HeteroRAG: A Heterogeneous Retrieval-Augmented Generation Framework for Medical Vision Language Tasks

Zhe Chen[1,3], **Yusheng Liao**[1,3], **Zhiyuan Zhu**[1], **Haolin Li**[2,3], **Hongcheng Liu**[1]
Yanfeng Wang[1,3], **Yu Wang**[1,3]✉
[1]Shanghai Jiao Tong University [2]Fudan University
[3]Shanghai Artificial Intelligence Laboratory
{chenzhe2018,yuwangsjtu}@sjtu.edu.cn

Abstract

Medical large vision-language Models (Med-LVLMs) have shown promise in clinical applications but suffer from factual inaccuracies and unreliable outputs, posing risks in real-world diagnostics. While RAG has emerged as a potential solution, current medical multimodal RAG systems are unable to perform effective retrieval across heterogeneous sources. The irrelevance of retrieved reports undermines the factuality of analysis, while insufficient knowledge affects the credibility of clinical decision-making. To bridge the research gap, we construct MedAtlas, which includes extensive multimodal report repositories and diverse text corpora. Based on it, we present HeteroRAG, a novel framework that enhances Med-LVLMs through heterogeneous knowledge sources. The framework introduces Modality-specific CLIPs for effective report retrieval and a Multi-corpora Query Generator for tailoring queries to diverse corpora. Incorporating knowledge from such multifaceted sources, Heterogeneous Knowledge Preference Tuning is performed to achieve cross-modality and multi-source knowledge alignment. Extensive experiments across 11 datasets and 3 modalities demonstrate that HeteroRAG achieves state-of-the-art performance in most medical vision language benchmarks, significantly improving factual accuracy and reliability of Med-LVLMs[1].

1 Introduction

Large vision-language models (LVLMs) have made significant strides in integrating multimodal information and generating natural responses (Chen et al., 2024c,b; Liu et al., 2024a; Comanici et al., 2025; Bai et al., 2025). Similarly, medical LVLMs (Med-LVLMs) show increasing promise for multimodal diagnosis and clinical decision support (Chen et al., 2024a; Liu et al., 2024b; Lin et al., 2025; Xu et al., 2025). However, despite these advances, current Med-LVLMs still struggle with critical challenges related to factual accuracy and reliability (Sun et al., 2025; Xia et al., 2025), which is essential in diagnosis (Xiong et al., 2024; Li et al., 2025a). This limitation poses serious risks in medical applications, where errors could lead to misdiagnosis or harmful treatment recommendations.

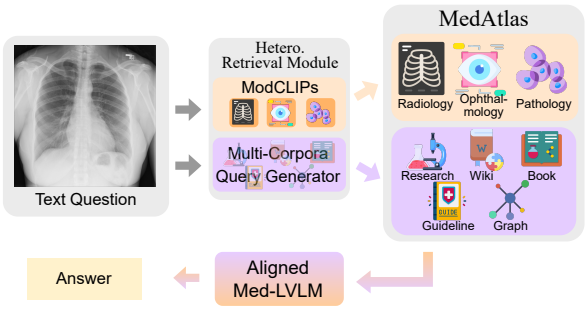

Figure 1: Overview of HeteroRAG Framework. The HRM retrieves reports and documents from MedAtlas for a knowledge-aligned Med-LVLM.

To mitigate these limitations, recent scholarly efforts have prioritized multimodal retrieval-augmented generation (MMRAG) frameworks, which augment Med-LVLMs with medical knowledge to enhance diagnostic accuracy and epistemic reliability (Ranjit et al., 2023; Sun et al., 2025; Choi et al., 2025; Shaaban et al., 2025). Predominant methodologies employ multimodal retrievers, such as the medical modality-aware CLIP models, to retrieve relevant reports using input images, owing to the strong semantic similarity between medical imaging and textual reports (Sun et al., 2025; Xia et al., 2024, 2025). **However, the training data used to enhance the modality awareness of these retrievers is typically limited to the training splits of only a few datasets.** This constraint leads to inferior retrieval performance and irrelevant retrieved reports, undermining the factuality

✉: Corresponding author.
[1]Project website: https://github.com/Jack-ZC8/HeteroRAG-Med

Findings of the Association for Computational Linguistics: ACL 2026, pages 3592–3613
July 2-7, 2026

of the Med-LVLMs.

Moreover, medical corpora, such as research articles and clinical guidelines, are crucial for enhancing the reliability of Med-LVLMs. However, the multimodal retrievers mentioned above fail when applied to corpora, since these corpora lack direct visual semantics and exhibit diverse linguistic characteristics. Current efforts (Wu et al., 2025; Hamza et al., 2025) typically conduct cross-modality document retrieval using the original multimodal query. **Though straightforward, they fail in corpus-specific retrieval as they neglect the alignment between queries and corpus linguistic characteristics as noted above.** MIRA (Wang et al., 2025), which employs zero-shot LLM-rewritten queries, still lacks this customizability due to limited information presented in the rewriting prompt. **In summary, current medical MMRAG works remain ineffective in retrieving across heterogeneous imaging reports and literature corpora, resulting in a significant knowledge gap.**

A key bottleneck in addressing these limitations is the lack of a diverse and sufficient knowledge base. To fill the gap, we construct MedAtlas, which comprises broad multimodal report repositories and rich text corpora. The report repositories contain image-text reports in radiology, ophthalmology, and pathology. The text corpora are compiled from research articles, Wikipedia entries, medical textbooks, clinical guidelines, and knowledge graphs.

Building upon MedAtlas, we propose HeteroRAG, a framework designed to significantly enhance the factual accuracy and reliability of Med-LVLMs. As illustrated in Figure 1, we develop the Heterogeneous Retrieval Module (HRM), which integrates Modality-specific CLIPs (ModCLIPs) and a Multi-corpora Query Generator (MQG). ModCLIPs are trained on large-scale data to ensure effective cross-modality report retrieval. The MQG module is trained in two stages to capture corpus-specific characteristics and generate tailored queries. Finally, we propose the Heterogeneous Knowledge Preference Tuning (HKPT) method to achieve two types of alignment: (1) cross-modality alignment, which aligns visual inputs with retrieved textual content; and (2) multi-source knowledge alignment, which aligns the model's internal knowledge with external knowledge from diverse sources.

We evaluate HeteroRAG on medical visual question answering and report generation tasks across 3 modalities and 11 datasets. Empirical results show that our framework achieves state-of-the-art performance on most benchmarks, demonstrating its strong factuality and reliability. Notably, HeteroRAG surpasses public Med-LVLMs, which contain 4–5× parameters, highlighting the value of effective knowledge integration and alignment.

Our contributions are summarized as follows:

- We introduce MedAtlas, a newly curated comprehensive medical database that provides rich multimodal knowledge for Med-LVLMs and establishes a robust foundation for medical MMRAG research.
- Leveraging MedAtlas, we propose HeteroRAG, a novel medical MMRAG framework that performs accurate heterogeneous knowledge retrieval and fine-grained knowledge alignment.
- Extensive experiments validate HeteroRAG's capability to precisely retrieve and effectively integrate multi-source knowledge, demonstrating SOTA performance across most benchmarks. The framework also consistently outperforms substantially larger Med-LVLMs and establishes a trustworthy and reliable foundation for medical applications.

2 Related Work

2.1 Report Retrieval in Medical MMRAG

Existing medical MMRAG approaches primarily utilize the medical images to retrieve relevant reports (He et al., 2024; Sun et al., 2025; Xia et al., 2024, 2025). For instance, FactMM-RAG (Sun et al., 2025) enhances report generation by incorporating high-quality reference reports. Similarly, RULE (Xia et al., 2024) and MMed-RAG (Xia et al., 2025) integrate reference reports and employ preference fine-tuning to improve model utilization of retrieved reports. Although these approaches improve the factual accuracy of responses, they neglect the retrieval of medical documents, which are crucial for Med-LVLM's reliable inference.

2.2 Document Retrieval in Medical MMRAG

Acknowledging the limitations of report-only retrieval, recent medical MMRAG studies have increasingly emphasized medical documents as knowledge sources (Choi et al., 2025; Shaaban et al., 2025; Wu et al., 2025; Hamza et al., 2025).

Among them, MKGF (Wu et al., 2025) and K-LLaVA (Hamza et al., 2025) both employ multimodal retrievers to fetch documents from the database, aiming to mitigate hallucination issues in language models. ChatCAD+ (Zhao et al., 2024b) and MIRA (Wang et al., 2025) utilize a zero-shot query rewriting module for retrieval. Nevertheless, these retrieval methods overlook the substantial content differences among various corpora, lacking corpus-specific retrieval mechanisms. While corpus-specific query generation has shown promise in text-only RAG (Chen et al., 2025), such methods lack cross-modal perception of visual evidence, and applying them to multimodal clinical scenarios remains an open challenge.

3 MedAtlas Knowledge Base

The MedAtlas knowledge base comprises comprehensive multimodal report repositories covering three modalities and rich textual corpora from five distinct sources.

3.1 Multimodal Report Repository

Existing medical image-report repositories are limited in scale and diversity, typically being derived from the training splits of a few report generation datasets (Sun et al., 2025; Xia et al., 2024, 2025). To address this issue, we collect image-report pairs from a wide range of datasets. Specifically, the Radiology subset includes 1,104,313 pairs from 6 datasets; the Ophthalmology subset includes 111,991 pairs from 5 datasets; and the Pathology subset includes 1,514,058 pairs from 5 datasets. To ensure data quality, duplicate pairs are removed using the image perceptual hashing algorithm (Du et al., 2020). More details are provided in Appendix B.1. For the retrieval method, we use images as queries and reports in the library as keys to retrieve the top-k reports, following Sun et al. (2025); Xia et al. (2024, 2025).

3.2 Textual Corpora

To ensure the richness, we collect corpora from representative sources, following Xiong et al. (2024); Chen et al. (2025). The **Research** corpus is drawn from the 2025 PubMed Annual Baseline. The **Wiki** corpus is collected from the Wikipedia dumps. The **Book** corpus contains E-books, MedQA Textbook, and StatPearls, providing foundational medical knowledge (Xiong et al., 2024; Fan et al., 2025; Chen et al., 2025). The **Guideline** corpus contains clinical guidelines crawled from authoritative websites following Chen et al. (2023). For the above four corpora, they are chunked into chunks of no more than 1000 characters, with an overlap of 200 characters following Xiong et al. (2024).

For the retrieval of unstructured corpora, each query is formatted as "query", and the MedCPT models (Jin et al., 2023) are used for vector search and reranking. For the structured Graph corpus, each query is formatted as "query_term, query_relation". Given the "query_term", its definition and all one-hop relationships are retrieved, followed by filtering relevant relationships by reranking with "query_relation" (Yang et al., 2024).

