Path-RAG: Knowledge-Guided Key Region Retrieval for Open-ended Pathology Visual Question Answering

Awais Naeem AWAIS.NAEEM@UTEXAS.EDU
University of Texas at Austin, USA

Tianhao Li TIANHAO@UTEXAS.EDU
University of Texas at Austin, USA

Huang-Ru Liao MIMILIAO2000@UTEXAS.EDU
University of Texas at Austin, USA

Jiawei Xu JIAWEIXU@UTEXAS.EDU
University of Texas at Austin, USA

Aby Mammen Mathew ABYMMATHEW@UTEXAS.EDU
University of Texas at Austin, USA

Zehao Zhu ZEHAOZHU@UTEXAS.EDU
University of Texas at Austin, USA

Zhen Tan ZTAN36@ASU.EDU
Arizona State University, USA

Ajay Kumar Jaiswal AJAYJAISWAL@UTEXAS.EDU
University of Texas at Austin, USA

Raffi A. Salibian RSALIBIAN@AOL.COM
University of California, Los Angeles, USA

Ziniu Hu BULL@CS.UCLA.EDU
University of California, Los Angeles, USA

Tianlong Chen TIANLONG@MIT.EDU
Massachusetts Institute of Technology, USA

Ying Ding YING.DING@ISCHOOL.UTEXAS.EDU
University of Texas at Austin, USA

Abstract

Accurate diagnosis and prognosis assisted by pathology images are essential for cancer treatment selection and planning. Despite the recent trend of adopting deep-learning approaches for analyzing complex pathology images, they fall short as they often overlook the domain-expert understanding of tissue structure and cell composition. In this work, we focus on a challenging Open-ended Pathology VQA (PathVQA-Open) task and propose a novel framework named Path-RAG, which leverages HistoCartography to retrieve relevant domain knowledge from pathology images and significantly improves performance on PathVQA-Open. Admitting the complexity of pathology image analysis, Path-RAG adopts a human-centered AI approach by retrieving domain knowledge using HistoCartography to select the relevant patches from pathology images. Our experiments suggest that domain guidance can significantly boost the accuracy of LLaVA-Med from 38% to 47%, with a notable gain of 28% for H&E-stained pathology images in the PathVQA-Open dataset. For longer-form question and answer pairs, our model consistently achieves significant improvements of 32.5% in ARCH-Open PubMed and 30.6% in ARCH-Open Books on H&E images. All our relevant codes and datasets will be open-sourced.

Keywords: Pathology, Medical Imaging, LLMs, HistoCartography

Data and Code Availability We use PathVQA He et al. (2020) dataset and introduce a custom dataset namely ARCH-Open (https://github.com/embedded-robotics/path-rag/tree/main/ARCH-OPEN), both of which are publicly available to other researchers for use and reproducibility. The code for PathRAG implementation is available here: https://github.com/embedded-robotics/path-rag.

Institutional Review Board (IRB) This research does not require IRB approval as it does not involve human subjects or data that would necessitate IRB oversight.

1. Introduction

Pathology image analysis plays an essential role in the diagnosis, treatment, and study of cancer Wang et al. (2019b). For example, reading hematoxylin and eosin (H&E)-stained slides is a clinical standard of care for the diagnosis and staging of many cancers Luo et al. (2017). The advent of whole-slide image (WSI) scanning, capable of capturing intricate histological details at high resolutions, has enabled remarkable advancements in pathology image analysis through deep learning algorithms to detect cancer regions Kather et al. (2016), classify tissue subtypes Javed et al. (2020), and identify diagnostically relevant structures and regions (e.g., nuclei Gamper et al. (2019), glands Graham et al. (2019), vessels and nerves Fraz et al. (2020)).

With recent advance of foundation model, a natural thinking is to utilize Visual Language Models (VLMs), like GPT4-V and LLAVA, for helping pathology analysis. However, studies show that on Open-ended Pathology visual question answering (VQA), recent state-of-the-art VLM LLAVA, even after fine-tuning, only achieves a recall of 38% Li et al. (2024).

The reason is because these VLMs are hard to identify fine-grained visual objects and textual entities, especially the intricate domain knowledge in pathology images. Therefore, in this paper, we resort to domain knowledge in pathology literature to help narrow down and identify key regions, and then use VLM as soft knowledge base to retrieve most important information from the image to be diagnosed. Specifically, we propose Path-RAG, refer to Retrieval Augmentation Generation for Pathology VQA.

