VInT, outperforms other models, RULE achieves an average accuracy improvement of 47.4% over it. Whether in radiology or ophthalmology, RULE demonstrates remarkable performance, significantly surpassing other open-source Med-LVLMs. This indicates that RULE is generally applicable and effective in the medical multimodal diagnosis, providing consistent improvements across various medical image modalities.

(a)

Comparison of Med-LVLM w/ or w/o KBPT. We report the Error (1-ACC) and Over-Reliance Ratio (ORR) (%).

Datasets		w/o	w/
IU-Xray	Error	22.85	**15.93**
	ORR	47.42	**27.16**
FairVLMed	Error	33.79	**15.19**
	ORR	47.44	**22.43**
MIMIC-CXR	Error	32.65	**19.86**
	ORR	58.69	**31.35**

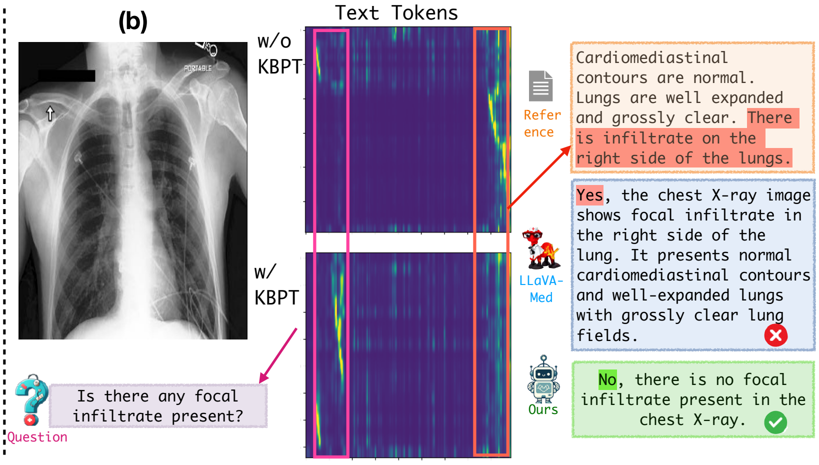

Figure 3: Comparison of over-reliance metrics and attention maps. After optimizing the model with knowledge balanced preference tuning, first, (a) the Med-LVLM's error (1-acc) and over-reliance ratio significantly decrease. Second, (b) the attention scores for the latter half of the text tokens, i.e., the retrieved contexts, are significantly reduced, while the attention scores for the first half of the text tokens, i.e., the questions, have increased. It indicates that RULE effectively mitigates the model's over-reliance on retrieved contexts and enhances factual accuracy.

Table 4: Comparison with other open-sourced Med-LVLMs. Here "FairVLMed": Harvard-FairVLMed.

Models	IU-Xray	FairVLMed	MIMIC-CXR
Med-Flamingo	26.74	42.06	61.27
MedVInT	73.34	35.92	66.06
RadFM	26.67	52.47	69.30
RULE (Ours)	**87.84**	**87.12**	**83.92**

Table 5: Results of ablation study. Here, "R": retrieval; "FRC": factuality risk control, "KBPT": knowledge balanced preference tuning.

Models	IU-Xray	FairVLMed	MIMIC-CXR
LLaVA-Med-1.5	75.47	63.03	75.79
+ R	77.15	66.21	67.35
+ FRC	78.62	80.61	76.54
+ KBPT + R	84.07	84.81	80.14
+ KBPT + FRC (Ours)	**87.84**	**87.12**	**83.92**

4.3 How Does RULE Improve the Performance?

In this section, we conduct a set of analyses demonstrate how different components contribute to the performance and illustrate how RULE enhances overall performance, which are details as follows:

Ablation Studies. To further illustrate the effectiveness of the components of RULE, we conduct ablation experiments on three datasets. The results are shown in Table 5. We find that the basic RAG strategy ("R") slightly improves factual accuracy on two datasets but decreases it on MIMIC-CXR. The limited retrieved contexts can not cover the fine-grained features of medical images, resulting in unstable factual accuracy improvements. With the aid of the factuality risk control strategy ("FRC"), retrieval performance see a stable increase, outperforming the original Med-LVLM. Considering the model's over-reliance on retrieved contexts, the knowledge balanced preference tuning ("KBPT") further enhances the model's reliability and significantly improves its performance. Ultimately, by combining these two strategies, RULE achieves optimal performance.