4 HeteroRAG Framework

In this section, we present the HeteroRAG framework, as illustrated in Figure 2. First, we introduce Modality-specific CLIPs (ModCLIPs), which are trained on large-scale image-text pairs for accurate report retrieval. Next, a Multi-corpora Query Generator (MQG) is developed to enable tailored retrieval for multimodal questions based on corpus characteristics. Finally, we propose a Heterogeneous Knowledge Preference Tuning (HKPT) method to realize cross-modality and multi-source knowledge alignment.

4.1 Modality-specific CLIPs

The ModCLIPs are initialized from BiomedCLIP (Zhang et al., 2023). For each modality, the report retrieval base is independently split into training, development, and test sets to fine-tune CLIP models, following Xia et al. (2024, 2025). Specifically, all samples of each modality are randomly split into 2000 development samples, 2000 test samples, and the remainder for training. This results in 1.10M image-text training pairs in radiology, 0.11M in ophthalmology, and 1.51M in pathology. Contrastive learning (Radford et al., 2021) is performed on single-modality image-text pairs for each ModCLIP. Compared to previous work (Sun et al., 2025; Xia et al., 2024, 2025), which relied solely on training splits from a limited number of datasets, the significantly scaled-up training data enables more accurate cross-modal report retrieval.

4.2 Multi-corpora Query Generator

For each multimodal question including the image v and text question t, the module generates query set for each corpus $Q = \{(i, j, q_j^i) \mid i =$

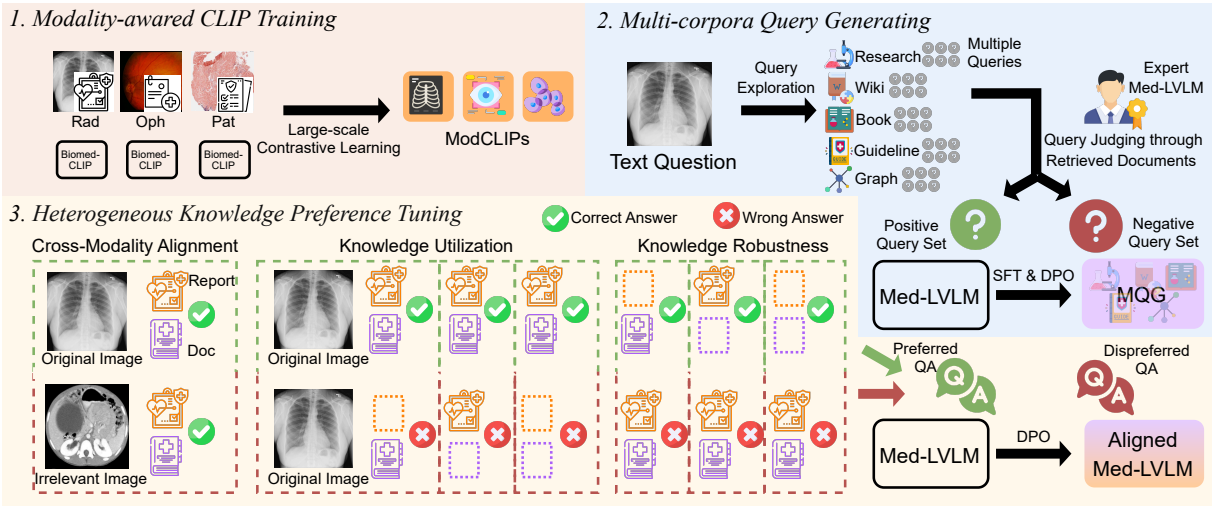

Figure 2: Overview of HeteroRAG framework. It introduces the Modality-specific CLIPs for effective report retrieval. Then, the Multi-corpora Query Generator is developed for tailored retrieval for different corpora. Finally, HKPT is conducted to achieve the cross-modality and multi-source knowledge alignment.

$1, 2, ..., N_C,\ j = 1, 2, ..., N_q^i\}$, where q_j^i denotes the j_{th} query for the i_{th} corpus, N_C denotes the number of corpora, and N_q^i denotes the number of queries for the i_{th} corpus. Each query is then used to retrieve documents that collectively support answering (v, t). Inspired by and extending prior work on text-only tailored query generation (Chen et al., 2025), we design the following pipeline for the medical MMRAG tasks.

Since annotations for documents supporting medical multimodal questions are generally unavailable, we use the LVLMs to generate proxy labels as inspired by Gu et al. (2024); Li et al. (2025b); Chen et al. (2025). We begin with a query exploration phase to identify potential retrieval strategies. Lingshu-32B (Xu et al., 2025) is selected as it consistently achieves strong performances across medical vision language tasks. We prompt the expert Med-LVLM to generate multiple queries for each source, with the prompt shown in Prompt D.3. The prompts are designed to encourage intra-corpus diversity and align with the characteristics of corpora. To control the cost, the number of exploration queries per corpus is fixed to 6.

Subsequently, the same expert model evaluates the documents retrieved by each query by judging whether they support the reference answer, with the prompt shown in Prompt D.4. To evaluate the annotation quality, manual judging is conducted on a 500-item subset[2] by medical researchers, also following the instructions in Prompt D.4. They were informed that the results would be used only for research, and their participation was voluntary. The results show that Lingshu-32B achieves an accuracy of 0.836 and an F1 score of 0.855 against manual expert judgments, demonstrating the reliability of VLM-as-a-judge. Based on Lingshu-32B's judgments, queries are categorized as either positive, denoted q_w, or negative, denoted q_l.

For each corpus, we select up to N_q^i instances of q_w and q_l to form positive queries Q_w and negative queries Q_l, respectively. A two-stage training strategy is applied to MQG. First, supervised fine-tuning (SFT) is performed:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(v,t,Q_w)\sim\mathcal{D}_w} \log \mathcal{M}_\theta \left(Q_w \mid v, t\right). \quad (1)$$

Then, direct preference optimization (DPO) is applied to further align the retrieval strategies with corpora:

$$\mathcal{L}_{\text{DPO}}(\mathcal{M}_\theta; \mathcal{M}_{\text{ref}}) = -\mathbb{E}_{(v,t,Q_w,Q_l)\sim\mathcal{D}_{wl}} \left[\log \sigma \left(\beta \log \frac{\mathcal{M}_\theta(Q_w|v,t)}{\mathcal{M}_{\text{ref}}(Q_w|v,t)} - \beta \log \frac{\mathcal{M}_\theta(Q_l|v,t)}{\mathcal{M}_{\text{ref}}(Q_l|v,t)}\right)\right]. \quad (2)$$

4.3 Heterogeneous Knowledge Preference Tuning

Despite retrieving relevant reports and reliable documents, Med-LVLMs still suffer from severe knowledge misalignment issues. Inspired by RULE (Xia et al., 2024) and MMed-RAG (Xia

[2]A sample size of 500 was determined using Cochran's Formula, yielding a margin of error of ±4.38% at a 95% confidence level.

Algorithm 1: Heterogeneous Knowledge Preference Tuning (HKPT)

Input: $\mathcal{D} = \{v^i, t^i, K^i, y^i\}_{i=1}^N$: Training dataset; $K = \{k_r, k_d\}$: Retrieved knowledge; $\mathcal{M}_\theta$: Med-LVLM; $\mathcal{D}_{cm}, \mathcal{D}_{mk}$: Preference datasets.
Output: $\mathcal{M}$: Preference tuned model.

Initialize $\mathcal{D}_{cm}, \mathcal{D}_{mk}$ with empty sets
foreach $(v, t, K, y) \in \mathcal{D}$ **do**
  Retrieve the image v^* irrelevant to v
  ▷ `Cross-Modality Alignment`
  if $\mathcal{M}(v, t, K) = y$ *and* $\mathcal{M}(v^*, t) \neq y$ *and* $\mathcal{M}(v^*, t, K) = y$ **then**
    $x_w \leftarrow (v, t, K)$; $y_w \leftarrow y$
    $x_l \leftarrow (v^*, t, K)$; $y_l \leftarrow \mathcal{M}(v^*, t, K)$
    Put $\{x_w, x_l, y_w, y_l\}$ into $\mathcal{D}_{cm}$
  ▷ `Multi-Source Knowledge Alignment`
  foreach $k \in \{\{k_r\}, \{k_d\}, \{k_r, k_d\}\}$ **do**
    ▷ `Knowledge Utilization`
    if $\mathcal{M}(v, t, K) = y$ *and* $\mathcal{M}(v, t, K_{\setminus k}) \neq y$ **then**
      $x_w \leftarrow (v, t, K)$; $y_w \leftarrow y$
      $x_l \leftarrow (v, t, K)$; $y_l \leftarrow \mathcal{M}(v, t, K_{\setminus k})$
      Put $\{x_w, x_l, y_w, y_l\}$ into $\mathcal{D}_{mk}$
    ▷ `Knowledge Robustness`
    if $\mathcal{M}(v, t, K_{\setminus k}) = y$ *and* $\mathcal{M}(v, t, K) \neq y$ **then**
      $x_w \leftarrow (v, t, K)$; $y_w \leftarrow y$
      $x_l \leftarrow (v, t, K)$; $y_l \leftarrow \mathcal{M}(v, t, K)$
      Put $\{x_w, x_l, y_w, y_l\}$ into $\mathcal{D}_{mk}$
foreach $(x_w, x_l, y_w, y_l) \in \mathcal{D}_{cm} \cup \mathcal{D}_{mk}$ **do**
  Compute the loss and update $\mathcal{M}$ following Eq. 3

et al., 2025), which introduce the preference fine-tuning strategy for aligning Med-LVLMs with external reports, we propose Heterogeneous Knowledge Preference Tuning (HKPT) to enable alignment with knowledge from more sources. The HKPT process is detailed in Algorithm 1.

Cross-Modality Alignment. The incorporation of external knowledge may cause Med-LVLM to ignore visual information and directly copy retrieved contents (Xia et al., 2025). To mitigate this, we construct preference pairs from the training set to improve modality alignment. Each training sample is denoted as $\{v, t, K, y\}$, where v is the medical image, t is the text question, K is the retrieved knowledge (including reports k_r and documents k_d), and y is the gold answer. For each v, we retrieve the least similar image based on the Mod-CLIP image feature from the same modality training samples as an irrelevant image v^*. Preferred responses are selected when $\mathcal{M}$ correctly answers using v, while dispreferred ones are selected when $\mathcal{M}$ correctly answers using irrelevant v^*, indicating that $\mathcal{M}$ ignores v and relies solely on K. For open-ended generation tasks, correctness is defined as the average metric exceeding a threshold α_r. The criterion also applies below. This process forms the preference dataset $\mathcal{D}_{cm}$.