The key question is which domain knowledge we'd like to retrieve. Since pathology images are complex and tightly connected with different tissue structures, normal deep learning approaches, that typically involve patch generation, patch-level feature extraction, and feature aggregation to generate image-level embeddings for pathology tasks, fall short as they often overlook crucial tissue structure and cell composition Jaume et al. (2021). The critical step for diagnosing pathology images is the pathologist's ability to identify tissue structure and examine the cellular composition, organization, and architecture of tissues to characterize features of normal structures and abnormal cells in histologic sections Bauer et al. (2020).

For analyzing this, cell graphs of cancer have been generated based on pathology images showing the distinct features compared with healthy cells at the cellular level and at the tissue level Gunduz et al. (2004). Cell graphs are one kind of entity graph where nodes and edges of the graphs denote tissue entities and their interactions. The entities can be nuclei, tissue regions, and glands. Entity graphs can simultaneously capture local entity environment and global tissue composition, illustrating an ideal way to capture domain-specific knowledge in pathology images. Entity graphs can improve the efficiency in diagnosis considering that only diagnostically relevant entities have been analyzed, instead of analyzing the entire pathology images Shaban et al. (2020). Generating entity graphs requires a complex workflow including stain normalization, tissue detection, entity detection, entity encoding, and constructing graph topology. HistoCartography is an open-source Python library that unifies a set of histology image manipulation tools to build entity graphs Graham et al. (2019); Ahmedt-Aristizabal et al. (2022), including stain normalization, tissue detection, nuclei detection, tissue component detection, feature extraction, and cell graph builders.

In Path-RAG, we utilize the existing HistoCartography tool to help decompose given image into several key region patches, and then use LLAVA-Med to summarize key information as well as candidate answer for each patch. The retrieved knowledge are fed into the final GPT-4 model for reasoning to provide the answer. In addition to using LLAVA-MED fine-tuned on PathVQA He et al. (2020), we also fine-tune LLAVA-MED on a custom dataset consisting of detailed open-ended question-answer pairs and pathology images which aids us in evaluating the efficacy of our knowledge-guided approach on detailed answers

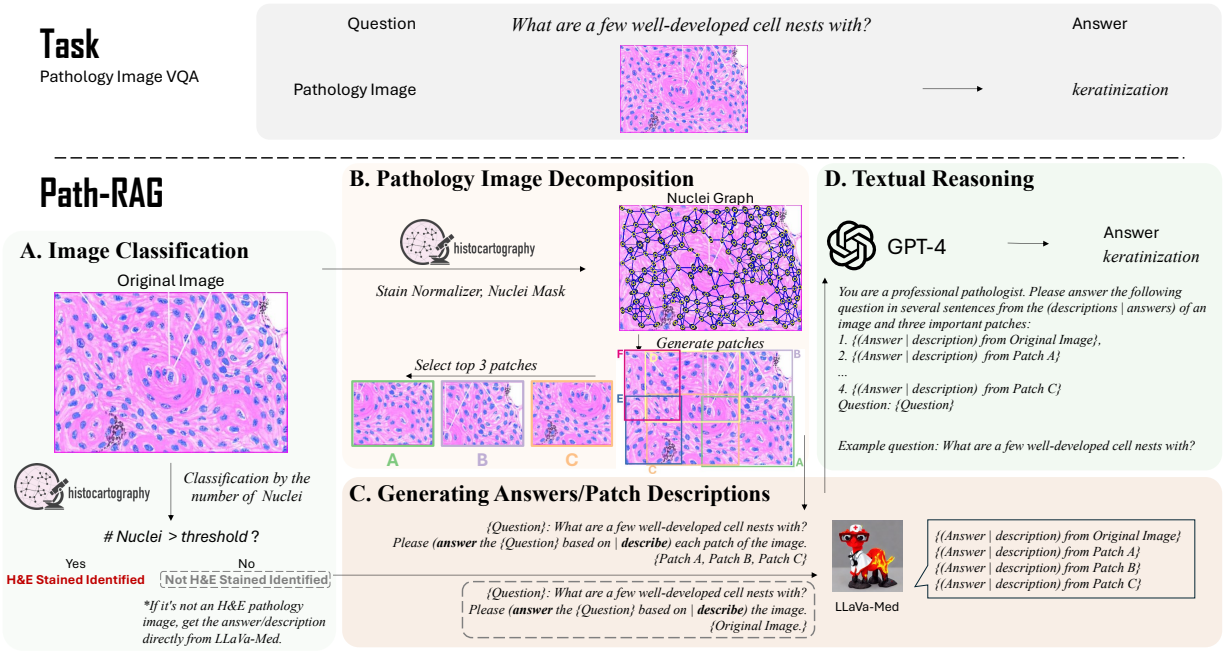

Figure 1: The overview of our Path-RAG framework.

