Patho-AgenticRAG: Towards Multimodal Agentic Retrieval-Augmented Generation for Pathology VLMs via Reinforcement Learning

Wenchuan Zhang[1,2*], Jingru Guo[3*], Hengzhe Zhang[4*], Penghao Zhang[5], Jie Chen[2], Shuwan Zhang[6], Zhang Zhang[1], Yuhao Yi[1,2†], Hong Bu[1,2]

[1]Department of Pathology, West China Hospital, Sichuan University
[2]Institute of Clinical Pathology, West China Hospital, Sichuan University
[3]University of Toronto
[4]School of Engineering and Computer Science, Victoria University of Wellington
[5]Independent Researcher
[6]Department of Pathology, Shengjing Hospital of China Medical University
zhangwenchuan@stu.scu.edu.cn, yuhaoyi@scu.edu.cn

Abstract

Although Vision Language Models (VLMs) have shown generalization in medical imaging, pathology presents unique challenges due to ultra-high resolution, complex tissue structures, and nuanced semantics. These factors make pathology VLMs prone to hallucinations, i.e., generating outputs inconsistent with visual evidence, which undermines clinical trust. Existing RAG approaches in this domain largely depend on text-based knowledge bases, limiting their ability to leverage diagnostic visual cues. To address this, we propose Patho-AgenticRAG, a multimodal RAG framework with a database built on page-level embeddings from authoritative pathology textbooks. Unlike traditional text-only retrieval systems, it supports joint text–image search, enabling retrieval of textbook pages that contain both the queried text and relevant visual cues, thus avoiding the loss of critical image-based information. Patho-AgenticRAG also supports reasoning, task decomposition, and multi-turn search interactions, improving accuracy in complex diagnostic scenarios. Experiments show that Patho-AgenticRAG significantly outperforms existing multimodal models in complex pathology tasks like multiple-choice diagnosis and visual question answering.

Code —
https://github.com/Wenchuan-Zhang/Patho-AgenticRAG
Extended version — https://arxiv.org/abs/2508.02258

1 Introduction

With the continuous development of large-scale vision-language models (VLMs), multimodal learning has made breakthrough progress in many fields such as natural image understanding, image-text generation, and medical image analysis. Compared with other medical images (such as X-rays, CT, and MRI), pathological images, with their ultra-high resolution, fine-grained structure, and complex semantic relationships, have put forward higher requirements on the perception, reasoning, and factual consistency capabilities of the model.

In recent years, many studies have introduced VLMs into digital pathology tasks, such as diagnostic assistance (Chen et al. 2025a), risk stratification (Liu et al. 2025), and question-answering systems (Lu et al. 2024). However, existing pathology VLMs face challenges like hallucinations and lack of structured semantic control in retrieval mechanisms, particularly in tasks requiring factual support and traceable evidence. Building a reliable, multimodally pathology VLM with consistency assurance is a key issue. While Retrieval-Augmented Generation (RAG) has been applied to medical multimodal tasks to enhance the accuracy and credibility of medical reasoning, limitations remain in pathological image analysis. MMed-RAG proposed a general multimodal medical RAG system (Xia et al. 2025), but its image vector library lacks fine-grained annotation for organ or tissue systems, categorizing only by imaging modality (e.g., CT, X-ray). Some studies rely solely on text retrieval, ignoring the crucial role of images in supporting VLM reasoning, especially in image-text consistency and visual evidence scenarios (Jabal et al. 2024; Cheetirala et al. 2025). Methods such as MedRAG and Medical Graph RAG introduce complex reasoning guided by knowledge graphs but suffer from complicated design, lack of scalability, and poor adaptation to clinical tasks (Zhao et al. 2025; Wu et al. 2024). The RAG module in Liu et al. also struggles with insufficient instruction-following, hindering stable knowledge extraction (Liu et al. 2023). This study aims to build an intelligent retrieval augmented generation framework for pathology VLMs, to improve the credibility and interpretability of the model in complex question-answering and reasoning tasks.

Our method focuses on three dimensions: multimodal knowledge retrieval, intelligent planning capabilities, and adaptability to pathology scenarios. Unlike previous RAG framework based on static prompts or text retrieval, Patho-AgenticRAG introduces a multimodal image-text retrieval module and an agent mechanism with planning capabili-

*These authors contributed equally.
†Corresponding author.