How does RULE Mitigate the Issue of Over-Reliance on Retrieved Contexts? To better understand how RULE mitigates the Med-LVLM's over-reliance on retrieved contexts, we measure the Med-LVLM's error and over-reliance ratios, and visualize the text and image attention maps of the models before and after fine-tuning using a randomly selected case, as shown in Figure 3. The quantitative results in Figure 3(a) demonstrate the significant positive impact of RULE in mitigating the model's over-reliance on retrieved contexts, with the error rate and over-reliance rate decreasing by an average of 42.9% and 47.3%, respectively. Attention maps Figure 3(b) illustrate the model's attention scores for text and image tokens. We find that, on the text side, the model with knowledge balanced preference tuning shows a significantly reduced focus on retrieved contexts, effectively mitigating over-reliance on such information. The model focuses more on the question and leverages its own knowledge to answer, rather than relying solely on the retrieved contexts, effectively enhancing factual accuracy.

Analyzing Preference Data Type in KBPT. We further conduct a thorough analysis of the data types used in constructing preference data for KBPT. Three formats are considered: medical image captioning (prompted as "Please describe

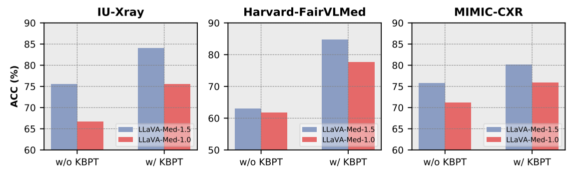

Figure 4: Results of RULE on different backbones. “KBPT": knowledge balanced preference tuning.

this medical image"), visual question-answering (VQA), and a mixture of both. The selected data are samples where the model makes errors due to over-reliance on retrieved contexts. The results are shown in Table 6. We observe that models fine-tuned using VQA data perform the best across all three datasets. This indicates that when retrieved contexts are incorporated into VQA questions, the Med-LVLM, through KBPT, can learn this paradigm of integrating and balancing its own knowledge with retrieved context to maximize factual accuracy. However, when the data is in the form of captioning, it may enhance the model’s ability to describe medical facts, but it merely distances the model’s answers from the retrieved contexts. The model fails to understand how to balance retrieval content with its own knowledge.

Table 6: Results of models fine-tuned on different formats of data.

Format	IU-Xray	FairVLMed	MIMIC-CXR
LLaVA-Med-1.5	75.47	63.03	75.79
Captioning	<u>81.61</u>	67.49	77.42
VQA	**84.07**	**84.81**	**80.14**
Merged	76.33	<u>67.96</u>	<u>78.99</u>

4.4 Compatibility Analysis

To demonstrate the compatibility of RULE, we conduct KBPT on LLaVA-Med-1.0 as well. The experimental results on three datasets are shown in Figure 4. We find that our knowledge balanced preference tuning method demonstrates good compatibility across different models, significantly improving factual accuracy across multiple datasets. Based on LLaVA-Med-1.0, RULE increases accuracy by an average of 16.7%. This indicates that RULE has a noticeable positive effect on mitigating over-reliance on retrieved contexts, thereby enhancing the Med-LVLM’s factual accuracy.

4.5 Case Study

Figure 5 presents two representative case results, demonstrating that RULE can effectively enhance the factual accuracy of med-LVLMs. In case 1, LLaVA-Med provides a factually incorrect answer. After applying the RAG strategy, the model still exhibits factual issues, whereas our method effectively addresses this and improves accuracy. In case 2, LLaVA-Med initially provides a correct answer, but due to the model’s over-reliance on retrieved contexts, it subsequently produces an incorrect response. RULE balances the weight of inherent knowledge and retrieved contexts, enhancing factual accuracy.