unlike PathVQA where majority of the open-ended answers contain just 2-3 words. Our primary contributions can be summarized as:

- **Significant Performance Improvement with up to 28% increase in the recall.** Path-RAG is a human-centered AI approach by retrieving domain knowledge using HistoCartography to select the most relevant patches for a given pathology image, which can significantly boost the recall of LlaVa-Med for PathVQA Open-ended questions from 38% to 47%, especially with a significant improvement of 28% for H&E-stained pathology images. For longer-form question and answer pairs, our model consistently achieves a significant improvement of 32.5% in ARCH-Open PubMed and 30.6% in ARCH-Open Books on H&E images.

- **Consistent performance gain with and without GPT-4.** Path-RAG has demonstrated consistent performance benefits over LLaVa-Med for open-ended QA with and without GPT-4 component indicating the significance of domain guidance in pathology image analysis.

- **Performance improvement for non-finetuned LLaVa-Med.** Path-RAG can improve the recall by 12% when using LLaVa-Med not finetuned on PathVQA to generate descriptions as a prompt for GPT-4 to get the answer on H&E-stained slides.

- **Dataset having detailed Open-Ended Pathology Visual Question-Answer Pairs.** ARCH-Open dataset contains extensive open-ended question-answer pairs about pathology images which can be used for VQA systems and training/fine-tuning multi-modal LLMs for pathology related tasks.

2. Methodology

In our approach to handling pathology VQA, we are employing a careful ensemble of tools built specifically to process pathology images with large language models (LLMs) pre-trained on the pathology data. Our approach obviates any need to further fine-tune LLMs which incur high computational and memory overhead. We are *the first* to illustrate that using an off-the-shelf domain-aware tool can significantly benefit the intricate task of pathology open-ended VQA rather than investing heavily in compute resources, which happens to be the norm when dealing with multi-modal data in the pathology domain. Furthermore, we convert the multi-modal problem (including image and text) into a solely text-based problem and then use the textual reasoning capabilities of GPT-4 as the last step. This technique gives us

the leverage to use multiple LLMs; one fine-tuned to handle pathology VQA in a multi-modal setting Li et al. (2024) and another one specialized for textual reasoning OpenAI (2024).

Pathology images can be categorized and represented in discrete units such as tissues, nuclei, cells, glands, etc. To answer any question related to pathology, it is considered essential to understand the image regions which are heavily concentrated in any discrete unit under consideration. In fact, this approach is used by radiologists when they are reasoning about the question asked about any diagnostic image Bauer et al. (2020). To follow a similar practice of analyzing pathology images, we are utilizing HistoCartography Jaume et al. (2021) which is a toolkit containing pre-trained models for graph analytics in digital pathology. We are using this tool for two purposes: (1) Differentiate between pathology and non-pathology images (2) Extract image patches which are concentrated in nuclei structures.

LLaVA-Med Li et al. (2024) is a multi-modal language model fine-tuned on biomedical data and exhibits conversational capabilities to answer open-ended questions about a biomedical image. Since it can follow open-ended instructions to give relevant details about any biomedical image, we employ LLaVA-Med to generate descriptions/answers about the pathology image and/or its relevant patches. Finally, we use GPT-4 OpenAI (2024) to reason about the question using the textual descriptions generated via LLaVA-Med.

Our method uses a sequential approach to use these components as shown in Figure 1 and subsequently detailed in the following sections.

2.1. H&E stained Pathology Image Identification

As our approach relies on generating the patches of the pathology image depending on the presence of nuclei in the image, firstly we need to determine whether an image is an H&E stained pathology image or not. Using state-of-the-art approaches, a binary image classifier can be used to differentiate pathology images from non-pathology images. However, to avoid the use of an extra classifier, we have used HistoCartography Jaume et al. (2021) tool to determine this classification dynamically. We use a nuclei extractor API in HistoCartography to detect the number of nuclei in the image. For any non-pathology image, the extractor will result in zero or miniature number of nuclei-like components in the image. For our use case, we have used a hard threshold value of minimum 5 nuclei to differentiate between pathology and non-pathology image. Identifying any image as non-pathology gives us leverage to not use HistoCartography to divide this image into patches and directly advance towards LLaVA-Med to get the description/answer about the whole image. For the pathology images, however, the next step is to extract the relevant patches. This approach aligns with our focus on cancer-related pathology, since H&E stained images are gold standard for tumor diagnosis.