Multi-Source Knowledge Alignment. To improve $\mathcal{M}$'s alignment with external knowledge K, which includes reports k_r and documents k_d, we design preference pairs from two aspects: **knowledge utilization and robustness**. Taking k_r as an example: For knowledge utilization, preferred responses are selected when $\mathcal{M}$ correctly answers by properly using k_r, while dispreferred ones are selected when $\mathcal{M}$ fails without k_r. For knowledge robustness, preferred responses are selected when $\mathcal{M}$ correctly answers without k_r, while dispreferred ones are selected when $\mathcal{M}$ misuses k_r and produces incorrect answers. The dual-aspect strategy is also applied to k_d, and a combination of k_r and k_d, ensuring fine-grained alignment across all knowledge sources.

The resulting D_{mk}, together with D_{cm}, are employed in HKPT, enabling unified alignment across modalities and knowledge sources:

$$\mathcal{L}_{\text{HKPT}}(\mathcal{M}_{\theta'}; \mathcal{M}_{\text{ref}'}) = -\mathbb{E}_{(x_w, x_l, y_w, y_l) \sim \mathcal{D}_{cm} \cup \mathcal{D}_{mk}} \left[\log \sigma \left(\beta \log \frac{\mathcal{M}_{\theta'}(y_w|x_w)}{\mathcal{M}_{\text{ref}'}(y_w|x_w)} - \beta \log \frac{\mathcal{M}_{\theta'}(y_l|x_l)}{\mathcal{M}_{\text{ref}'}(y_l|x_l)} \right) \right]. \tag{3}$$

5 Experiments

5.1 Experimental Setups

Datasets and Metrics. The medical VQA datasets include VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021), OMVQA-Rad (Hu et al., 2024), DME-VQA (Tascon-Morales et al., 2022), OMVQA-Oph (Hu et al., 2024), PathMMU (Sun et al., 2024a), PathVQA (He et al., 2020), and Quilt-VQA (Seyfioglu et al., 2024). Medical report generation datasets include MIMIC-CXR (Johnson et al., 2019), IU-Xray (Demner-Fushman et al., 2015), and Harvard-FairVLMed (Luo et al., 2024). We have excluded any retrieval report samples overlapping with these datasets using the image perceptual hashing and carefully checked to ensure no overlap between them. This guarantees that the dataset samples are unseen during Mod-CLIPs' training and prevents the retrieval results from containing instances identical to the samples. **Note that the performance on OMVQA-Rad, OMVQA-Oph, and Quilt-VQA can be seen as out-of-distribution results, as they do not include a training split.** Additional dataset details are provided in Appendix B.2.

Methods	Retrieval	Radiology			Ophthalmology		Pathology		
		VQA-RAD	SLAKE	OMVQA-Rad†	DME-VQA	OMVQA-Oph†	PathMMU	PathVQA	Quilt-VQA†
Original	-	72.79	83.65	74.92	81.92	80.83	57.36	77.38	49.27
Beam Search	-	73.16	83.17	75.17	81.92	80.75	57.69	76.76	45.77
DoLa	-	76.47	82.21	73.67	79.33	80.17	55.35	80.86	66.76
VCD	-	71.69	81.73	73.42	80.32	80.42	56.86	75.05	52.77
AVISC	-	73.16	83.65	75.50	81.62	81.33	57.69	77.41	50.15
M3ID	-	72.79	82.93	74.83	81.85	80.83	57.53	76.59	45.77
MedDr	Report	73.16	83.65	75.08	79.71	79.50	61.20	77.74	53.35
FactMM-RAG	Report	76.84	83.89	75.58	81.92	81.50	73.58	<u>91.98</u>	<u>69.68</u>
RULE	Report	73.16	84.38	74.67	82.61	79.50	65.72	81.95	60.35
MMed-RAG	Report	75.74	<u>86.06</u>	76.33	80.70	79.08	68.06	85.67	67.35
MKGF	Doc	74.63	84.86	74.25	82.07	82.33	66.22	80.06	58.02
K-LLaVA	Doc	<u>77.21</u>	84.62	76.00	<u>88.48</u>	<u>83.75</u>	73.75	87.76	61.81
MIRA	Report+Doc	76.84	84.38	<u>76.58</u>	87.95	82.50	<u>74.25</u>	**92.10**	68.80
HeteroRAG (Ours)	Report+Doc	**81.99**	**87.50**	**80.42**	**88.56**	**86.00**	**75.59**	90.83	**72.89**

Table 1: Model performance of different methods based on Lingshu-7B on the medical VQA task. The best results and second-best results are highlighted in **bold** and <u>underlined</u>, respectively. †: out-of-distribution datasets.

For evaluation metrics, accuracy is used for medical VQA tasks. Radiology report generation is evaluated using BLEU[3] (Papineni et al., 2002), ROUGE-L (Lin, 2004), and RaTEScore (Zhao et al., 2024a), while ophthalmology reports are evaluated using BLEU, ROUGE-L, and METEOR[4] (Banerjee and Lavie, 2005).

Implementation details regarding report and corpus retrieval, as well as the training of the HeteroRAG framework, are provided in Appendix B.4.

Baselines. Four categories of baselines are introduced: (1) decoding-based methods aiming for improving factuality including Beam Search (Sutskever et al., 2014), DoLa (Chuang et al., 2024), VCD (Leng et al., 2024), AVISC (Woo et al., 2024), and M3ID (Favero et al., 2024); (2) report-retrieval methods including MedDr (He et al., 2024), FactMM-RAG (Sun et al., 2025), RULE (Xia et al., 2024), and MMed-RAG (Xia et al., 2025); (3) document-retrieval methods including MKGF (Wu et al., 2025) and K-LLaVA (Hamza et al., 2025) and (4) a more recent work that retrieves both reports and documents, MIRA (Wang et al., 2025). **To ensure fair comparison, retrievable reports and documents remain consistent across all baselines. Medical CLIPs for report retrieval also remain consistent across all baselines, with the impact of CLIP training data analyzed separately in Section 5.4.** We also introduce widely-used Med-LVLMs: LLaVA-Med-7B (Li et al., 2023), MedGemma-4B (Sellergren et al., 2025), HuatuoGPT-V-34B (Chen et al., 2024a), HealthGPT-32B (Lin et al., 2025), and Lingshu-32B (Xu et al., 2025). More baseline details are shown in Appendix B.3.

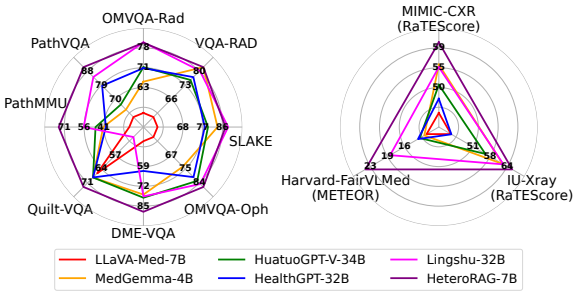

Figure 3: Comparison of HeteroRAG with other Med-LVLMs. Effective retrieval and fine-grained integration of knowledge enables the HeteroRAG to surpass larger Med-LVLMs with greater parameter efficiency.

5.2 Main Results

The experimental results of different methods based on Lingshu-7B are presented in Table 1 and Table 4. A comparison between widely used Med-LVLMs and HeteroRAG is illustrated in Figure 3. These results lead to the following key observations: **(1) Effectiveness of incorporating multi-source knowledge:** HeteroRAG achieves superior performance compared to approaches under different retrieval settings. This demonstrates our effectiveness in retrieving and integrating heterogeneous knowledge. Highly relevant reports enhance the factual accuracy of Med-LVLMs, while evidence documents improve their reliability. **(2) Generaliz-**

[3]In this work, BLEU refers to the average of the BLEU-1, BLEU-2, BLEU-3, and BLEU-4 scores.

[4]RaTEScore is not used for evaluating ophthalmology report generation, as it is specifically for radiology.

Methods	OMVQA-Rad	OMVQA-Oph	Quilt-VQA
Original	74.92	80.83	49.27
SFT	75.00	82.17	63.85
HeteroRAG	**80.42**	**86.00**	**72.89**
w/o Reports	75.08	81.33	62.97
w/o Doc	74.17	79.08	53.94
w/o Research	79.42	84.33	68.80
w/o Wiki	77.75	84.00	67.64
w/o Book	75.17	80.92	67.06
w/o Guideline	79.33	84.58	69.68
w/o Graph	78.58	83.25	66.47

Table 2: Performance comparison of HeteroRAG against Original and SFT baselines, along with ablation results on various knowledge sources and corpora.

Models	Rad.	Oph.	Pat.
CLIP	6.45	2.90	5.25
Jina-embeddings-v4	8.25	4.85	6.25
VLM2Vec-V2-2B	6.00	3.05	4.45
GME-Qwen2VL-7B	15.90	7.80	11.95
Ops-MM-embedding-v1-7B	16.20	9.80	15.20
Seed-1.6-embedding	16.00	7.95	12.30
BiomedCLIP	30.20	13.45	28.85
PMC-CLIP	30.00	19.80	23.35
PubMedCLIP*	13.35	-	-
MM-Retinal*	-	4.65	-
QuiltNet*	-	-	39.65
FactMM-RAG*	44.25	-	-
RULE*	31.80	19.25	-
MMed-RAG*	31.80	19.25	30.20
ModCLIPs*	**79.40**	**47.55**	**77.35**

Table 3: Image-to-text recall@5 of different retrievers. The asterisks (*) denote the modality-specific retrievers.

ability of our framework: HeteroRAG achieves the best performance on nearly all datasets across three modalities. Notably, this superiority holds not only for closed-ended VQA tasks but also for open-ended report generation and OOD datasets, which requires more sophisticated multimodal understanding and generation capabilities. **(3) Superiority over larger Med-LVLMs:** Figure 3 shows that HeteroRAG, with a 7B parameter size, outperforms most advanced Med-LVLMs, which contain 4 to 5 times more parameters across multiple datasets. This indicates that the proposed framework advances the medical multimodal capabilities of existing Med-LVLMs to a higher level.

5.3 Effectiveness of Retrieved Knowledge

We conduct ablation studies to evaluate the contribution of knowledge sources, as shown in Table 2. The "Original" and "SFT" settings represent the performance of the original Lingshu-7B and Lingshu-7B after SFT on the original training set, which does not include reports and documents. The other configurations examine HeteroRAG's performance when either reports or documents are removed. The results show that retrieved knowledge significantly improves Med-LVLM's performance compared to the Original baseline. The performance improvements from supervised fine-tuning alone are insufficient to compensate for the absence of knowledge. When reports or documents are excluded, the performance degradation confirms that both sources are important for HeteroRAG's knowledge-intensive inference. Furthermore, all five corpora contribute to Med-LVLM's capacity.