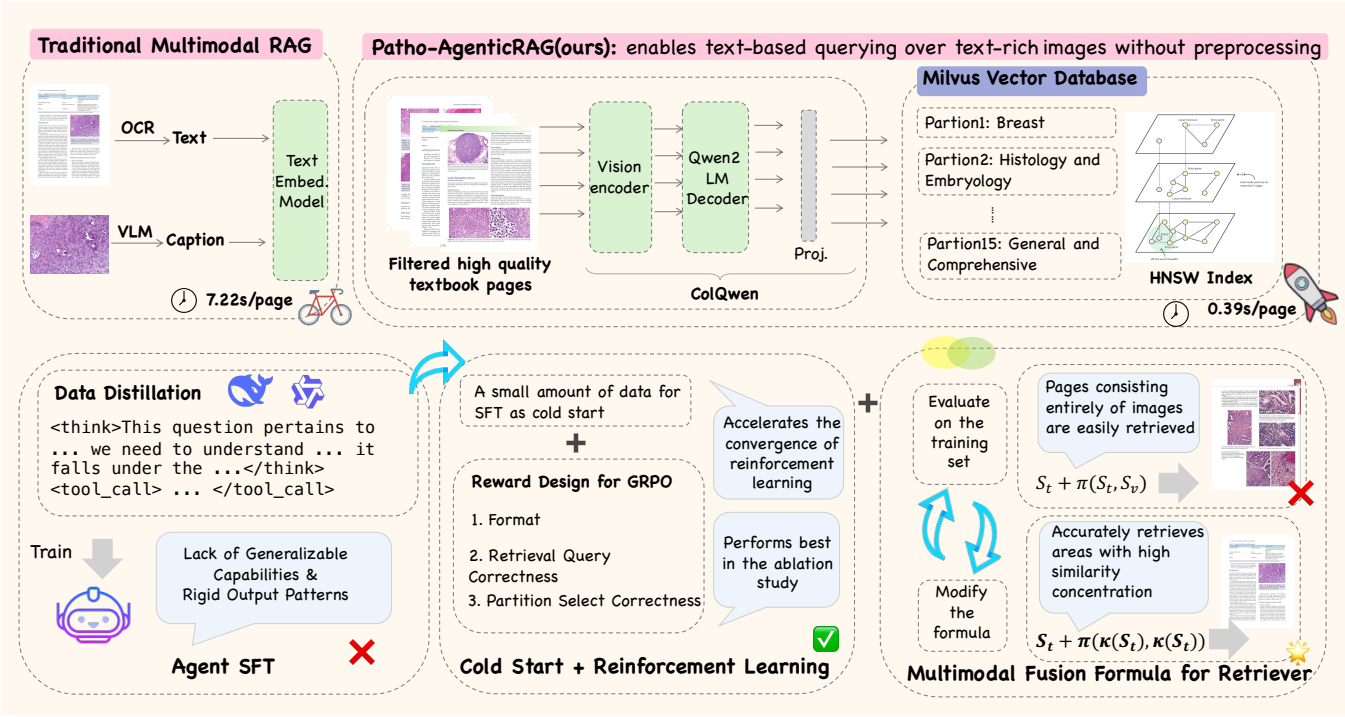

Figure 1: Knowledge Base Construction and Agent Training Method in the Patho-AgenticRAG

ties, which enables the model to more effectively retrieve target images and corresponding knowledge content from the structured pathology knowledge base, and to reasonably integrate and reason. We also incorporate a reinforcement learning optimization strategy to make the agent more robust and generalizable in the complex and uncertain question-answering environment (see Figure 1). The main contributions are summarized as follows:

- We proposed **a novel Multimodal Retrieval Mechanism** that combines multimodal (image-text) vector space modeling with a tissue-aware retrieval strategy. This significantly improves the recall rate of the target knowledge fragment while ensuring accuracy, providing a guarantee for fine-grained knowledge alignment in pathology diagnosis tasks.
- We built **a planning-capable intelligent agent within the Agentic RAG system**, which autonomously plans multi-round retrieval and reasoning trajectories in response to complex natural language pathology questions. It dynamically invokes relevant multimodal knowledge and effectively supports long-term dependency modeling and multi-hop reasoning in diagnostic tasks.
- We proposed **a Tool-Integrated Reasoning training paradigm tailored for medical diagnostics**, built upon GRPO. This paradigm enables the agent to make fine-grained decisions, such as whether to invoke retrieval, how to reformulate questions, and how to assign domainspecific tools or classifiers within complex pathology question answering scenarios. It addresses the high-stakes nature of medical reasoning by promoting robust decision-making and reliable tool coordination.

2 Related Work

2.1 Multimodal Agentic RAG

Retrieval-Augmented Generation enhances large language models by integrating external knowledge to improve completeness and factual accuracy (Lewis et al. 2020; Huang et al. 2025). Traditional RAG systems—such as Na"ive RAG (Zhai et al. 2023; Lee et al. 2024) and Advanced RAG (Yu et al. 2024; Cho et al. 2024)—follow a fixed "retrieve-then-read" pipeline that works for simple queries but falls short on tasks requiring multi-step reasoning, adaptive retrieval, or tool use (Singh et al. 2025). Agentic RAG advances this paradigm by introducing autonomous agents (Gao et al. 2023) that can plan retrieval (Joshi et al. 2024), call tools (Chen et al. 2025b), refine outputs (Ravuru, Sakhinana, and Runkana 2024), and collaborate across roles or modalities (Wang et al. 2025; Jin et al. 2025; Wu et al. 2025). However, in the medical domain, studies have largely centered on text or structured data (Thakrar, Basavatia, and Daftardar 2025; Zeggai et al. 2025), overlooking the rich diagnostic information embedded in medical images—especially in visually driven fields like pathology, where tissue morphology, staining patterns, and spatial structures are essential.

2.2 Reinforcement Learning for Medical VLMs

Reinforcement learning (RL) offers a promising way to align VLM outputs with clinical accuracy requirements, particularly in high-risk settings where hallucinations can be harm-

ful (Zhang et al. 2025). A key challenge is maintaining factual consistency between visual evidence and textual outputs (Sun et al. 2025; Chen et al. 2024). Direct end-to-end RL for large VLMs, however, is impractical due to limited high-quality physician-annotated rewards (Pham and Ngo 2025), training instability, and poor interpretability of learned behaviors (Zhu et al. 2025). Recent work instead adopts agent-centric approaches, applying RL to external agents rather than model parameters. These agents optimize decision-making—such as formulating better queries, validating responses, or selecting external knowledge (Xia et al. 2025; Jin et al. 2025; Wu et al. 2025)—providing safer, more transparent, and more controllable optimization.

3 Methodology

3.1 Overall Framework

The overall architecture adopts a modular design and contains four main components. 1. Multimodal pathology knowledge base: This is a specialized vector database containing a rich collection of pathology textbook pages. It acts as an external storage for agent queries to collect relevant evidence, such as images of similar cases and their corresponding diagnostic descriptions; 2. Intelligent agentic router: This is the central processing unit of our framework. It accepts the initial diagnosis query, decomposes it into logically sequential subtasks, and plans; 3. VRAG Agent (Chen et al. 2025b): This module supports multi-turn retrieval and summarization. It interacts with the knowledge base and distills the returned textbook images into concise, useful information. 4. Core vision language model (inference engine) (Zhang et al. 2025): We use the pretrained pathology VLM as the basic inference engine. With the contextual summaries provided by the VRAG agent, it performs inference to address the diagnostic query.