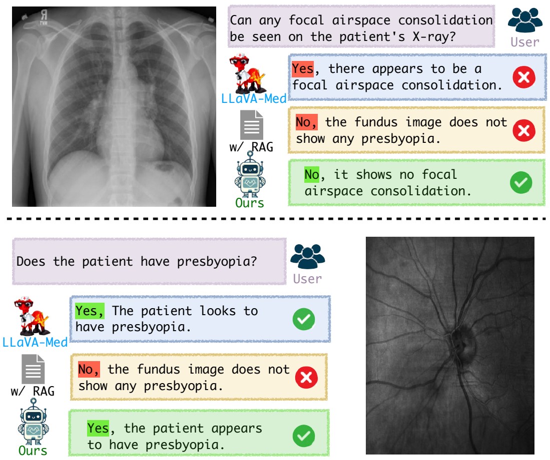

Figure 5: Illustrations of factuality enhancement by RULE in radiology and ophthalomology.

5 Related Work

Factuality in Med-LVLMs. The rapid development of Large Vision and Language Models (LVLMs) (Liu et al., 2023b,a; Zhu et al., 2023; Alayrac et al., 2022; Zhou et al., 2024a,b; Xia et al., 2024c, 2023) has begun to impact medical diagnosis. A series of Med-LVLMs (Li et al., 2023; Moor et al., 2023; Wu et al., 2023; Zhang et al., 2023), represented by LLaVA-Med, have emerged, demonstrating impressive performance across various medical image modalities. However, Med-LVLMs still exhibit significant factual errors, producing medical responses that conflict with the visual medical information (Xia et al., 2024a; Su et al., 2024). This could potentially lead to misdiagnoses or missed diagnoses. Recently, several benchmarks (Royer et al., 2024; Xia et al., 2024a) have been established to evaluate the accuracy of Med-LVLMs in tasks such as VQA or report generation. Beyond evaluating factuality, improving the factual accuracy of Med-LVLMs remains an underexplored area.

Retrieval Augmented Generation. RAG has recently been recognized as a promising solution (Gao et al., 2023; Sun et al., 2024). It enhances the model's ability to generate accurate facts by incorporating contextual information from external datasets. In medical multimodal analysis, the RAG approach has been applied to various tasks such as medical VQA (Yuan et al., 2023) and report generation (Kumar and Marttinen, 2024; Tao et al., 2024; He et al., 2024). However, in Med-LVLMs, applying RAG-based approaches overlook two critical issues: the number of retrieved contexts and whether the model overly relies on these reference. These factors can significantly affect the model's performance and may even degrade it. In RULE, we systematically address these challenges and enhance the factuality of Med-LVLMs.

6 Conclusion

In this work, we aim to enhance the factuality of Med-LVLM by addressing two key challenges in medical RAG. Specifically, we first introduce a provably effective strategy for controlling factuality risk through the calibrated selection of retrieved contexts. Second, we develop a preference optimization strategy that addresses errors stemming from the model's excessive dependence on retrieved contexts, aiming to balance its intrinsic knowledge and the retrieved information. Experiments on three medical imaging analysis datasets demonstrate the effectiveness of RULE.

Limitations

This work explores a reliable multimodal RAG method for Med-LVLMs to enhance factual accuracy. Our primary focus is on factual accuracy. Future research can explore other issues related to deploying Med-LVLMs in clinical settings, such as safety, fairness, robustness, and privacy.

Acknowledgement

This research was supported by Cisco Faculty Research Award.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems, 35:23716–23736.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. arXiv preprint arXiv:1904.03323.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. 2021. Learn then Test: Calibrating Predictive Algorithms to Achieve Risk Control. arXiv:2110.01052.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, pages 65–72.

Frank Bretz, Willi Maurer, Werner Brannath, and Martin Posch. 2009. A graphical approach to sequentially rejective multiple test procedures. Statistics in medicine, 28(4):586–604.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. arXiv preprint arXiv:2309.03883.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association, 23(2):304–310.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 770–778.

Sunan He, Yuxiang Nie, Zhixuan Chen, Zhiyuan Cai, Hongmei Wang, Shu Yang, and Hao Chen. 2024. Meddr: Diagnosis-guided bootstrapping for large-scale medical vision-language learning. arXiv preprint arXiv:2404.15127.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.

Ming Hu, Lin Wang, Siyuan Yan, Don Ma, Qingli Ren, Peng Xia, Wei Feng, Peibo Duan, Lie Ju, and Zongyuan Ge. 2024a. Nurvid: A large expert-level video database for nursing procedure activity understanding. Advances in Neural Information Processing Systems, 36.