5.4 Effectiveness of ModCLIPs

We evaluate ModCLIPs against other retrievers on image-to-text report retrieval tasks, as shown in Table 3. Open-domain retrievers include CLIP (Radford et al., 2021), Jina-embeddings-v4 (Günther et al., 2025), VLM2Vec-V2-2B (Meng et al., 2025), GME-Qwen2VL-7B (Zhang et al., 2024), Ops-MM-embedding-v1-7B (Alibaba Cloud OpenSearch-AI Team, 2025), and Seed-1.6-embedding (ByteDance SEED Team, 2025). Generalist medical retrievers include BiomedCLIP and PMC-CLIP (Lin et al., 2023). Modality-specific medical retrievers include PubMedCLIP (Eslami et al., 2023), MM-Retinal (Wu et al., 2024b), QuiltNet (Ikezogwo et al., 2023), FactMM-RAG, RULE and MMed-RAG. For FactMM-RAG, RULE, and MMed-RAG, they are reproduced by fine-tuning BiomedCLIP on their data.

Using the test set described in Section 4.1 with recall@5 as our evaluation metric, our experiments demonstrate that ModCLIPs consistently outperform competing methods across all three modalities. This superior performance can be attributed to two key advantages: (1) Single-modality training yields significantly better modality-specific understanding compared to mixed-modality approaches, and (2) our training data offers more comprehensive coverage and greater diversity within each modality.

5.5 Effectiveness of MQG

We further investigate the effectiveness of MQG in Table 5. First, the MQG in HeteroRAG is replaced with a CLIP retrieval module. Specifically, for each medical visual question, the ModCLIPs are employed to retrieve documents through both image-to-text and text-to-text retrieval. The two re-

Methods	Retrieval	Radiology						Ophthalmology		
		MIMIC-CXR			IU-Xray			Harvard-FairVLMed		
		BLEU	ROUGE-L	RaTEScore	BLEU	ROUGE-L	RaTEScore	BLEU	ROUGE-L	METEOR
Original	-	10.31	30.39	53.30	18.50	41.00	57.95	4.21	14.30	15.75
Beam Search	-	10.46	30.04	50.02	24.35	42.77	65.31	2.78	11.68	13.74
DoLa	-	10.05	30.22	52.99	18.85	40.81	58.01	5.05	16.24	18.35
VCD	-	11.49	31.41	53.44	21.50	40.76	63.03	4.52	13.97	16.63
AVISC	-	12.78	32.59	54.14	22.16	41.33	62.00	3.58	12.19	14.91
M3ID	-	13.13	32.47	54.44	22.17	41.95	62.38	3.64	12.61	14.22
MedDr	Report	16.63	33.92	56.32	23.04	41.92	63.00	6.98	18.91	19.96
FactMM-RAG	Report	16.94	36.03	56.86	22.74	42.93	60.70	9.42	22.58	18.40
RULE	Report	16.49	34.02	56.47	23.95	42.69	63.51	8.39	20.94	19.91
MMed-RAG	Report	17.10	35.48	57.76	24.36	42.96	64.15	8.17	20.70	19.64
MKGF	Doc	11.56	32.33	53.66	19.97	41.32	59.64	5.87	16.18	16.83
K-LLaVA	Doc	16.51	35.41	56.07	20.27	41.41	57.48	9.27	22.73	18.38
MIRA	Report+Doc	16.71	36.06	57.02	19.70	41.06	57.29	9.57	22.96	18.81
HeteroRAG (Ours)	Report+Doc	**20.49**	**39.20**	**60.73**	**27.88**	**46.18**	**65.92**	**10.69**	**24.62**	**23.63**

Table 4: Model performance of different methods based on Lingshu-7B on the medical report generation task. The best results and second-best results are highlighted in **bold** and underlined, respectively.

Methods	OMVQA-Rad	OMVQA-Oph	Quilt-VQA
CLIP	75.67	83.92	66.76
MQG	**80.42**	**86.00**	**72.89**
w/o DPO	78.33	83.92	69.10
w/o SFT	75.67	82.17	65.89

Table 5: Performance of HeteroRAG under two ablation settings: replacing MQG with CLIP-based retrieval, and removing the training stages of MQG.

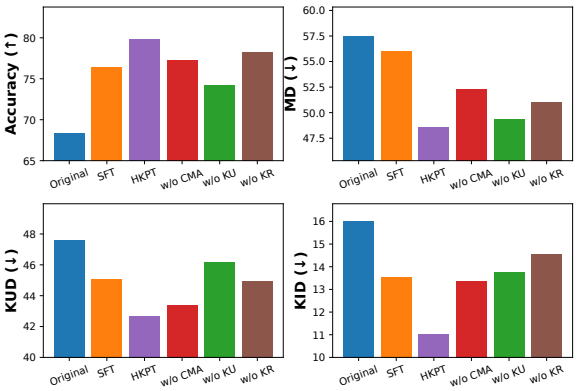

Figure 4: Accuracy and disalignment metrics of Lingshu-7B trained with different methods and data.

trieval results are combined using RRF. We also ablate the DPO and SFT training stages of the MQG. The retrieval quality is implicitly evaluated through the downstream performance of Med-LVLMs. We adopt this approach because multimodal medical questions typically lack gold document labels, making classic metrics like MRR or nDCG infeasible. Furthermore, manual assessment of retrieved documents for every query across all baseline methods is impractical due to the large scale of data. Our experiments demonstrate that MQG retrieves more relevant documents compared to standard CLIP methods. This improvement can be attributed to better alignment of MQG and each corpus's characteristics. Furthermore, both the SFT and DPO training stages prove essential in developing MQG.

5.6 Alignment Effectiveness of HKPT

To evaluate the alignment effectiveness of HKPT, we introduce three metrics besides answer accuracy: Modality Disalignment (MD), Knowledge Usage Disalignment (KUD), and Knowledge Interference Disalignment (KID). MD corresponds to CMA in Section 4.3, KUD corresponds to KU, and KID corresponds to KR. MD measures the proportion that the Med-LVLM directly references the retrieved knowledge and correctly answers with the irrelevant image among cases where it originally fails with the irrelevant image. KUD measures the proportion that the Med-LVLM fails when retrieved contents (reports and documents) are introduced among cases where it fails without retrieval. KID measures the proportion that the Med-LVLM fails when retrieved contents are introduced among cases where it succeeds without retrieval.

Figure 4 shows the average metrics on OMVQA-Rad, OMVQA-Oph, and Quilt-VQA. The "SFT" method refers to SFT using the training dataset with documents and reports added. The results

Models	OMVQA-Rad	OMVQA-Oph	Quilt-VQA
LLaVA-Med-7B	53.67	56.83	66.18
+ HeteroRAG	**59.67**	**68.42**	**69.10**
HuatuoGPT-V-7B	72.08	81.83	66.18
+ HeteroRAG	**76.92**	**84.33**	**70.55**
Lingshu-7B	74.92	80.83	49.27
+ HeteroRAG	**80.42**	**86.00**	**72.89**

Table 6: Model performance when the HeteroRAG framework is applied to different Med-LVLMs.

demonstrate that HKPT improves overall accuracy compared to both the original and SFT models. Moreover, all three types of disalignment are significantly reduced by the HKPT method. We further conduct ablation studies on each type of preference pair in HKPT, including CMA, KU, and KR. The results confirm that each component contributes effectively to both question answering performance and overall knowledge alignment. Moreover, each component enhances its corresponding alignment capability as expected.

5.7 Compatibility Analysis

To analyze the compatibility of the HeteroRAG framework with different Med-LVLMs, we apply it to LLaVA-Med-7B and HuatuoGPT-V-7B besides Lingshu-7B. Specifically, the ModCLIPs and MQG in HRM are kept unchanged, as they are universal across different downstream readers. The HKPT process is performed separately for each Med-LVLM. Results in Table 6 show that HeteroRAG brings consistent improvements over all Med-LVLMs. This shows the compatibility of the HeteroRAG framework and indicates that HeteroRAG can be transferred to diverse Med-LVLMs.

6 Conclusion

This work addresses the critical challenges of effective retrieval and multi-aspect alignment for heterogeneous knowledge in the Medical MMRAG field. MedAtlas provides a rich, multi-source knowledge base for medical multimodal tasks. The HeteroRAG framework enables precise report retrieval and multi-corpus retrieval, followed by the alignment of heterogeneous knowledge sources. Extensive experiments demonstrate that our framework achieves state-of-the-art performance across multiple medical VQA and report generation benchmarks. Our work paves the way for effectively integrating multi-source knowledge, advancing the reliability of Med-LVLMs in clinical scenarios.

Limitations

This work focuses on improving the factual accuracy of Med-LVLMs through effective retrieval and aggregation from heterogeneous medical knowledge sources. Our current framework adopts a one-step retrieval approach. Future work may explore multi-round tool usage and reasoning, such as in OpenAI o3-style systems, to further scale up reasoning capabilities.

Moreover, our primary goal is to enhance factual correctness. Other aspects of practical Med-LVLM deployment, such as fairness, privacy, and safety, are not the focus of this study and remain to be investigated in future work.

Ethical Consideration

The multimodal reports in MedAtlas, including radiology, ophthalmology, and pathology subsets, are publicly available and licensed for academic research. For the text corpora in MedAtlas, Research articles, Wiki content, MedQA Textbook, StatPearls, selected clinical guidelines, and Graph data are openly accessible. We will provide access instructions to these publicly available subsets. The remaining materials, including E-books and certain clinical guidelines, can not be directly distributed in their raw form due to data licensing restrictions. However, detailed lists will be publicly available for researchers to obtain these resources.

The medical visual question answering (VQA) and report generation datasets used in this study are publicly available and widely adopted in the research community. The proposed HeteroRAG framework significantly improves the accuracy and reliability of model responses, thereby enhancing the trustworthiness of multimodal medical AI systems. We emphasize responsible data use and follow ethical guidelines for research with publicly available, non-sensitive information.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62576209) and STCSM (No. 2025SHZDZX025G05).

References

Alibaba Cloud OpenSearch-AI Team. 2025. Ops-mm-embedding-v1-7b. Hugging Face Model Hub.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie

Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72. Association for Computational Linguistics.

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, 32(Database-Issue):267–270.

ByteDance SEED Team. 2025. Seed-1.6-embedding. SEED Model Blog, VolcEngine.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P Langlotz. 2024. Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. *arXiv preprint arXiv:2405.19538*.

Junying Chen, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, and Benyou Wang. 2024a. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale. *CoRR*, abs/2406.19280.

Zeming Chen, Alejandro Hernández-Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. MEDITRON-70B: scaling medical pretraining for large language models. *CoRR*, abs/2311.16079.

Zhe Chen, Yusheng Liao, Shuyang Jiang, Pingjie Wang, Yiqiu Guo, Yanfeng Wang, and Yu Wang. 2025. Towards Omni-RAG: Comprehensive retrieval-augmented generation for large language models in medical applications. *arXiv preprint arXiv:2501.02460*.