2.2. Pathology Image Decomposition

To extract the relevant patches of pathology images, we use different tools available in HistoCartography Jaume et al. (2021). Firstly, we use stain normalizer to overcome stain variability in the input image. Secondly, we detect nuclei present in the processed image using a nuclei extractor module to come up with the pixel locations of nuclei and their centers. Third, a deep feature extractor based on ResNet34 He et al. (2016) is used to extract the features of each nucleus by taking input from all the neighborhood image patches around each nucleus Jaume et al. (2021). The encoded nuclei features are then fed into the KNNGraphBuilder which builds a graph based on DGL Wang et al. (2019a) such that each nucleus is connected to 5 nearest neighbors with a maximum distance threshold of 50 pixels as per the configuration defined in HistoCartography Jaume et al. (2021) tool. In the resulting nuclei graph, each nuclei center is represented by node whereas the edges connect the nearest neighbor nuclei. The procedure is illustrated in part B of Figure 1.

Once the graph is obtained, a custom approach is taken to extract only those patches of the image which have maximum number of nuclei. In this approach, the pathology image is divided into a total of 9 patches having 20% overlap between them irrespective of the total dimensions of the image. It results in varying patch size for different images because of overlapping between successive patches up/-down and avoids miniature/big patches having fixed size. Finally, all the patches are ranked as per the number of nuclei centers contained in each patch and different numbers of top-ranked patches are considered for experiments and evaluation. To investigate if our domain-aware patch selection strategy is effec-

tive, we additionally perform ablation studies with random patches.

2.3. Patch Captioning and Candidate Answer Extraction

To generate the textual descriptions and candidate of each patch (we abbreviate as *answer/description* later), we resort to LLaVA-Med Li et al. (2024) which can generate an open-ended description of biomedical images. Once the key image regions are extracted as per the domain knowledge Jaume et al. (2021), For the pathology images, we provide each patch to LLaVA-Med to *generate the open-ended descriptions*, and then *generate a candidate answer* independently. This results into multiple *answer/description* pairs of each key region along with the original image, which will be later provided to GPT-4 for the final-stage reasoning. For the non-pathology images, we only use the input image to generate a single *answer/description* of the complete image. The procedure is illustrated in part C of Figure 1.

2.4. Reasoning over Per-Patch Extracted Knowledge

As the last step, we are using GPT-4 OpenAI (2024) to reason about the final answer given question and all *answer/description* of the patches generated in the last step. For the non-pathology images, we pass in a question along with the complete description of the input image to GPT-4 in order to get a final response. For the pathology images, we have the answers/descriptions of each patch along with the original image, so we consider two different approaches for experimentation/evaluation: *Path-RAG (answer)* and *Path-RAG (description)*. We pass in candidate answers and descriptions returned by LLaVA-Med, respectively, along with the question to GPT-4 and ask it to reason about the final answer. This response is the final answer which is returned as the final output of the system given a question and a pathology image as an input. The prompts given to GPT-4 for textual reasoning are added in supplmentary material. We also set a ablation baseline (*Path-RAG w/o GPT-4 (answer)*), which only consider our final answer as the concatenation of different answers returned by LLaVA-Med for each image/patch, without using GPT-4 for reasoning.