Ming Hu, Peng Xia, Lin Wang, Siyuan Yan, Feilong Tang, Zhongxing Xu, Yimin Luo, Kaimin Song, Jurgen Leitner, Xuelian Cheng, et al. 2024b. Ophnet: A large-scale video benchmark for ophthalmic surgical workflow understanding. arXiv preprint arXiv:2406.07471.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2023. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. arXiv preprint arXiv:2311.17911.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. 2019. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042.

Yogesh Kumar and Pekka Marttinen. 2024. Improving medical multi-modal contrastive learning with expert annotations. arXiv preprint arXiv:2403.10153.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2023. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. arXiv preprint arXiv:2311.16922.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out, pages 74–81.

Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-clip: Contrastive language-image pre-training using biomedical documents. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 525–536. Springer.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. arXiv preprint arXiv:2304.08485.

Yan Luo, Min Shi, Muhammad Osama Khan, Muhammad Muneeb Afzal, Hao Huang, Shuaihang Yuan, Yu Tian, Luo Song, Ava Kouhana, Tobias Elze, et al. 2024. Fairclip: Harnessing fairness in vision-language learning. arXiv preprint arXiv:2403.19949.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. Med-flamingo: a multimodal medical few-shot learner. In Machine Learning for Health (ML4H), pages 353–367. PMLR.

OpenAI. 2023. Gpt-4 technical report. https://arxiv.org/abs/2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pages 311–318.

Xiaoye Qu, Qiyuan Chen, Wei Wei, Jishuo Sun, and Jianfeng Dong. 2024a. Alleviating hallucination in large vision-language models with active retrieval augmentation. arXiv preprint arXiv:2408.00555.

Xiaoye Qu, Jiashuo Sun, Wei Wei, and Yu Cheng. 2024b. Look, compare, decide: Alleviating hallucination in large vision-language models via multi-view multi-path reasoning. arXiv preprint arXiv:2408.17150.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In Thirty-seventh Conference on Neural Information Processing Systems.

Corentin Royer, Bjoern Menze, and Anjany Sekuboyina. 2024. Multimedeval: A benchmark and a toolkit for evaluating medical vision-language models. arXiv preprint arXiv:2402.09262.

Zhaochen Su, Jun Zhang, Xiaoye Qu, Tong Zhu, Yanshu Li, Jiashuo Sun, Juntao Li, Min Zhang, and Yu Cheng. 2024. Conflictbank: A benchmark for evaluating the influence of knowledge conflicts in llm. arXiv preprint arXiv:2408.12076.

Jiashuo Sun, Jihai Zhang, Yucheng Zhou, Zhaochen Su, Xiaoye Qu, and Yu Cheng. 2024. Surf: Teaching large vision-language models to selectively utilize retrieved information. arXiv preprint arXiv:2409.14083.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In Advances in neural information processing systems, pages 3104–3112.

Yitian Tao, Liyan Ma, Jing Yu, and Han Zhang. 2024. Memory-based cross-modal semantic alignment network for radiology report generation. IEEE Journal of Biomedical and Health Informatics.

Alexandra-Maria Tăuţan, Bogdan Ionescu, and Emiliano Santarnecchi. 2021. Artificial intelligence in neurodegenerative diseases: A review of available tools with a focus on machine learning techniques. Artificial Intelligence in Medicine, 117:102081.

Aad W Van der Vaart. 2000. Asymptotic statistics, volume 3. Cambridge university press.

Chunhao Wang, Xiaofeng Zhu, Julian C Hong, and Dandan Zheng. 2019. Artificial intelligence in radiotherapy treatment planning: present and future. Technology in cancer research & treatment, 18:1533033819873922.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024. Pmc-llama: toward building open-source language models for medicine. Journal of the American Medical Informatics Association, page ocae045.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Towards generalist foundation model for radiology. arXiv preprint arXiv:2308.02463.

Peng Xia, Ze Chen, Juanxi Tian, Yangrui Gong, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, et al. 2024a. Cares: A comprehensive benchmark of trustworthiness in medical vision language models. arXiv preprint arXiv:2406.06007.

Peng Xia, Ming Hu, Feilong Tang, Wenxue Li, Wenhao Zheng, Lie Ju, Peibo Duan, Huaxiu Yao, and Zongyuan Ge. 2024b. Generalizing to unseen domains in diabetic retinopathy with disentangled representations. In arXiv preprint arXiv:2406.06384.