Zhe Chen, Hongcheng Liu, and Yu Wang. 2024b. Dialogmcf: Multimodal context flow for audio visual scene-aware dialog. *IEEE ACM Trans. Audio Speech Lang. Process.*, 32:753–764.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Kyoyun Choi, Byungmu Yoon, Soobum Kim, and Jonggwon Park. 2025. Leveraging llms for multimodal retrieval-augmented radiology report generation via key phrase extraction. *arXiv preprint arXiv:2504.07415*.

Yun-Wei Chu, Kai Zhang, Christopher Malon, and Martin Renqiang Min. 2025. Reducing hallucinations of medical multimodal large language models with visual retrieval-augmented generation. *CoRR*, abs/2502.15040.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Etienne Decenciere, Guy Cazuguel, Xiwei Zhang, Guillaume Thibault, J-C Klein, Fernand Meyer, Beatriz Marcotegui, Gwénolé Quellec, Mathieu Lamard, Ronan Danno, and 1 others. 2013. Teleophta: Machine learning and image processing methods for teleophthalmology. *Irbm*, 34(2):196–203.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2015. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Ling Du, Anthony T. S. Ho, and Runmin Cong. 2020. Perceptual hashing for image authentication: A survey. *Signal Process. Image Commun.*, 81.

Sedigheh Eslami, Christoph Meinel, and Gerard De Melo. 2023. Pubmedclip: How much does clip benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1151–1163.

Run-Ze Fan, Zengzhi Wang, and Pengfei Liu. 2025. Megascience: Pushing the frontiers of post-training datasets for science reasoning. *Preprint*, arXiv:2507.16812.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. 2024. Multi-modal hallucination control by visual information grounding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 14303–14312. IEEE.

Jevgenij Gamper and Nasir Rajpoot. 2021. Multiple instance captioning: Learning representations from histopathology textbooks and articles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16549–16559.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. 2024. A survey on llm-as-a-judge. *CoRR*, abs/2411.15594.

Michael Günther, Saba Sturua, Mohammad Kalim Akram, Isabelle Mohr, Andrei Ungureanu, Bo Wang, Sedigheh Eslami, Scott Martens, Maximilian Werk, Nan Wang, and 1 others. 2025. jina-embeddings-v4: Universal embeddings for multimodal multilingual retrieval. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 531–550.

Ameer Hamza, Abdullah, Yong Hyun Ahn, Sungyoung Lee, and Seong Tae Kim. 2025. Llava needs more knowledge: Retrieval augmented natural language generation with knowledge graph for explaining thoracic pathologies. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 3311–3319. AAAI Press.

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. *CoRR*, abs/2404.15127.

Xuehai He, Yichen Zhang, Luntian Mou, Eric P. Xing, and Pengtao Xie. 2020. Pathvqa: 30000+ questions for medical visual question answering. *CoRR*, abs/2003.10286.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. 2024. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical LVLM. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 22170–22183. IEEE.

Jia-Hong Huang, C-H Huck Yang, Fangyu Liu, Meng Tian, Yi-Chieh Liu, Ting-Wei Wu, I Lin, Kang Wang, Hiromasa Morikawa, Hernghua Chang, and 1 others. 2021. Deepopht: medical report generation for retinal images via deep models and visual explanation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2442–2452.

Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. 2023. Quilt-1m: One million image-text pairs for histopathology. *Advances in neural information processing systems*, 36:37995–38017.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. 2023. MedCPT: contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651.

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*.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13872–13882. IEEE.

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 *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Haolin Li, Tianjie Dai, Zhe Chen, Siyuan Du, Jiangchao Yao, Ya Zhang, and Yanfeng Wang. 2025a. RAD: towards trustworthy retrieval-augmented multi-modal clinical diagnosis. *CoRR*, abs/2509.19980.

Jun Li, Tongkun Su, Baoliang Zhao, Faqin Lv, Qiong Wang, Nassir Navab, Ying Hu, and Zhongliang Jiang. 2024. Ultrasound report generation with cross-modality feature alignment via unsupervised guidance. *IEEE Transactions on Medical Imaging*.

Mingjie Li, Wenjia Cai, Rui Liu, Yuetian Weng, Xiaoyun Zhao, Cong Wang, Xin Chen, Zhong Liu, Caineng Pan, Mengke Li, and 1 others. 2021. FFA-IR: towards an explainable and reliable medical report generation benchmark. In *Thirty-fifth conference*

on neural information processing systems datasets and benchmarks track (round 2).

Yuan Li, Qi Luo, Xiaonan Li, Bufan Li, Qinyuan Cheng, Bo Wang, Yining Zheng, Yuxin Wang, Zhangyue Yin, and Xipeng Qiu. 2025b. R3-RAG: learning step-by-step reasoning and retrieval for llms via reinforcement learning. *CoRR*, abs/2505.23794.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wanggui He, Hao Jiang, Mengze Li, Xiaohui Song, Siliang Tang, Jun Xiao, Hui Lin, Yueting Zhuang, and Beng Chin Ooi. 2025. Healthgpt: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. *CoRR*, abs/2502.09838.

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.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *18th IEEE International Symposium on Biomedical Imaging, ISBI 2021, Nice, France, April 13-16, 2021*, pages 1650–1654. IEEE.

Hongcheng Liu, Zhe Chen, Hui Li, Pingjie Wang, Yanfeng Wang, and Yu Wang. 2024a. MSG-BART: multi-granularity scene graph-enhanced encoder-decoder language model for video-grounded dialogue generation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 10516–10520. IEEE.

Hongcheng Liu, Yusheng Liao, Siqv Ou, Yuhao Wang, Heyang Liu, Yanfeng Wang, and Yu Wang. 2024b. Med-pmc: Medical personalized multi-modal consultation with a proactive ask-first-observe-next paradigm. *arXiv preprint arXiv:2408.08693*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Yan Luo, Min Shi, Muhammad Osama Khan, Muhammad Muneeb Afzal, Hao Huang, Shuaihang Yuan, Yu Tian, Luo Song, Ava Kouhana, Tobias Elze, and 1 others. 2024. FairCLIP: Harnessing fairness in vision-language learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12289–12301.

Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, and 1 others. 2025. Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. *arXiv preprint arXiv:2507.04590*.

NCBI. 2025. Pubmed baseline data. `https://ftp.ncbi.nlm.nih.gov/pubmed/baseline/`.

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, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Mériaudeau. 2018. Indian diabetic retinopathy image dataset (idrid): A database for diabetic retinopathy screening research. *Data*, 3(3):25.

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. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Mercy Prasanna Ranjit, Gopinath Ganapathy, Ranjit Manuel, and Tanuja Ganu. 2023. Retrieval augmented chest x-ray report generation using openai GPT models. In *Machine Learning for Healthcare Conference, MLHC 2023, 11-12 August 2023, New York, USA*, volume 219 of *Proceedings of Machine Learning Research*, pages 650–666. PMLR.

Johannes Rückert, Louise Bloch, Raphael Brüngel, Ahmad Idrissi-Yaghir, Henning Schäfer, Cynthia S Schmidt, Sven Koitka, Obioma Pelka, Asma Ben Abacha, Alba G. Seco de Herrera, and 1 others. 2024. ROCOv2: radiology objects in context version 2, an updated multimodal image dataset. *Scientific Data*, 11(1):688.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Mehmet Saygin Seyfioglu, Wisdom Oluchi Ikezogwo, Fatemeh Ghezloo, Ranjay Krishna, and Linda G. Shapiro. 2024. Quilt-llava: Visual instruction tuning by extracting localized narratives from open-source histopathology videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13183–13192. IEEE.

Mai A Shaaban, Tausifa Jan Saleem, Vijay Ram Papineni, and Mohammad Yaqub. 2025. Motor: Multimodal optimal transport via grounded retrieval in medical visual question answering. *arXiv preprint arXiv:2506.22900*.

StatPearls. 2024. Statpearls. https://www.ncbi.nlm.nih.gov/books/NBK430685/.

Liwen Sun, James Jialun Zhao, Wenjing Han, and Chenyan Xiong. 2025. Fact-aware multimodal retrieval augmentation for accurate medical radiology report generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 643–655.

Yuxuan Sun, Hao Wu, Chenglu Zhu, Sunyi Zheng, Qizi Chen, Kai Zhang, Yunlong Zhang, Dan Wan, Xiaoxiao Lan, Mengyue Zheng, Jingxiong Li, Xinheng Lyu, Tao Lin, and Lin Yang. 2024a. Pathmmu: A massive multimodal expert-level benchmark for understanding and reasoning in pathology. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXII*, volume 15120 of *Lecture Notes in Computer Science*, pages 56–73. Springer.

Yuxuan Sun, Chenglu Zhu, Sunyi Zheng, Kai Zhang, Lin Sun, Zhongyi Shui, Yunlong Zhang, Honglin Li, and Lin Yang. 2024b. Pathasst: A generative foundation ai assistant towards artificial general intelligence of pathology. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5034–5042.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Sergio Tascon-Morales, Pablo Márquez-Neila, and Raphael Sznitman. 2022. Consistency-preserving visual question answering in medical imaging. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2022 - 25th International Conference, Singapore, September 18-22, 2022, Proceedings, Part VIII*, volume 13438 of *Lecture Notes in Computer Science*, pages 386–395. Springer.

Masayuki Tsuneki and Fahdi Kanavati. 2022. Inference of captions from histopathological patches. In *International Conference on Medical Imaging with Deep Learning*, pages 1235–1250. PMLR.

Jinhong Wang, Tajamul Ashraf, Zongyan Han, Jorma Laaksonen, and Rao Mohammad Anwer. 2025. Mira: A novel framework for fusing modalities in medical rag. *arXiv preprint arXiv:2507.07902*.

Wikimedia. 2023. Wikimedia wikipedia. https://huggingface.co/datasets/wikimedia/wikipedia.

Sangmin Woo, Donguk Kim, Jaehyuk Jang, Yubin Choi, and Changick Kim. 2024. Don't miss the forest for the trees: Attentional vision calibration for large vision language models. *CoRR*, abs/2405.17820.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2024a. Pmc-llama: toward building open-source language models for medicine. *J. Am. Medical Informatics Assoc.*, 31(9):1833–1843.

Ruiqi Wu, Chenran Zhang, Jianle Zhang, Yi Zhou, Tao Zhou, and Huazhu Fu. 2024b. Mm-retinal: Knowledge-enhanced foundational pretraining with fundus image-text expertise. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 722–732. Springer.

Yinan Wu, Yuming Lu, Yan Zhou, Yifan Ding, Jingping Liu, and Tong Ruan. 2025. MKGF: A multi-modal knowledge graph based RAG framework to enhance lvlms for medical visual question answering. *Neurocomputing*, 635:129999.

Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. 2025. MMed-RAG: versatile multimodal rag system for medical vision language models. In *The Thirteenth International Conference on Learning Representations*.