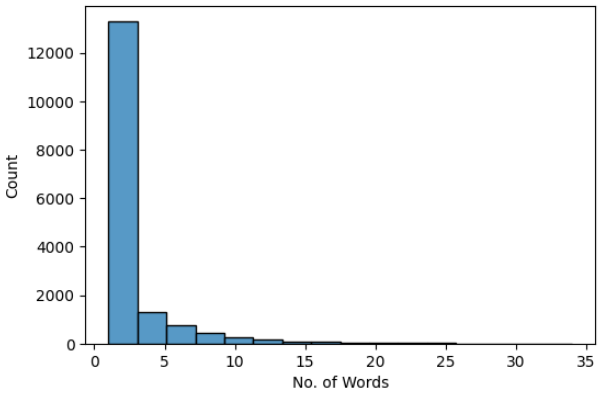

Figure 2: Word Count distribution of open-ended answers in PathVQA

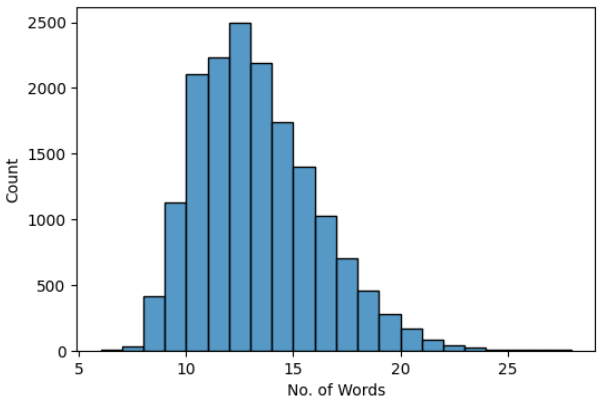

Figure 3: Word Count distribution in ARCH-Open Questions - PubMed

3. Experimental Setup

3.1. PathVQA Dataset.

PathVQA He et al. (2020) contains 32,799 question-answer pairs and 4,998 images from pathology textbooks. Each question has been manually checked to ensure correctness. For each image, several questions may be asked regarding multiple aspects such as location, shape, color, appearance, etc. All of the questions can be categorized into two types by whether the answer is yes or no (close-ended questions) and open-ended questions. We only consider open-ended questions in this work because SoTA methods Li et al. (2024); Saygin Seyfioglu et al. (2023) excel at yes/no questions, while performance on open-ended VQA re-

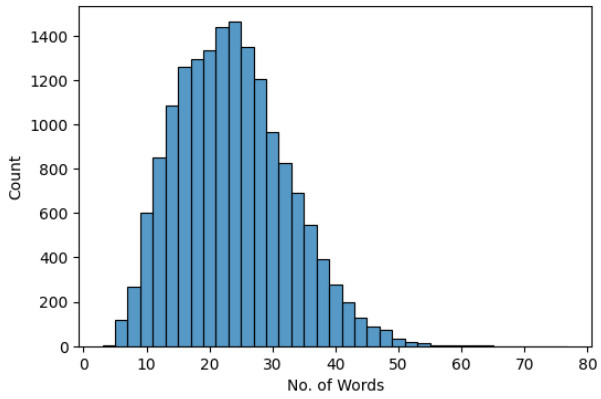

Figure 4: Word Count distribution in ARCH-Open Answers - PubMed

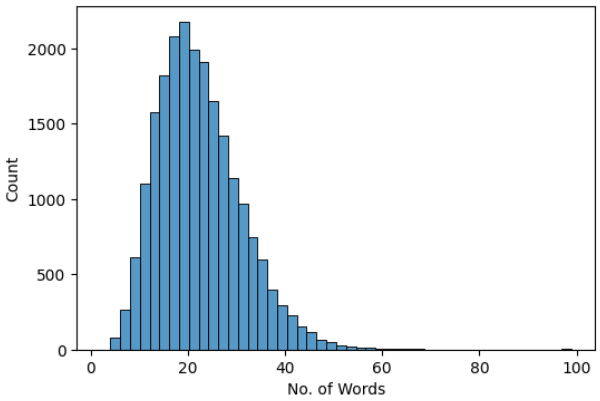

Figure 6: Word Count distribution in ARCH-Open Answers - Books

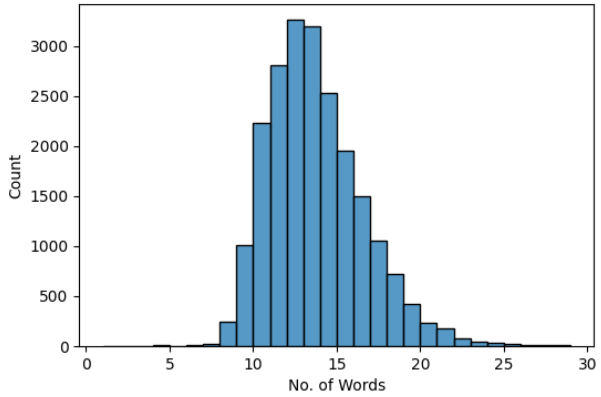

Figure 5: Word Count distribution in ARCH-Open Questions - Books

mains low. We use the official training, validation, and test splits in the dataset. Since we do not perform fine-tuning in our work, we only evaluate open-ended questions in the test set to establish consistency with other baselines. There are 849 images and 3,370 questions in the test set, in which 308 images with 1,127 questions are H&E-stained pathology images and 541 images with 2,243 questions are not H&E-stained pathology images.