3.2 Construction of a Multimodal Pathology Knowledge Base

To support retrieval-augmented reasoning in pathology, we construct a high-quality multimodal knowledge base that integrates authoritative textual and visual information. We curated a large corpus by collecting over 600 authoritative pathology textbooks—approximately 300,000 pages in total and, after removing irrelevant content, retained more than 200,000 high-quality pages, which were converted into image-based samples for diagnostic relevance. Using the ColQwen2 model (Faysse et al. 2025), we embed image-text pairs into a unified vector space that captures both visual and semantic signals. The embeddings are indexed with the HNSW algorithm (Malkov and Yashunin 2018) and stored in Milvus (Wang et al. 2021) to support efficient high-dimensional retrieval during reasoning. Full construction details are provided in Appendix A.

3.3 Multimodal Fusion

Method Let $S_t \in \mathbb{R}^{N_t \times N_d}$ denote the text–document similarity matrix and $S_v \in \mathbb{R}^{N_v \times N_d}$ the image–document similarity matrix, where N_t is the number of text tokens, N_v the number of image patches, and N_d the number of document tokens. These two modalities are fused by the following expression:

$$\begin{aligned} &\mathrm{std}\big(\mathrm{std}(S_t[i,:])\big) \times \big[\mathrm{mean}\big(\kappa(S_t[i,:])\big)\big]^2 \\ &\times \mathrm{mean}\big(\kappa(S_v[i,:])\big) + \mathrm{mean}\big(\max(S_t[i,:])\big), \end{aligned} \tag{1}$$

where $\kappa(S_t[i,:])$ denotes the kurtosis of the similarity scores between all document tokens and text token i, and $\kappa(S_v[i,:])$ denotes the kurtosis of the similarity scores between all document tokens and image patch i. The first term captures how the standard deviation $\mathrm{std}(\cdot)$ and kurtosis $\kappa(\cdot)$ reflect the variation of similarity scores across tokens or image patches with respect to the database document, represented as rows in S_t or S_v. This encourages the retrieval results to have different importance for various tokens in the document, indexed by $j = 1, \ldots, N_d$. The reason for this is, in practice, only a portion of a page is typically relevant to the database document, meaning only some j-th elements in $S_t[i,j]$ or $S_v[i,j]$ contribute meaningfully. When all parts of a document exhibit high responsiveness, resulting in a low standard deviation, this may indicate a noisy document with problematic embeddings, which should be deprioritized. The second term, $\mathrm{mean}\big(\max(S_t[i,:])\big)$, as in CoPaLi (Faysse et al. 2025), quantifies the maximum relevance $\max(S_t[i,:])$ of each token to any token in the document.

It is important to note that the similarity matrix of the image modality with respect to the document embedding, i.e., $S_v \in \mathbb{R}^{N_v \times N_d}$, is used only to calculate the kurtosis $\kappa(S_v[i,:])$, and the average similarity information from the image modality, such as $\mathrm{mean}(S_v[i,:])$, is not incorporated. This design emphasizes attention to the most relevant portions of a page, identified via large values of $\kappa(S_v[i,:])$. When a token or patch shows high similarity to all parts of a document, it is treated as a noisy retrieval result and is assigned a lower score during re-ranking.

Explanation To intuitively understand how Equation (1) influences the re-ranking process, we present the similarity matrices of the first- and second-ranked documents with respect to the query text embedding and image embedding in Figure 3. As shown in Figure 3, the second-ranked document exhibits a more uniform similarity, or attention, across the tokens of the document embedding. In contrast, the first-ranked document demonstrates more concentrated and higher similarity on fewer tokens of the document embedding. In this case, the high similarity of the first-ranked document to the query is meaningful, as it is caused by focused attention on the most informative tokens. On the other hand, the high similarity of the second-ranked document to the query results from more diffuse attention, likely caused by noise. Thus, although the second-ranked feature initially has the highest text similarity in the retrieval process, after re-ranking by the fusion formula, it drops to second place, while the original second-ranked document is promoted to the top rank due to its combination of high similarity and concentrated attention.