Peng Xia, Kangyu Zhu, Haoran Li, Hongtu Zhu, Yun Li, Gang Li, Linjun Zhang, and Huaxiu Yao. 2024. RULE: Reliable multimodal rag for factuality in medical vision language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1081–1093.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 6233–6251. Association for Computational Linguistics.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, and 1 others. 2025. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*.

Rui Yang, Haoran Liu, Edison Marrese-Taylor, Qingcheng Zeng, Yuhe Ke, Wanxin Li, Lechao Cheng, Qingyu Chen, James Caverlee, Yutaka Matsuo, and Irene Li. 2024. Kg-rank: Enhancing large language models for medical QA with knowledge graphs and ranking techniques. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing, BioNLP@ACL 2024, Bangkok, Thailand, August 16, 2024*, pages 155–166. Association for Computational Linguistics.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, and 1 others. 2023. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. Gme: Improving universal multimodal retrieval by multimodal llms. *arXiv preprint arXiv:2412.16855*.

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024a. Ratescore: A metric for radiology report generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15004–15019.

Zihao Zhao, Sheng Wang, Jinchen Gu, Yitao Zhu, Lanzhuju Mei, Zixu Zhuang, Zhiming Cui, Qian Wang, and Dinggang Shen. 2024b. Chatcad+: Toward a universal and reliable interactive cad using llms. *IEEE Transactions on Medical Imaging*, 43(11):3755–3766.

A Additional Analysis

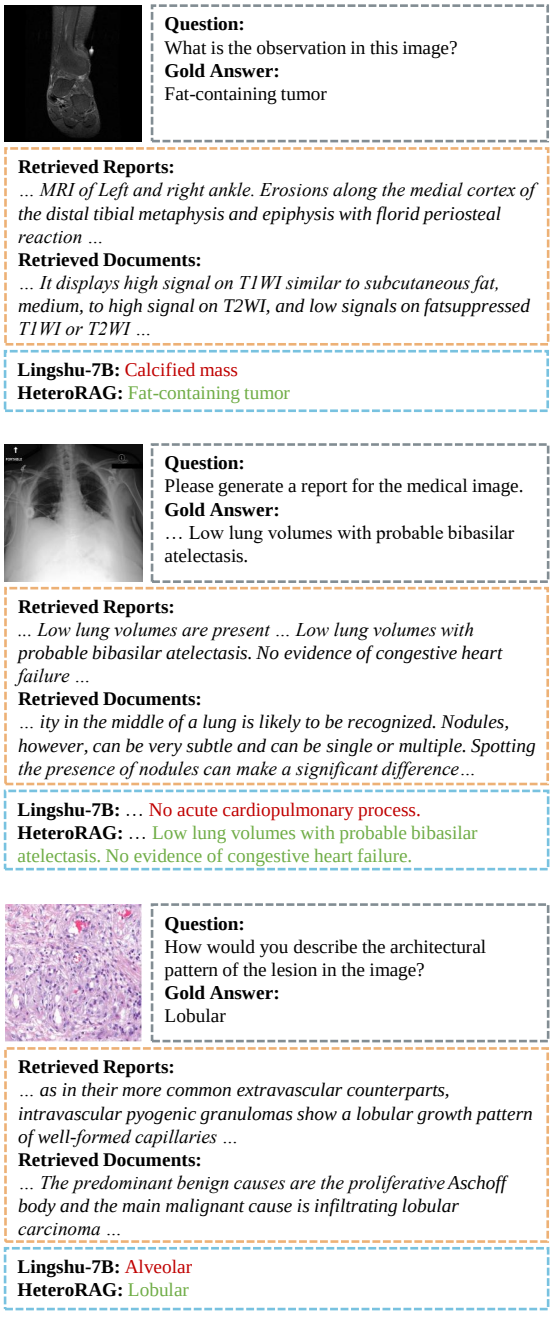

Figure 5: Qualitative analysis I: HeteroRAG effectively leveraging external knowledge.

A.1 Qualitative Analysis

HeteroRAG significantly improves the factual accuracy of Med-LVLM by effectively leveraging medical reports and documents, as shown in Figure 5. In the first case, HeteroRAG utilizes the retrieved document's description of typical MRI signal characteristics of fat-containing tumors to recognize imaging features indicative of fat content in the lesion, thereby supporting the

Question:
What is the term used to describe the anomaly evident in the image?
Gold Answer:
Airspace opacity

Retrieved Reports:
... multifocal patchy ground glass opacity and consolidation in both lungs ... Radiological observations may indicate non-specific infiltrates to nodular and/or cavity formation ...
Retrieved Documents:
... Diagnosis of pleural effusion requires combining clinical evaluation with imaging and laboratory studies, particularly pleural fluid analysis. Thoracentesis is both diagnostic...

Lingshu-7B: Nodule formation
HeteroRAG: Airspace opacity

Question:
is there a mass?
Gold Answer:
Yes

Retrieved Reports:
Bilateral pneumofibrotic changes in the chest X ray ... Postoperative chest radiography. There were no abnormalities on postoperative chest radiography.
Retrieved Documents:
Chest x-ray: A chest x-ray (CXR) is often the initial imaging study performed and may reveal the presence of a mediastinal mass, its location, size, and any associated pulmonary abnormalities...

Lingshu-7B: No
HeteroRAG: Yes

Figure 6: Qualitative analysis II: HeteroRAG resolving conflicting information across multiple knowledge sources.

correct answer. For the second case, HeteroRAG outperforms Lingshu-7B by effectively leveraging retrieved reports. It refers to key phrases: "Low lung volumes are present," and the impression: "No evidence of congestive heart failure." The similar reports enable clinically accurate conclusions for HeteroRAG. For the third case, HeteroRAG effectively leveraged both retrieved contents. Retrieved reports state, "intravascular pyogenic granulomas show a lobular growth pattern of well-formed capillaries," where "lobular growth pattern" directly corresponds to the "architectural pattern". Additionally, the retrieved documents further complement information for "lobular", which is a well-established histopathological architectural pattern. HeteroRAG integrated the multi-source knowledge to confirm "lobular" as the correct answer.

HeteroRAG handles conflicting information acquired from various knowledge sources, as illustrated in Figure 6. In the first case, the image clearly shows airspace opacity, but the retrieved reports contain the misleading term "nodular formation". HeteroRAG, via HKPT, successfully integrated other correct evidence (e.g., "ground glass opacity") and filtered the conflicting "nodular" term, demonstrating robust discrimination in multi-source knowledge competition. The second case features a strong conflict where the image explicitly shows a mass (Answer: Yes), yet some of the retrieved reports state "no abnormalities". HeteroRAG successfully rejected this noise, prioritizing the clear visual evidence of the mass.

A.2 Impact of Retrieved Report Images

Methods	OMVQA-Rad	OMVQA-Oph	Quilt-VQA	Avg.
Original	74.92	80.83	49.27	68.34
HeteroRAG	**80.42**	**86.00**	72.89	**79.77**
+ Ret. Images	77.58	84.33	**73.76**	78.56
+ Blurring	78.92	85.25	72.59	78.92
+ Masking	**80.42**	85.00	72.01	79.14

Table 7: Model performance when the retrieved report images are incorporated.

We further explore the integration of retrieved report images into Med-LVLMs inspired by V-RAG (Chu et al., 2025). Specifically, we incorporate retrieved report images in constructing the preference pairs and training models. Results in Table 7 indicate that adding report images does not improve model performance and even leads to degradation on most datasets.

To further investigate the cause of this degra-

Models	Ret.	OMVQA-Rad	OMVQA-Oph	Quilt-VQA
HuatuoGPT-V-34B	N	71.08	82.42	**68.22**
HealthGPT-32B	N	70.75	80.25	**68.22**
Lingshu-32B	N	**80.58**	**84.25**	48.40
HeteroRAG-7B	N	74.33	80.58	49.85
HeteroRAG-7B	Y	80.42	86.00	72.89

Table 8: Model performance when the Retrieval (Ret.) component of HeteroRAG-7B is dropped.

dation, we conduct test-time perturbation experiments. Perturbation settings are introduced: Blurring, which applies a Gaussian blur to the retrieved images to obscure fine-grained visual details; and Masking, which completely obscures the retrieved images with black masks. Results are also shown in Table 7.

"Blurring" improves the average performance over the "+Ret. Images", and "Masking" further restores it to 79.14, close to HeteroRAG without report images (79.77). This validates our hypothesis that visual information in retrieved report images is largely redundant with the report text and hinders the model's ability to align and integrate external knowledge. Therefore, we exclude retrieved report images in our main methods.

A.3 HeteroRAG's Superiority over Larger-Parameter Models

To investigate why HeteroRAG (built on a 7B backbone) often outperforms 4 to 5 times larger models, we conduct an ablation study by removing its retrieval component. This turns HeteroRAG into a closed-book 7B model. As shown in Table 8, removing the retrieved knowledge causes a significant performance drop, often resulting in performance weaker than the 4–5× larger models. This proves that the gain is directly attributable to heterogeneous knowledge retrieval and is not merely due to the parametric knowledge of the 7B model. Moreover, HeteroRAG operates as an open-book system that leverages precise external non-parametric knowledge to overcome the inherent capacity limitations of its 7B backbone.

A.4 Analysis of Query Diversity and Inter-corpus Similarity

To assess the diversity of queries from different methods, we calculate the ratio of unique n-grams to total generated n-grams across the entire test set. We hope to prevent query collapse, where the model generates the same query term for different samples. The metric is formulated as:

$$\text{Dist-corpus} = \frac{1}{2}\sum_{n=1}^{2}\text{Distinct-}n, \tag{4}$$

$$\text{Distinct-}n = \frac{\text{Count(unique } n\text{-grams)}}{\text{Count(total } n\text{-grams)}}. \tag{5}$$

Table 9 reports the average metric across OMVQA-Rad, OMVQA-Oph, and Quilt-VQA for different baseline methods. Overall, we can observe that MQG significantly improves the diversity of query terms compared to baseline methods.

Furthermore, we also measure the inter-corpus similarity of queries from different methods. In particular, we measure how similar the queries for different sources are to each other for the same input to avoid query redundancy among queries from different sources. MedCPT-Query-Encoder is used to encode the query for each source and calculate the average cosine similarity of each source's query with the other four sources' queries, respectively. The metric is formulated as:

$$\text{Sim-corpus} = \frac{1}{M-1}\sum_{j=1,j\neq i}^{M}\cos(q_i, q_j), \tag{6}$$

where q_i is the encoded query for source i, M is the total number of sources (here $M = 5$), and cos denotes the cosine similarity between the two encoded queries. Table 10 shows the average metric across OMVQA-Rad, OMVQA-Oph, and Quilt-VQA for different baseline methods. It can be observed that MQG greatly reduces the inter-corpus query similarity, avoiding redundancy among the source queries.