3.2. ARCH-Open Dataset.

To address the lack of datasets accommodating the open-ended nature of pathology VQA, we constructed an open-ended VQA dataset namely ARCH-Open using ARCH Gamper and Rajpoot (2021). ARCH is a multiple instance captioning dataset for pathology images in which caption-image pairs are provided from relevant PubMed articles and books. There are a total of 4,270 captions and 4,305 images from PubMed articles, whereas a total of 3,309 image-caption pairs are extracted from the pathology books.

This dataset addressed the shortcomings of PathVQA He et al. (2020) and was subsequently used to assess the effectiveness of our PathRAG approach along with PathVQA. While PathVQA is the gold standard for open-ended VQA about pathology, it has significant flaws: answers are typically 2-4 words long (Figure: 2), and some questions can be answered without any image. We ensured each ARCH-Open question requires an image for a valid answer and the answer lengths align with the nature of open-ended (long-form) VQA task.

To construct ARCH-Open, we employed instruction tuning by providing image captions to GPT-4 OpenAI (2024) and then asked it to generate five open-ended question-answer pairs for each image-caption pair considering that caption text belongs to a pathology image (Appendix C). As a result, we got 21,350 VQA pairs from image-caption pairs extracted from PubMed articles (termed as ARCH-Open PubMed), and a total of 16,545 VQA pairs relating to image-caption pairs extracted from pathology books (termed as ARCH-Open Books). In ARCH-Open PubMed, the questions and answers have the average word length of 12.81 and 23.12, respectively (Figure: 3, 4), whereas the average word length for questions and answers in ARCH-Open

Books is 13.23 and 22.61, respectively (Figure: 5, 6). Both ARCH-Open PubMed and ARCH-Open Books datasets are split into train/test sets having 80/20 percent instances with only the test set being used for evaluation in our experiments. The quality checks for ARCH-Open are part of our ongoing efforts, involving human pathologists for thorough evaluation.

3.3. Implementation Settings

We use LLaVa-Med model and LLaVa-Med model fine-tuned on PathVQA training set to get image descriptions from image input and answers from image and question input. Prompts used are given in Appendix A, B. All inference processes were done on a single NVIDIA RTX A6000 GPU. We use GPT-4 (gpt-4-0125-preview) OpenAI (2024) from OpenAI API for the experiments that need to generate final answers after doing inference on LLaVa-Med model. The temperature of GPT-4 is set to be 0. Because LLMs' responses are stochastic and depend on the temperature, replication of the scores may be slightly different.

4. Results and Ablation Study

4.1. Comparison with SOTA:

In Table 1, we compare our method with existing generation-based methods on the open set of PathVQA He et al. (2020) datasets. Following previous works Li et al. (2024); Saygin Seyfioglu et al. (2023), we evaluate our performance using recall. To compare with our main baseline LLaVA-Med Li et al. (2024), we use both finetuned and not fine-tuned models on PathVQA, and explored different forms of textual input to the GPT-4 model. To highlight the effectiveness of our HistoCartography Jaume et al. (2021) patches retrieval, we categorized the testing data into H&E-stained pathology images or non-H&E-stained pathology images. Our major findings are: ① We achieve SoTA on PathVQA open-set dataset with **47.4% recall**. ② Our method demonstrates a significant improvement of **27.7% over LLaVA-Med** with fine-tuned model on H&E-stained pathology images, highlighting the importance of incorporating HistoCartography for pathology image analysis. ③ For models not fine-tuned on PathVQA, our method using description **outperforms LLaVA-Med by 11.6%**, providing further evidence of effectiveness of our approach.

In Table 2, we further evaluate our models on the ARCH-Open dataset. As mentioned in Section 3.2, ARCH-Open features longer forms of questions and answers, which better aligns with the real-world setting of VQA. It is worth noting that our method performs consistently well. On H&E images, we achieve a significant improvement of 32.5% in ARCH-Open PubMed and 30.6% in ARCH-Open Books with 3 patches from HistoCartography compared to the baseline Li et al. (2024) without patches from HistoCartography, demonstrating the effectiveness of our HistoCartography patches retrieval.