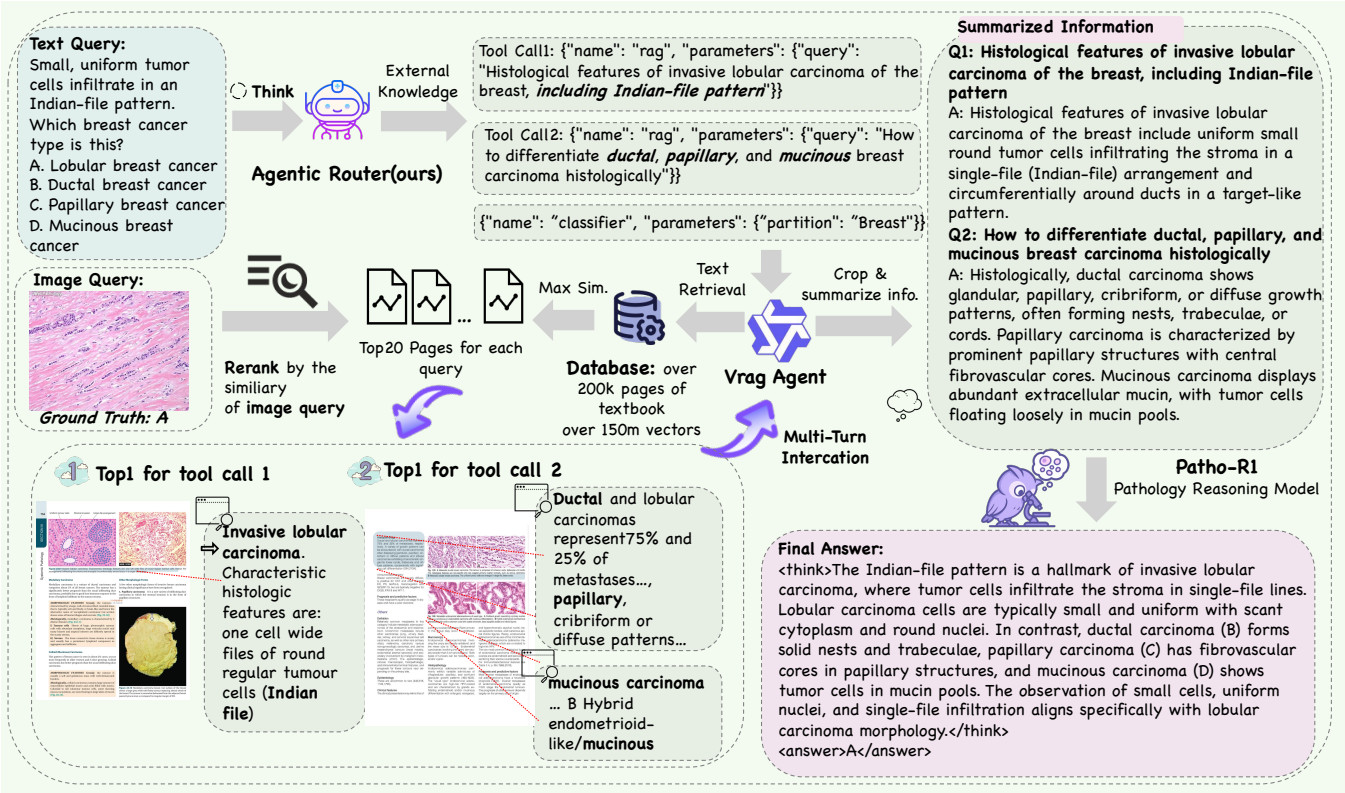

Figure 2: An illustration of the multi-turn retrieval and summarization process.

3.4 Agentic Diagnostic Workflow with Multimodal Evidence Tracing

Our system employs a multi-agent workflow that transforms diagnostic queries into evidence-based conclusions. The **Agentic Router** module parses the query, breaks it down into sub-tasks for each diagnostic candidate, and creates a high-level retrieval plan. It delegates the retrieval and evidence aggregation to the **VRAG Agent**, which performs a multi-turn retrieval process on the multimodal knowledge base. Initially, it conducts text-based retrieval with candidate-specific keywords, then re-ranks the results based on image-text similarity. Through iterative evidence refinement, as shown in Figure 2, the agent constructs a structured prompt containing the top visual evidence for each candidate. This prompt is passed to a vision-language model, which performs contrastive reasoning to generate a diagnosis and evidence-supported report. Further details are in Appendix B.

3.5 Tool Integrated RL for Agentic Router

Traditional RAG systems are static and don't adapt to query complexity. To address this, we introduce a reinforcement learning framework that allows an agent to learn dynamic routing strategies (Qian et al. 2025). The agent's task is to generate a decision path specifying whether and how to call the RAG system. Formally, given an input query Q_{orig}, the agent's policy π outputs a path P, optimized to maximize the expected similarity to a ground-truth decision path P_{gt}

$$\max_{\pi} \mathbb{E}_{P \sim \pi(Q_{\text{orig}})} \left[R_{\text{final}}(P, P_{\text{gt}}) \right] \tag{2}$$

The hierarchical reward R_{final} compares the generated path to the target path step by step (see Algorithm 1). The agent performs a sequence of decisions to construct the path:

- **Decision 1: Whether to invoke RAG?**
  - **Path A (No RAG Invocation):** If the agent decides `False`, the decision process terminates. This is for simple queries that can be answered without external knowledge. Final path: $\{\texttt{rag} : \texttt{False}\}$
  - **Path B (Invoke RAG):** If `True`, proceed to the next decision.
- **Decision 2: How to decompose the task?** (Only applies if RAG is invoked)
  The agent may choose to rewrite the query one or more times to better surface its core semantics and improve retrieval quality. This is not a mechanical rewriting process, but a targeted transformation to better align with the retrieval engine.
- **Decision 3: Whether to use a tissue-specific classifier?** (Only applies if RAG is invoked)
  The agent decides whether to enable a classifier to restrict retrieval to a relevant knowledge partition.
  - **Path B.1 (Global Retrieval):** If `False`, RAG retrieves documents from the full corpus. Fi-

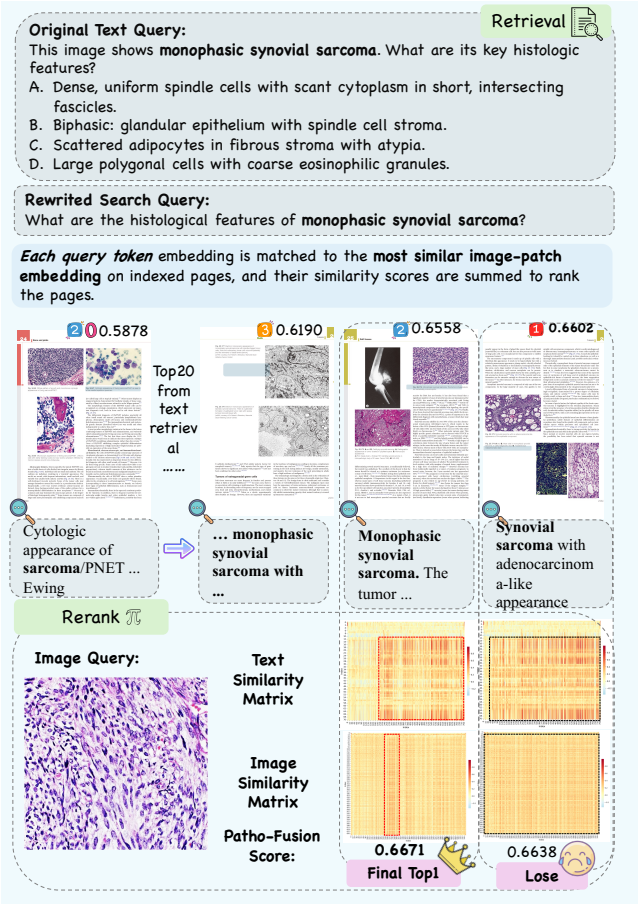

Figure 3: Illustration of the reranking process utilizing modality fusion.

nal path: $\{\texttt{rag} : \texttt{True}, \texttt{rewrite\_count} : n, \texttt{classifier} : \texttt{False}\}$

  - **Path B.2 (Classifier-Based Retrieval):** If $\texttt{True}$, proceed to the next decision.