Peng Xia, Di Xu, Ming Hu, Lie Ju, and Zongyuan Ge. 2024c. Lmpt: Prompt tuning with class-specific embedding loss for long-tailed multi-label visual recognition. In Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR), pages 26–36, Bangkok, Thailand. Association for Computational Linguistics.

Peng Xia, Xingtong Yu, Ming Hu, Lie Ju, Zhiyong Wang, Peibo Duan, and Zongyuan Ge. 2023. Hg-clip: Exploring vision-language models with graph representations for hierarchical understanding. arXiv preprint arXiv:2311.14064.

Qing Ye, Chang-Yu Hsieh, Ziyi Yang, Yu Kang, Jiming Chen, Dongsheng Cao, Shibo He, and Tingjun Hou. 2021. A unified drug–target interaction prediction framework based on knowledge graph and recommendation system. Nature communications, 12(1):6775.

Zheng Yuan, Qiao Jin, Chuanqi Tan, Zhengyun Zhao, Hongyi Yuan, Fei Huang, and Songfang Huang. 2023. Ramm: Retrieval-augmented biomedical visual question answering with multi-modal pre-training. In Proceedings of the 31st ACM International Conference on Multimedia, pages 547–556.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-vqa: Visual instruction tuning for medical visual question answering. arXiv preprint arXiv:2305.10415.

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. 2024a. Aligning modalities in vision large language models via preference fine-tuning. arXiv preprint arXiv:2402.11411.

Yiyang Zhou, Zhiyuan Fan, Dongjie Cheng, Sihan Yang, Zhaorun Chen, Chenhang Cui, Xiyao Wang, Yun Li, Linjun Zhang, and Huaxiu Yao. 2024b. Calibrated self-rewarding vision language models. arXiv preprint arXiv:2405.14622.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592.

A Data

A.1 Data statistics

The quantities of all the data used are shown in Table 7 and Table 8. It is notable to note that for training the retriever, this refers to the number of image-text pairs; for fine-tuning, it refers to the number of QA items. “All" represents the total quantity used to construct the preference dataset, where only the samples with correct original answers that become incorrect after adding retrieved contexts are included in the training of knowledge balanced preference tuning (“KBPT").

Dataset	Train (R)	All (KBPT)	Train (KBPT)
IU-Xray	1035	6761	1579
FairVLMed	7000	6271	2259
MIMIC-CXR	3000	4951	1106

Table 7: Data statistics of training set. Here, the number of data for the training of retriever (“R") means the number of image-caption pairs. The number of data for knowledge balanced preference tuning (“KBPT") means the number of question-answering pairs. FairVLMed: Harvard-FairVLMed.

Dataset	# Images	# QA Items
IU-Xray	589	2573
Harvard-FairVLMed	713	4285
MIMIC-CXR	700	3470

Table 8: Data statistics of test set. # Images and # QA items mean the number of images and QA pairs, respectively.

A.2 Instructions

We convert the medical reports into a series of closed-ended questions with yes or no answers. To ensure the quality of the VQA data, we perform a round of self-checks using GPT-4 (OpenAI, 2023). Finally, we conduct an round of manual filtering to remove questions with obvious issues or those related to multiple images or patient histories. The prompt templates used are shown in Table 9.

Instruction [Round1]
You are a professional medical expert. I will provide you with some medical reports. Please generate some questions with answers (the answer should be yes or no) based on the provided report. The subject of the questions should be the medical image or patient, not the report.
Below are the given report:
[REPORT]
Instruction [Round2]
Please double-check the questions and answers, including how the questions are asked and whether the answers are correct. You should only generate the questions with answers and no other unnecessary information.
Below are the given report and QA pairs in round1:
[REPORT]
[QA PAIRS R1]

Table 9: The instruction to GPT-4 for generating QA pairs.

A.3 Involved Datasets

We utilize three open-source medical vision-language datasets, i.e., MIMIC-CXR (Johnson et al., 2019), IU-Xray (Demner-Fushman et al., 2016), Harvard-FairVLMed (Luo et al., 2024).