A.5 Joint Error Analysis of Retrieval and Answer Generation

To analyze cases where the model fails or misuses knowledge sources, and to identify whether the primary bottlenecks reside in the retrieval phase or in the alignment of the Med-LVLM, we conduct a comprehensive joint error analysis. We define four typical error types from the perspectives of input, perception, and integration. First, *Retrieval Failure (Type a)* occurs when the retrieved contents lack the necessary factual evidence to support the gold answer. Second, *Visual Perception Limitation (Type b)* arises when a question requires precise spatial orientations or fine-grained visual features that exceed the model's inherent visual perception capacity, even if the retrieved content is sufficient.

Method	Dist-Research (↑)	Dist-Wiki (↑)	Dist-Book (↑)	Dist-Guideline (↑)	Dist-Graph (↑)	Avg. (↑)
Original Question	0.1504	0.1504	0.1504	0.1504	0.1504	0.1504
Zero-shot Rewriting	0.2357	0.3239	0.3048	0.3307	0.3357	0.3062
MQG (Ours)	**0.4105**	**0.4175**	**0.3933**	**0.3936**	**0.3774**	**0.3985**

Table 9: Query diversity across knowledge sources. Higher values (↑) indicate greater lexical diversity.

Method	Sim-Research (↓)	Sim-Wiki (↓)	Sim-Book (↓)	Sim-Guideline (↓)	Sim-Graph (↓)	Avg. (↓)
Original Question	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Zero-shot Rewriting	0.9627	0.9656	0.9645	0.9611	0.9156	0.9539
MQG (Ours)	**0.7394**	**0.7390**	**0.7358**	**0.7156**	**0.7238**	**0.7307**

Table 10: Inter-corpus query similarity across knowledge sources. Lower values (↓) indicate higher query distinctiveness across sources.

Failure Type	OMVQA-Rad	OMVQA-Oph	Quilt-VQA	Average
Type a (%)	19.57	23.21	**44.09**	28.96
Type b (%)	3.40	4.76	10.75	6.30
Type c (%)	**75.74**	**70.24**	**44.09**	**63.36**

Table 11: Distribution of failure types.

Third, *Knowledge Alignment Failure (Type c)* describes instances where both textual evidence and visual cues are explicit, but the model fails to successfully integrate them or is misled. Finally, remaining anomalies are categorized as *Others (Type d)*. The detailed prompt is shown in D.6.

Based on these failure types, we annotated error cases from three representative out-of-domain datasets: OMVQA-Rad, OMVQA-Oph, and Quilt-VQA. To ensure reliable and scalable evaluation, we employed the expert medical model in the main method for automated annotation. The quantitative results are summarized in Table 11.

Based on these results, we draw three key conclusions. First, the system bottleneck has successfully shifted from information retrieval to knowledge reasoning; as ModCLIPs and MQG minimize retrieval failures (Type a), alignment and integration (Type c) naturally become the primary challenge. Second, deep medical alignment represents a challenging "last-mile" problem for Med-LVLMs, because adjudicating subtle logic conflicts between internal and external heterogeneous knowledge remains difficult for 7B-scale backbones, even with HKPT mitigating disalignment. Finally, heterogeneous knowledge alignment is the core key to medical AI reliability; while high-quality retrieval is a prerequisite, the most critical gains stem from enhancing multi-source alignment through methods like HKPT, pointing the way for future medical AI development.

A.6 Analysis of Computational Cost

For the training phase, we focused primarily on the offline data construction stage for MQG training, as the overhead for query generation, document retrieval, and expert model judging constitutes the majority of the HeteroRAG construction time. We also measured the costs for the online inference stage. Specifically, we randomly sampled 200 instances from the training set to measure the time spent on retrieval and inference for data construction. For testing, we sampled 200 VQA and 200 report generation samples from the test sets. Table 12 shows the measurement results.

B Additional Details

B.1 MedAtlas Details

The statistics of the multimodal report knowledge base and textual corpora in MedAtlas are shown in Table 13 and Table 14, respectively. For the multimodal report knowledge base, its radiology subset includes IU-Xray (Demner-Fushman et al., 2015), PLA (Li et al., 2024), CheXpert-Plus (Chambon et al., 2024), MIMIC-CXR (Johnson et al., 2019), ROCOv2 (Rückert et al., 2024), and PMC-OA-Rad (Lin et al., 2023). The ophthalmology subset includes Harvard-FairVLMed (Luo et al., 2024), DeepEyeNet (Huang et al., 2021), FFA-IR (Li et al., 2021), MM-Retinal (Wu et al., 2024b), and PMC-OA-Oph (Lin et al., 2023). The pathology subset includes ARCH (Gamper and Rajpoot, 2021), PathCap (Sun et al., 2024b), PatchGastric (Tsuneki and Kanavati, 2022), Quilt-1M (Ikezogwo et al., 2023), and PMC-OA-Pat (Lin et al., 2023).

The textual knowledge base of MedAtlas encompasses a diverse collection of biomedical and general-domain corpora. The Research corpus in-

Stage	Per-sample Retrieval Latency (s)	Per-sample Inference Latency (s)
Offline Data Construction	4.2	17.4
Online Testing (VQA)	0.5	1.4
Online Testing (Report Generation)	0.4	2.8

Table 12: Latency analysis across different stages.

Source	Modality	# Pairs	# Total
IU-Xray	Rad.	495	1.1M
PLA		14.7k	
CheXpert-Plus		187.6k	
MIMIC-CXR		209.6k	
ROCOv2		79.8k	
PMC-OA-Rad		612.2k	
Harvard-FairVLMed	Oph.	5.0k	112.0k
DeepEyeNet		2.9k	
FFA-IR		44.7k	
MM-Retinal		4.4k	
PMC-OA-Oph		55.1k	
ARCH	Pat.	6.8k	1.5M
PathCap		221.3k	
PatchGastric		262.8k	
Quilt-1M		433.9k	
PMC-OA-Pat		589.3k	

Table 13: Statistics of multimodal report knowledge base in MedAtlas.

Source	Corpus	# Chunks	# Total
PubMed	Research	48.0M	48.0M
Wikipedia	Wiki	29.7M	29.7M
E-books	Book	13.7M	14.1M
MedQA		125.8k	
StatPearls		322.7k	
Meditron	Guideline	657.9k	657.9k
-	-	**# Terms**	**# Relations**
UMLS	Graph	1.7M	2.9M

Table 14: Statistics of textual corpora in MedAtlas.

cludes PubMed Annual Baseline (NCBI, 2025), a comprehensive collection of biomedical literature. The Wiki corpus includes Wikipedia (Wikimedia, 2023), providing broad-domain textual knowledge. The Book corpus comprises E-books (Chen et al., 2025; Wu et al., 2024a), MedQA Textbooks (Jin et al., 2020), and StatPearls (StatPearls, 2024), offering in-depth medical knowledge from authoritative sources. The Guideline corpus includes Meditron Guidelines (Chen et al., 2023), which contains curated clinical practice guidelines. The Graph corpus is from UMLS Metathesaurus (Bodenreider, 2004), a comprehensive semantic network that integrates concepts and relationships from multiple biomedical vocabularies.

B.2 Dataset Details

The datasets used in our work include medical VQA datasets and medical report generation datasets. The VQA datasets are introduced as follows:

- **VQA-RAD** (Lau et al., 2018) is the first manually curated VQA dataset in radiology, where clinical questions were naturally formulated by medical professionals based on radiological images, along with reference answers. We employ the closed-ended subset. We use the official training split of size 1,027 and the official test split of size 272.
- **SLAKE** (Liu et al., 2021) is a large bilingual medical VQA dataset featuring comprehensive semantic annotations by experienced physicians, accompanied by a structured medical knowledge base. We employ the English closed-ended subset. We use the official training split of size 1,943 and the official test split of size 416.
- **OMVQA-Rad** (Hu et al., 2024) is the radiology subset of the OmniMedVQA dataset, which aggregates data from multiple medical classification datasets and converts them into a VQA format. We employ the open-access subset. We randomly select 1,200 samples for the test set.
- **DME-VQA** (Tascon-Morales et al., 2022) is built upon two public retinal image datasets, IDRiD (Porwal et al., 2018) and e-Ophta (Decenciere et al., 2013), containing questions related to diabetic macular edema (DME) and other eye conditions. The contours of the original image masks are extracted and rendered as red outlines on the original images to form the question images for each sample. We randomly select 5,000 samples from the official training split for the training set and use the official test split of size 1,311.

- **OMVQA-Oph** (Hu et al., 2024) is the ophthalmology subset derived from the OmniMedVQA dataset. We employ the open-access subset. We randomly select 1,200 samples for the test set.

- **PathMMU** (Sun et al., 2024a) is a high-quality, diverse pathology VQA dataset designed to assess the reasoning and understanding capabilities of large multimodal models in pathology. We employ its PathCLS and Atlas subsets, as they are not included in the pre-training data of Lingshu-7B to the best of our knowledge. Then we randomly select 2,095 samples for the training set and 598 samples for the test set.

- **PathVQA** (He et al., 2020) is the first VQA dataset in pathology, constructed using a semi-automated pipeline that extracts question-answer pairs from pathology textbooks and digital libraries. We employ the closed-ended subset. We randomly select 5,000 samples from the official training split for the training set and use the official test split of size 3,391.

- **Quilt-VQA** (Seyfioglu et al., 2024) is an organic evaluation dataset created by extracting real-world medical questions and answers from QUILT educational videos. We employ the closed-ended subset. We randomly select 343 samples for the test set.

The medical report generation datasets are described as follows:

- **MIMIC-CXR** (Johnson et al., 2019) is a large, publicly available collection of chest radiographs in DICOM format, paired with free-text radiology reports from studies conducted at the Beth Israel Deaconess Medical Center in Boston, MA. We exclude the samples that do not contain findings or impressions. We randomly select 5,000 samples from the official training split for the training set and use the official test split of size 1,624.

- **IU-Xray** (Demner-Fushman et al., 2015) consists of chest X-ray images linked to their corresponding clinical diagnostic reports. We exclude the samples that do not contain findings or impressions. We use the official training split of size 2,445 and the official test split of size 296.

- **Harvard-FairVLMed** (Luo et al., 2024) includes patient records with SLO fundus images and clinical notes for glaucoma diagnosis. We randomly select 3,500 samples from the official training split for the training set and 1,000 samples from the official test split for the test set.

B.3 Baseline Details

Decoding-based methods aiming to improve factuality are described as follows:

- **Original** uses greedy decoding, which selects the token with the highest probability at each generation step, favoring locally optimal choices without considering long-term sequence quality.

- **Beam Search** (Sutskever et al., 2014) improves upon greedy decoding by keeping track of multiple partial sequences (beams) at each step, exploring a wider range of potential outputs and often yielding more coherent and accurate generations.