- **Decision 4: Whether the classifier assigns the query to the correct partition?** (Only applies if a classifier is enabled)
  The agent selects a classifier from the available set $\{C_1, \ldots, C_m\}$ and attempts to assign the query to the correct partition. The effectiveness of this decision depends on both the selection of the classifier and the correctness of the classification. Final path: $\{\texttt{rag} : \texttt{True}, \texttt{rewrite\_count} : n, \texttt{classifier} : \texttt{True}, \texttt{partition} : C_j\}$

RL Training with GRPO We use the GRPO algorithm (Shao et al. 2024) to train the policy π_θ. For each query Q_{orig}, the agent generates multiple decision paths:

$$G_Q = \{(P_1, r_1), (P_2, r_2), \ldots, (P_n, r_n)\} \tag{3}$$

where each P_i is a complete decision path and r_i is its corresponding reward score $r_i \in [0, 4]$ based on the hierarchical reward function. We normalize the rewards within the group to compute the advantage function $A_i(P_i \mid Q)$:

$$A_i(P_i \mid Q) = \frac{r_i - \mu_Q}{\sigma_Q + \eta} \tag{4}$$

where μ_Q and σ_Q are the mean and standard deviation of rewards within group G_Q, and η is a small constant for numerical stability.

To update the policy, we apply the GRPO objective, which extends PPO by group-wise normalized advantages and KL regularization with a reference model. Specifically, for each group of outputs $\{o_i\}_{i=1}^G$, from the same query, we optimize:

$$\begin{aligned}\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim P(Q),\{o_i\}_{i=1}^G\sim\pi_{\theta_{\text{old}}}(O|q)} \Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|} \\ \min\left(\frac{\pi_\theta(o_{i,t}\mid q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}\mid q,o_{i,<t})}\hat{A}_{i,t}, \right. \\ \left. \text{clip}\left(\frac{\pi_\theta(o_{i,t}\mid q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}\mid q,o_{i,<t})}, 1-\epsilon, 1+\epsilon\right)\hat{A}_{i,t}\right) \\ - \beta D_{\text{KL}}\left(\pi_\theta(o_{i,t}\mid q,o_{i,<t}) \,\middle\|\, \pi_{\text{ref}}(o_{i,t}\mid q,o_{i,<t})\right)\Bigg]\end{aligned} \tag{5}$$

Here, $\hat{A}_{i,t}$ is the advantage at step t within each output, computed relative to other outputs for the same query. This group-wise comparison helps the agent learn from relative improvements, leading to more effective decision-making in complex reasoning paths. See Appendix C for details.

Algorithm 1: Hierarchical Reward Computation

Require: Agent path P, Ground Truth P_{gt}

$R_{\text{final}} \leftarrow 0$
if $P.\texttt{rag} \neq P_{\text{gt}}.\texttt{rag}$ **then**
  return R_{final} ▷ Incorrect Decision 1
end if
if $P_{\text{gt}}.\texttt{rag} = \texttt{False}$ **then**
  $R_{\text{final}} \leftarrow 4$ ▷ Correct Decision 1 (Path A)
else
  $R_{\text{final}} \leftarrow 1$ ▷ Correct Decision 1 (Path B)
  if $P.\texttt{rewrite\_count} = P_{\text{gt}}.\texttt{rewrite\_count}$ **then**
    $R_{\text{final}} \leftarrow R_{\text{final}} + 1$ ▷ Correct Decision 2
  end if
  if $P.\texttt{classifier} = P_{\text{gt}}.\texttt{classifier}$ **then**
    if $P.\texttt{classifier} = \texttt{False}$ **then**
      $R_{\text{final}} \leftarrow R_{\text{final}} + 2$ ▷ Correct Decision 3 (Path B.1)
    else
      $R_{\text{final}} \leftarrow R_{\text{final}} + 1$ ▷ Correct Decision 3 (Path B.2)
      if $P.\texttt{partition} = P_{\text{gt}}.\texttt{partition}$ **then**
        $R_{\text{final}} \leftarrow R_{\text{final}} + 1$ ▷ Correct Decision 4
      end if
    end if
  end if
end if
return R_{final}

Method	Rec@1	Rec@5	MRR@1	MRR@5	MRR@20	NDCG@1	NDCG@5	NDCG@20
CoPaLi (Text)	0.640 (2)	**0.900 (1)**	0.640 (2)	0.734 (2)	0.736 (2)	0.740 (2)	0.804 (2)	0.796 (2)
CoPaLi (Image)	0.060 (3)	0.220 (3)	0.060 (3)	0.112 (3)	0.170 (3)	0.080 (3)	0.174 (3)	0.359 (3)
WeiMoCIR	0.060 (3)	0.200 (4)	0.060 (3)	0.102 (4)	0.158 (4)	0.080 (3)	0.163 (4)	0.342 (4)
Patho-Fusion (ours)	**0.720 (1)**	0.880 (2)	**0.720 (1)**	**0.777 (1)**	**0.784 (1)**	**0.820 (1)**	**0.824 (1)**	**0.827 (1)**

Table 1: Comparison of retrieval methods. Numbers in parentheses denote rank.

4 Experiments

4.1 Multimodal Fusion

Baseline Methods For baseline comparisons, we evaluate the proposed method against CoPaLi (Faysse et al. 2025) and Weighted Modality Fusion and Similarity for Composed Image Retrieval (WeiMoCIR) (Wu, Lin, and Yang 2024).

- For CoPaLi (Faysse et al. 2025), which supports only a single modality, we apply retrieval separately for each modality. The scoring function is

$$\sum_{i=1}^{N_q} \max_{j=1,\ldots,N_d} \big\langle E_q(i),\, E_d(j) \big\rangle, \tag{6}$$

  where E_q is either E_t or E_v so $N_q = N_t$ or $N_q = N_v$, and $\langle \cdot, \cdot \rangle$ denotes the inner product. Each query embedding $E_q(i)$ is matched to the most relevant document embedding $E_d(j)$, where $j = 1, \ldots, N_d$, and the results are aggregated over all query tokens or patches.
- In WeiMoCIR, the query embedding is computed as

$$\mathbf{q} = (1 - \alpha) \cdot \mathbf{e}_v + \alpha \cdot \mathbf{e}_t, \tag{7}$$

  where $\mathbf{e}_v = \text{mean}(E_v(i, :))$ is the average vision embedding over all patches $i = 1, \ldots, N_v$, and $\mathbf{e}_t = \text{mean}(E_t(i, :))$ is the average text embedding over all tokens $i = 1, \ldots, N_t$. The parameter $\alpha = 0.1$ represents the weighting coefficient for the text modality. The final similarity score between the query and a database document is computed using the inner product as

$$\frac{1}{N_d} \sum_{j=1}^{N_d} \langle \mathbf{q},\, \mathbf{e}_{d,j} \rangle\,, \tag{8}$$

  where $\mathbf{e}_{d,j}$ is the j-th token embedding of the document.