- MIMIC-CXR (Johnson et al., 2019) is a large publicly available dataset of chest X-ray images in DICOM format with associated radiology reports.
- IU-Xray (Demner-Fushman et al., 2016) is a dataset that includes chest X-ray images and corresponding diagnostic reports.
- Harvard-FairVLMed (Luo et al., 2024) focuses on fairness in multimodal fundus images, containing image and text data from various sources. It aims to evaluate bias in AI models on this multimodal data comprising different demographics.

B Evaluated Models

We evaluate four open-source Med-LVLMs, *i.e.*, LLaVA-Med (Li et al., 2023), Med-Flamingo (Moor et al., 2023), MedVInT (Zhang et al., 2023), RadFM (Wu et al., 2023). The selected models are all at the 7B level.

- LLaVA-Med (Li et al., 2023) is a vision-language conversational assistant, adapting the general-domain LLaVA (Liu et al., 2023b) model for the biomedical field. The model is fine-tuned using a novel curriculum learning method, which includes two stages: aligning biomedical vocabulary with figure-caption pairs and mastering open-ended conversational semantics. It demonstrates excellent multimodal conversational capabilities.
- Med-Flamingo (Moor et al., 2023) is a multimodal few-shot learner designed for the medical domain. It builds upon the Open-Flamingo (Alayrac et al., 2022) model, continuing pre-training with medical image-text data from publications and textbooks. This model

aims to facilitate few-shot generative medical visual question answering, enhancing clinical applications by generating relevant responses and rationales from minimal data inputs.

- RadFM (Wu et al., 2023) serve as a versatile generalist model in radiology, distinguished by its capability to adeptly process both 2D and 3D medical scans for a wide array of clinical tasks. It integrates ViT as visual encoder and a Perceiver module, alongside the MedLLaMA (Wu et al., 2024) language model, to generate sophisticated medical insights for a variety of tasks. This design allows RadFM to not just recognize images but also to understand and generate human-like explanations.
- MedVInT (Zhang et al., 2023), which stands for Medical Visual Instruction Tuning, is designed to interpret medical images by answering clinically relevant questions. This model features two variants to align visual and language understanding (Wu et al., 2024): MedVInT-TE and MedVInT-TD. Both MedVInT variants connect a pre-trained vision encoder ResNet-50 adopted from PMC-CLIP (Lin et al., 2023), which processes visual information from images. It is an advanced model that leverages a novel approach to align visual and language understanding.

C Implementation Details

Following the settings of CLIP (Radford et al., 2021), we adopt the same architecture and hyperparameters for the vision and text encoders. The vision encoder is a ResNet-50 (He et al., 2016), and the text encoder is a bio-bert-based model (Alsentzer et al., 2019). We use the AdamW optimizer with a learning rate of 10^{-3}, weight decay of 10^{-2} and a batch size of 32. The model is trained for 360 epochs. The reports available for retrieval are from the training set of the corresponding dataset. In our experiments, we apply cross-validation to tune all hyperparameters with grid search. All the experiments are implemented on PyTorch 2.1.2 using four NVIDIA RTX A6000 GPUs. It takes roughly 2.5 and 4 hours for fine-tuning CLIP and LLaVA-Med-1.5 7B, respectively.

D Proofs

Proof of Proposition 1: According to the definition, $\mathcal{M}(\cdot,\cdot)$ denotes the Med-LVLM. $\{T_k\}_{i=1}^N$ denotes the top$k$ retrieved contexts. The dataset is $\mathcal{D}_{Med} = \{x_i, y_i, q_i\}_{i=1}^N$, where x_i is the target image, y_i is the ground-truth answer, q_i is the target question. By the definition of $FR(k)$,

$$\begin{aligned} FR(k) =& 1 - \text{ACC}(\mathcal{M}(x,(q,\{T_k\}_{i=1}^N))) \\ =& 1 - \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{\mathcal{M}(x_i,(q_i,\{T_k\}_{i=1}^N)) \\ =& y_i\} \\ =& \frac{1}{N}\sum_{i=1}^{N}(1 - \mathbf{1}\{\mathcal{M}(x_i,(q_i,\{T_k\}_{i=1}^N)) \\ =& y_i\}) \end{aligned}$$

Therefore, $FR(k)$ can be written as the average value of a function evaluated at each data point (x_i, y_i, q_i) in $\mathcal{D}_{Med}$. Then, by combining Theorem 1, Proposition 1 and Proposition 2 of (Angelopoulos et al., 2021), we finish the proof.