- **DoLa** (Chuang et al., 2024) leverages the discrepancy between early and later layer representations in the model by comparing their projected logits onto the vocabulary space, guiding generation toward more accurate and contextually appropriate tokens.

- **VCD** (Leng et al., 2024) introduces a training-free decoding strategy that compares outputs from original and perturbed visual inputs, helping to mitigate model reliance on statistical bias and unimodal priors.

- **AVISC** (Woo et al., 2024) is a test-time decoding method that enhances visual understanding by dynamically recalibrating attention during token generation, specifically reducing over-attention to image tokens that lack task-relevant content.

- **M3ID** (Favero et al., 2024) strengthens the impact of the reference image during generation by amplifying tokens that have higher mutual information with the visual input.

Medical report-retrieval methods are described as follows:

- **MedDr** (He et al., 2024) employs a retrieval-augmented medical diagnosis strategy in the

inference process to improve the factuality of the model's responses.

- **FactMM-RAG** (Sun et al., 2025) feeds the multimodal question together with the retrieved report to the Med-LVLM, which is fine-tuned using standard SFT to better incorporate external reports.

- **RULE** (Xia et al., 2024) constructs a preference dataset focusing on cases where over-reliance on retrieved reports causes errors, aiming to balance the use of internal knowledge and external context.

- **MMed-RAG** (Xia et al., 2025) extends RULE (Xia et al., 2024) by introducing cross-modality alignment to ensure image utilization and proposing overall alignment to better incorporate external reports.

Medical document-retrieval methods are described as follows:

- **MKGF** (Wu et al., 2025) uses a multimodal retriever to fetch knowledge graphs and supplement knowledge for LVLMs. We reproduce it using ModCLIP for image-to-text and text-to-text retrieval to retrieve text corpora, combining results via Reciprocal Rank Fusion.

- **K-LLaVA** (Hamza et al., 2025) retrieves relevant KG triplets using a CLIP model and fine-tunes the LVLM to incorporate the knowledge. We also use ModCLIP for retrieval in this method.

A more recent work that retrieves both reports and documents is described as follows:

- **MIRA** (Wang et al., 2025) is a method that retrieves both medical reports and documents. To reproduce it, we use the input image to retrieve similar clinical cases and employ a zero-shot query-rewriting module (Lingshu-7B) for corpus retrieval. Then the downstream reader is fine-tuned, whose training data includes a chain-of-thought to guide the reader in analyzing the external knowledge.

We also introduce widely used Med-LVLMs, which are described as follows:

- **LLaVA-Med-7B** (Li et al., 2023) first aligns biomedical terminology using figure-caption pairs from scientific literature, then enhances conversational understanding through GPT-4-generated instruction-following data, simulating the way non-experts gradually acquire medical knowledge through.

- **MedGemma-4B** (Sellergren et al., 2025) is developed by Google and exhibits strong medical image and text understanding capabilities, significantly outperforming other generative models of similar size and approaching the performance of specialized task-specific models.

- **HuatuoGPT-V-34B** (Chen et al., 2024a) is trained on PubMedVision, a large-scale dataset of 1.3 million medical VQA samples constructed by refining image-text pairs from PubMed with the help of MLLMs (e.g., GPT-4V), showing superior performance in medical multimodal scenarios.

- **HealthGPT-32B** (Lin et al., 2025) integrates medical visual comprehension and generation into a unified autoregressive framework, progressively adapting heterogeneous multimodal knowledge to a pre-trained LLM through a bootstrapping approach.

- **Lingshu-32B** (Xu et al., 2025) is developed based on a carefully curated multimodal dataset enriched with comprehensive medical knowledge, undergoing multi-stage training to progressively embed domain expertise and improve task-solving abilities, consistently outperforming existing open-source models in most medical multimodal benchmarks.

B.4 Implementation Details

For report retrieval, we adopt the adaptive retrieval context selection method (Xia et al., 2025). For document retrieval, the MQG generates up to four queries in total for unstructured corpora (all except Graph). Each query retrieves the top-10 documents, which are then reranked to select the top-2 documents. For the Graph corpus, the MQG retrieves one term and re-ranks the top-10 relations. The correctness threshold α_r in HKPT is set to 50/3 for radiology reports and 20/3 for ophthalmology reports, which is tuned based on the external development set.

For the training of ModCLIPs, they are initialized from BiomedCLIP (Zhang et al., 2023). The learning rate is set to 2e-4, and the batch size is set to 512. The number of training epochs of radiology, ophthalmology, and pathology ModCLIP is set to 10, 100, and 10, respectively, for the different sizes of modality image-text pairs.

For the training of MQG, the Med-LVLM is initialized from Lingshu-7B (Xu et al., 2025). We use LoRA (Hu et al., 2022) for efficient fine-tuning. For the SFT process, its learning rate is set to 2e-4, the batch size is set to 64, and the number of epochs is 3. For the DPO process, its learning rate is set to 2e-5, the batch size is set to 64, and the number of epochs is set to 3.

For the training of HKPT, the Med-LVLM is initialized from Lingshu-7B. We also use LoRA (Hu et al., 2022) for efficient fine-tuning. Its learning rate is set to 2e-5, the batch size is set to 64, and the number of epochs is set to 3.

In our experiments, we use the development set, which has no overlap with the training and test sets, to tune the hyperparameters. We use the AdamW (Loshchilov and Hutter, 2019) as our optimizer. For the test stage, the temperature is set to 0 to ensure reproducibility. Detailed prompts are provided in Appendix D. Huggingface Trainer is adopted as the training framework for Med-LVLMs.

C AI Assistance Statement

LLM tools are employed solely for language refinement and grammar correction to improve overall clarity and readability.

D Prompt List

Prompt D.1: VQA with Retrieved Reports and Documents

{question_image}

Retrieved Contents:
{text_doc}

Reference Reports:
{mm_doc}

{question_text}
Please answer the question based on the Retrieved Contents. It should be noted that the diagnostic information in the Reference Reports cannot be directly used as the basis for diagnosis, but should only be used for reference and comparison.

Answer with the option's letter from the given choices directly.

Prompt D.2: Report Generation with Retrieved Reports and Documents

{question_image}

Retrieved Contents:
{text_doc}

Reference Reports:
{mm_doc}

Please answer the question based on the Retrieved Contents. It should be noted that the diagnostic information in the Reference Reports cannot be directly used as the basis for diagnosis, but should only be used for reference and comparison.

(For radiology) You are a helpful assistant. Please generate a report for the given image, including both findings and impressions. Return the report in the following format: Findings: {} Impression: {}.
(For ophthalmology) You are a helpful assistant. Please generate a short report for the given image in 100 words. Please only include the content of the report in your response.

Prompt D.3: Query Exploration by the Expert Med-LVLM

{question_image}

Question (based on the image)
{question_text}

Corpus Description
research: The corpus provides access to advanced biomedical research, facilitating access to specialized knowledge and resources.
wiki: The corpus provides access to general knowledge across a wide range of topics.
book: The corpus provides access to medical knowledge resource including various educational resources and textbooks.
guideline: The corpus provides access to clinical guidelines from leading health organizations.
graph: The corpus provides a structured knowledge graph that connects medical definitions and related terms.

Query Format
<research>{query0} ; {query1} ; ... (Use ; to separate the queries)</research>
<wiki>{query0} ; {query1} ; ... (Use ; to separate the queries)</wiki>
<book>{query0} ; {query1} ; ... (Use ; to separate the queries)</book>
<guideline>{query0} ; {query1} ; ... (Use ; to separate the queries)</guideline>
<graph>{query_term0} , {query_relation0} ; {query_term1} , {query_relation1} ; ... (Use ; to separate the queries. Each query should use , to separate the {query_term} and {query_relation})</graph>

To answer the question labeled as # Question, please construct appropriate queries to get the information

you need.
1. Each corpus in # Corpus Description must have search queries constructed.
2. Please give the search queries following the format in # Query Format. Each corpus should have 6 queries, separated by ';'.
3. The queries generated for each corpus should exhibit diversity and be closely aligned with the specific information needs and characteristics of that corpus.

Prompt D.4: Query Judging through Retrieved Documents by the Expert Med-LVLM

{question_image}

Question (based on the image)
{question_text}

Gold Answer
{gold}

Documents
{documents}

You are a professional medical expert. Please judge whether the information in the # Documents supports the # Gold Answer as a response to the # Question. Please judge whether # Documents supports the # Gold Answer in response to the # Question, rather than evaluating if the # Question's answer is the # Gold Answer. Please first think step-by-step and then show your judgement using the format <answer>yes/no</answer> at the end of your response. Please keep your entire response simple and complete, up to 100 words.

Prompt D.5: Query Generation by the Multi-corpora Query Generator

{question_image}

Question (based on the image)
{question_text}

Corpus Description
research: The corpus provides access to advanced biomedical research, facilitating access to specialized knowledge and resources.
wiki: The corpus provides access to general knowledge across a wide range of topics.
book: The corpus provides access to medical knowledge resource including various educational resources and textbooks.
guideline: The corpus provides access to clinical guidelines from leading health organizations.
graph: The corpus provides a structured knowledge graph that connects medical definitions and related terms.

Query Format
<research>{query}</research>
<wiki>{query}</wiki>
<book>{query}</book>
<guideline>{query}</guideline>
<graph>{query_term} , {query_relation} (Each query should use , to separate the {query_term} and {query_relation})</graph>

To answer the question labeled as # Question, please construct appropriate queries to get the information you need.
1. Please give the search queries following the format in # Query Format. For each corpus, if you think no information retrieval is needed, simply output an empty tag for that corpus, for example: <book></book>.
2. The queries generated for each corpus should be closely aligned with the specific information needs and characteristics of that corpus.

Prompt D.6: Joint Error Analysis of Retrieval and Answer Generation

{question_image}

Retrieved Contents
{reports_and_documents}

Question (based on the image)
{question_text}

Gold Answer
{gold}

Wrong Response
{wrong}

You are a professional medical expert. The current model has provided an incorrect answer # Wrong Response instead of the # Gold Answer to the question # Question (based on the image) despite having access to # Retrieved Contents. Categorize the model's failure into exactly one of these four mutually exclusive categories:
a. Retrieval Failure: The # Retrieved Contents lack the necessary factual evidence to support the # Gold Answer.
b. Visual Perception Limitation: Even if the retrieved content is sufficient, the question requires precise spatial orientations or fine-grained visual features that may exceed the model's inherent visual perception capacity.
c. Knowledge Alignment Failure: Both textual evidence and visual cues are explicit, but the model failed to integrate them or was misled.
d. Others

Please first think step-by-step and then show your judgement using the format <answer>a/b/c/d</answer> at the end of your response. Only one option can be chosen. Please keep your entire response simple and complete, up to 100 words.