Dataset and Evaluation Protocol The dataset consists of 100 pairs of images, questions, and answers curated by domain experts. We randomly split the dataset, using 50% for training and 50% for testing. The modality fusion function is optimized only on the training data to prevent potential data leakage. All modality fusion methods are evaluated on the test set, and recall, mean reciprocal rank (MRR), and normalized discounted cumulative gain (NDCG) metrics are reported.

Experimental Results The experimental results are shown in Table 1. Recall@20 is omitted since there are only 20 results in the re-ranking stage and the Recall@20 is identical for all algorithms. The results demonstrate a clear advantage for the proposed modality fusion method over the baseline approaches. While using only the text modality for retrieval can already achieve good performance, the retrieval results remain suboptimal without the proposed fusion strategy in Equation (1). Methods using only the image modality or WeiMoCIR perform worse than the proposed fusion by a significant margin. This is primarily because both heavily rely on the image modality for retrieval, whereas pathology images require strong domain expertise to interpret and general-purpose embedding methods may not provide optimal representations for this task. Although WeiMoCIR achieves strong results on general retrieval benchmarks, it underperforms in the medical multimodal retrieval setting. Overall, these findings demonstrate that general multimodal fusion strategies are not sufficient for the pathology domain and the specifically designed fusion mechanism proposed here offers superior effectiveness.

4.2 Patho-AgenticRAG Evaluation Results

Ablation Analysis We conducted three main ablation studies to investigate the necessity and data proportion of SFT before GRPO. The results show that skipping SFT leads to poor convergence during GRPO. However, performing SFT with a large amount of data causes the model to lack generalizable capabilities and exhibit rigid output patterns. Therefore, using a small amount of SFT data as a cold start before GRPO is the optimal strategy. Notably, adopting a lightweight SFT phase (e.g., SFT400) before GRPO achieves the best overall balance. This setting consistently outperforms both the "no-SFT" and "large-SFT" baselines across multiple datasets. For example, on the Path-VQA benchmark, using SFT400+GRPO4k improves performance from 77.51% (GRPO4k only) to 80.34%. Similarly, on the Quilt-VQA dataset, performance improves from 60.93% (GRPO4k) to 75.80%, a +14.87% increase, indicating that a small amount of supervised guidance before preference optimization significantly enhances model capability. These results suggest that SFT400 provides an effective "cold start" that guides the policy initialization without compromising flexibility or generalization, as shown by Figure 4.

Close-Ended Benchmarks Results Closed-ended questions play a crucial role in pathology-related tasks, particularly in diagnostic classification. To evaluate model performance on such tasks, we consider two types of close-ended question datasets: (1) Yes/No questions, selected from Path-VQA and Quilt-VQA; and (2) multiple-choice questions, sourced from PathMMU (Sun et al. 2024), MedXpertQA (Zuo et al. 2025), and OmniMedVQA (Hu et al. 2024).

The results on close-ended benchmarks are summa-

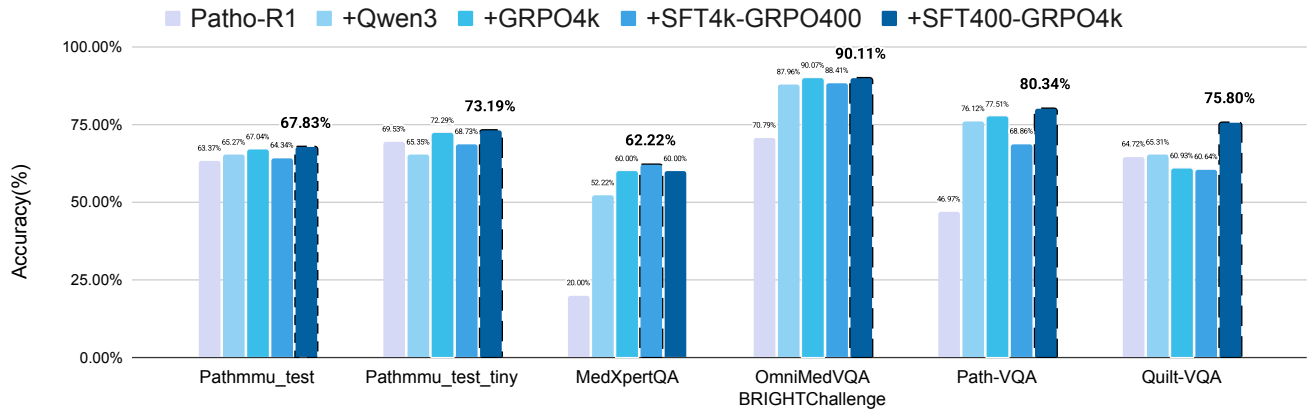

Figure 4: Ablation study results across multiple medical QA datasets.

Model	PathMMU-test					PathMMU-test-tiny				
	Atlas	EduContent	PathCLS	PubMed	SocialPath	Atlas	EduContent	PathCLS	PubMed	SocialPath
InternVL2-8B	43.68	44.86	23.77	44.56	45.40	46.63	50.59	21.47	49.11	51.38
InternVL2.5-8B	50.06	50.62	32.84	50.02	50.87	51.44	50.59	29.38	55.87	57.80
InternVL3-8B	54.07	50.80	39.09	54.04	53.32	58.17	54.90	42.94	57.65	60.55
Llama-3.2-11B-VI	41.05	37.49	26.72	38.82	39.21	45.19	38.04	29.38	39.50	41.74
Llama-3.2V-11B-cot	51.81	45.45	30.76	48.15	46.10	49.04	47.06	29.94	53.38	45.41
LLaVA-Onevision-7B	21.65	21.27	12.01	27.77	21.25	31.25	21.18	13.56	31.32	18.35
Qwen2.5VL-7B	41.18	43.20	24.82	42.77	39.67	44.23	49.41	24.86	44.84	40.83
Patho-R1-7B	75.34	66.43	45.40	66.06	67.93	**81.73**	75.29	44.63	72.24	67.89
Patho-AgenticRAG	**78.32**	**70.96**	**53.16**	**69.69**	**71.06**	79.33	**76.47**	**57.22**	**72.24**	**74.70**

Table 2: Comparison of model performance across multiple tasks. The left group shows results on PathMMU-test, and the right group on PathMMU-test-tiny. Best and second-best performances are bolded and underlined respectively.

Model	YorN		MedXpert	OmniMed
	Quilt	Path	Path	Bright
InternVL2-8B	60.56	61.36	10.00	40.56
InternVL2.5-8B	60.06	64.78	22.22	49.78
InternVL3-8B	33.82	18.56	15.56	65.28
Llama-3.2-11B-VI	63.27	63.50	13.33	47.08
Llama-3.2V-11B-cot	54.81	56.42	21.11	54.83
LLaVA-Onevision-7B	24.20	52.38	16.67	31.46
Qwen2.5VL-7B	52.19	41.82	12.22	43.60
Patho-R1-7B	64.72	46.97	22.00	70.79
Patho-AgenticRAG	**75.80**	**80.34**	**60.00**	**90.11**

Table 3: Performance comparison on Quilt-VQA, Path-VQA, MedXpert, and OmniMed.

rized in Tables 2 and 3. Patho-AgenticRAG achieves the best overall performance across most tasks, significantly outperforming both general-purpose vision-language models (e.g., InternVL3, Qwen2.5VL) and domain-specialized baselines such as Patho-R1-7B (Zhang et al. 2025). Specifically, Patho-AgenticRAG achieves +13.37% improvement on Quilt-VQA (75.80% vs. 64.72%) and +38.00% on MedXpertQA (60.00% vs. 22.00%) over Patho-R1. The largest margin appears on MedXpertQA, highlighting the importance of retrieval-augmented reasoning in knowledge-intensive tasks. On OmniMedVQA Bright Challenge, the model improves from 70.79% (Patho-R1) to 90.11%, a +19.32% increase, demonstrating substantial gains in both generalization and diagnostic precision. Details are in Appendix D.

5 Conclusion

We proposed Patho-AgenticRAG, a novel multimodal retrieval-augmented generation framework tailored for pathology diagnosis. By leveraging intelligent agents for dynamic querying of image-based vector databases, as well as employing task decomposition, query planning, and evidence aggregation, our approach significantly enhances the reasoning capabilities of vision-language models in pathology tasks. Our framework addresses the critical issue of hallucination in pathology diagnosis by promoting knowledge alignment, supporting evidence-based reasoning, and improving factual consistency in generated outputs. Patho-AgenticRAG demonstrates significant improvements over existing state-of-the-art multimodal models in key metrics, including answer precision and evidence traceability, representing a notable advancement in the integration of image content and reasoning for real-world pathology applications.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62303338).

References

Cheetirala, S. N.; Raut, G.; Patel, D.; Sanatana, F.; Freeman, R.; Levin, M. A.; Nadkarni, G. N.; Dawkins, O.; Miller, R.; Steinhagen, R. M.; et al. 2025. Less Context, Same Performance: A RAG Framework for Resource-Efficient LLM-Based Clinical NLP. arXiv:2505.20320.

Chen, C.; Weishaupt, L. L.; Williamson, D. F.; Chen, R. J.; Ding, T.; Chen, B.; Vaidya, A.; Le, L. P.; Jaume, G.; Lu, M. Y.; et al. 2025a. Evidence-based diagnostic reasoning with multi-agent copilot for human pathology. arXiv:2506.20964.

Chen, J.; Ouyang, R.; Gao, A.; Chen, S.; Chen, G. H.; Wang, X.; Zhang, R.; Cai, Z.; Ji, K.; Yu, G.; Wan, X.; and Wang, B. 2024. HuatuoGPT-Vision, Towards Injecting Medical Visual Knowledge into Multimodal LLMs at Scale. arXiv:2406.19280.

Chen, Y.; Shen, Y.; Huang, W.; Zhou, S.; Lin, Q.; Cai, X.; Yu, Z.; Shi, B.; and Qiao, Y. 2025b. Learning Only with Images: Visual Reinforcement Learning with Reasoning, Rendering, and Visual Feedback. arXiv:2507.20766.

Cho, J.; Mahata, D.; Irsoy, O.; He, Y.; and Bansal, M. 2024. M3docrag: Multi-modal retrieval is what you need for multi-page multi-document understanding. arXiv:2411.04952.

Faysse, M.; Sibille, H.; Wu, T.; Omrani, B.; Viaud, G.; Hudelot, C.; and Colombo, P. 2025. ColPali: Efficient Document Retrieval with Vision Language Models. In *ICLR*.

Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, H.; and Wang, H. 2023. Retrieval-augmented generation for large language models: A survey. arXiv:2312.10997.

Hu, Y.; Li, T.; Lu, Q.; Shao, W.; He, J.; Qiao, Y.; and Luo, P. 2024. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22170–22183.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2): 1–55.

Jabal, M. S.; Warman, P.; Zhang, J.; Gupta, K.; Jain, A.; Mazurowski, M.; Wiggins, W.; Magudia, K.; and Calabrese, E. 2024. Language Models and Retrieval Augmented Generation for Automated Structured Data Extraction from Diagnostic Reports. arXiv:2409.10576.

Jin, B.; Zeng, H.; Yue, Z.; Yoon, J.; Arik, S.; Wang, D.; Zamani, H.; and Han, J. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. arXiv:2503.09516.

Joshi, A.; Sarwar, S. M.; Varshney, S.; Nag, S.; Agrawal, S.; and Naik, J. 2024. Reaper: Reasoning based retrieval planning for complex rag systems. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, 4621–4628.

Lee, C.; Roy, R.; Xu, M.; Raiman, J.; Shoeybi, M.; Catanzaro, B.; and Ping, W. 2024. Nv-embed: Improved techniques for training llms as generalist embedding models. arXiv:2405.17428.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474.

Liu, H.; Son, K.; Yang, J.; Liu, C.; Gao, J.; Lee, Y. J.; and Li, C. 2023. Learning customized visual models with retrieval-augmented knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15148–15158.

Liu, P.; Ji, L.; Gou, J.; Fu, B.; and Ye, M. 2025. Interpretable Vision-Language Survival Analysis with Ordinal Inductive Bias for Computational Pathology. In *The Thirteenth International Conference on Learning Representations*.

Lu, M. Y.; Chen, B.; Williamson, D. F.; Chen, R. J.; Zhao, M.; Chow, A. K.; Ikemura, K.; Kim, A.; Pouli, D.; Patel, A.; et al. 2024. A multimodal generative AI copilot for human pathology. *Nature*, 634(8033): 466–473.

Malkov, Y. A.; and Yashunin, D. A. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(4): 824–836.

Pham, T.-H.; and Ngo, C. 2025. RARL: Improving Medical VLM Reasoning and Generalization with Reinforcement Learning and LoRA under Data and Hardware Constraints. arXiv:2506.06600.

Qian, C.; Acikgoz, E. C.; He, Q.; Wang, H.; Chen, X.; Hakkani-Tür, D.; Tur, G.; and Ji, H. 2025. Toolrl: Reward is all tool learning needs. arXiv:2504.13958.

Ravuru, C.; Sakhinana, S. S.; and Runkana, V. 2024. Agentic retrieval-augmented generation for time series analysis. arXiv:2408.14484.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv:2402.03300.

Singh, A.; Ehtesham, A.; Kumar, S.; and Khoei, T. T. 2025. Agentic retrieval-augmented generation: A survey on agentic rag. arXiv:2501.09136.

Sun, Y.; Si, Y.; Zhu, C.; Zhang, K.; Shui, Z.; Ding, B.; Lin, T.; and Yang, L. 2025. CPathAgent: An Agent-based Foundation Model for Interpretable High-Resolution Pathology Image Analysis Mimicking Pathologists' Diagnostic Logic. arXiv:2505.20510.

Sun, Y.; Wu, H.; Zhu, C.; Zheng, S.; Chen, Q.; Zhang, K.; Zhang, Y.; Wan, D.; Lan, X.; Zheng, M.; et al. 2024. Pathmmu: A massive multimodal expert-level benchmark for understanding and reasoning in pathology. In *European Conference on Computer Vision*, 56–73.

Thakrar, K.; Basavatia, S.; and Daftardar, A. 2025. Cultivating Multimodal Intelligence: Interpretive Reasoning and Agentic RAG Approaches to Dermatological Diagnosis. arXiv:2507.05520.

Wang, J.; Yi, X.; Guo, R.; Jin, H.; Xu, P.; Li, S.; Wang, X.; Guo, X.; Li, C.; Xu, X.; et al. 2021. Milvus: A Purpose-Built Vector Data Management System. In *Proceedings of the 2021 International Conference on Management of Data*, 2614–2627.

Wang, Q.; Ding, R.; Chen, Z.; Wu, W.; Wang, S.; Xie, P.; and Zhao, F. 2025. Vidorag: Visual document retrieval-augmented generation via dynamic iterative reasoning agents. arXiv:2502.18017.

Wu, J.; Deng, Z.; Li, W.; Liu, Y.; You, B.; Li, B.; Ma, Z.; and Liu, Z. 2025. MMSearch-R1: Incentivizing LMMs to Search. arXiv:2506.20670.

Wu, J.; Zhu, J.; Qi, Y.; Chen, J.; Xu, M.; Menolascina, F.; and Grau, V. 2024. Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation. arXiv:2408.04187.

Wu, R.-D.; Lin, Y.-Y.; and Yang, H.-F. 2024. Training-free Zero-shot Composed Image Retrieval via Weighted Modality Fusion and Similarity. In *International Conference on Technologies and Applications of Artificial Intelligence*, 77–90. Springer.

Xia, P.; Zhu, K.; Li, H.; Wang, T.; Shi, W.; Wang, S.; Zhang, L.; Zou, J.; and Yao, H. 2025. MMed-RAG: Versatile Multimodal RAG System for Medical Vision Language Models. In *The Thirteenth International Conference on Learning Representations*.

Yu, S.; Tang, C.; Xu, B.; Cui, J.; Ran, J.; Yan, Y.; Liu, Z.; Wang, S.; Han, X.; Liu, Z.; et al. 2024. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. arXiv:2410.10594.

Zeggai, A.; Traikia, I.; Lakehal, A.; and Boulesnane, A. 2025. AI-VaxGuide: An Agentic RAG-Based LLM for Vaccination Decisions. arXiv:2507.03493.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11975–11986.

Zhang, W.; Zhang, P.; Guo, J.; Cheng, T.; Chen, J.; Zhang, S.; Zhang, Z.; Yi, Y.; and Bu, H. 2025. Patho-R1: A Multimodal Reinforcement Learning-Based Pathology Expert Reasoner. arXiv:2505.11404.

Zhao, X.; Liu, S.; Yang, S.-Y.; and Miao, C. 2025. Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot. In *Proceedings of the ACM on Web Conference 2025*, 4442–4457.

Zhu, W.; Dong, X.; Li, X.; Qiu, P.; Chen, X.; Razi, A.; Sotiras, A.; Su, Y.; and Wang, Y. 2025. Toward Effective Reinforcement Learning Fine-Tuning for Medical VQA in Vision-Language Models. arXiv:2505.13973.

Zuo, Y.; Qu, S.; Li, Y.; Chen, Z.; Zhu, X.; Hua, E.; Zhang, K.; Ding, N.; and Zhou, B. 2025. MedXpertQA: Benchmarking Expert-Level Medical Reasoning and Understanding. arXiv:2501.18362.