4.2. Effectiveness of Retrieving Patches with HistoCartography:

To demonstrate the effectiveness of our patches retrieval, we conducted a comparative analysis of three methods: employing no patches, using random patches, and our method, which utilizes domain relevant patches retrieved by HistoCartography Jaume et al. (2021) as presented in Table 3. For textual input as answer, we first compared scenarios where no patches are employed, our method demonstrates a significant recall improvement of 9.3%, emphasizing the critical importance of extracting the region of interest patches. These patches provide essential pathology features that can enhance the performance. Secondly, our method outperforms the utilization of random patches by 1.2% in terms of recall improvement, highlighting the necessity of incorporating domain knowledge to extract relevant patches. Notably, the presence of features such as the number of nuclei in a patch emerges as significant regions of interest for pathology multi-modal models. Such results highlight the significant advantages brought about by retrieving domain relevant patches using HistoCartography.

4.3. Effect of textual reasoning:

We ablate the components used for textual reasoning. The results are shown in Table 1. Firstly, we investigated the form of textual input to GPT-4. For the fine-tuned model, instructing LLaVA-Med to generate the answer directly and sending it to GPT-4 yields a better recall improvement of 15.9% compared to generating open-ended descriptions and forwarding them to GPT-4. This suggests that leveraging the explicit answer generation capability of fine-tuned LLaVA-Med enhances the effectiveness of GPT-4 in textual reasoning tasks. However, for not fine-tuned

Table 1: Comparison with prior state-of-the-art supervised methods on PathVQA datasets. Please note that we report our method using 3 patches. w/o GPT-4 (answer) refers to the Path-RAG directly concatenating answers without using GPT-4. (description/answer) refers to different textual input passed to GPT-4 for further reasoning.

Method	**Recall**		
	Not H&E pathology	**H&E pathology**	**All**
Not Fine-tuned on PathVQA			
Quilt-LLaVA Saygin Seyfioglu et al. (2023)	-	-	15.3
LLaVA-Med Li et al. (2024)	11.3	11.6	11.4
Path-RAG w/o GPT-4 (answer)	11.3	19.2	13.9
Path-RAG (description)	20.3	**28.5**	**23.0**
Path-RAG (answer)	11.3	25.9	16.2
Fine-tuned on PathVQA			
LLaVA-Med Li et al. (2024)	39.0	36.4	38.1
Path-RAG w/o GPT-4 (answer)	39.0	51.2	43.1
Path-RAG (description)	28.7	37.0	31.5
Path-RAG (answer)	39.0	**64.1**	**47.4**

Table 2: Comparison with prior state-of-the-art supervised methods on ARCH-Open datasets. Please note that we report our method using 3 patches. w/o GPT-4 (answer) refers to the Path-RAG directly concatenating answers without using GPT-4. (description/answer) refers to different textual input passed to GPT-4 for further reasoning.

Method	**Recall**		
	Not H&E pathology	**H&E pathology**	**All**
Eval on ARCH-Open PubMed			
LLaVA-Med Li et al. (2024)	53.5	52.2	52.9
Path-RAG w/o GPT-4 (answer)	53.5	77.6	65.5
Path-RAG (description)	75.9	80.8	**78.4**
Path-RAG (answer)	53.5	**84.7**	69.0
Eval on ARCH-Open Books			
LLaVA-Med Li et al. (2024)	52.7	54.0	53.7
Path-RAG w/o GPT-4 (answer)	52.7	79.5	72.3
Path-RAG (description)	76.0	76.6	**76.4**
Path-RAG (answer)	52.7	**84.6**	76.1

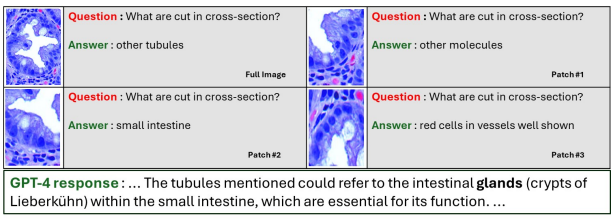

Figure 7: Example overview of Path-RAG with 3 patches. Detailed GPT-4 reasoning to come up with the correct answer (**bold**) is provided in Appendix A, B.

model, generating descriptions enhance the recall by 6.8% compared to generating answers directly. This indicates that, in the absence of fine-tuning, providing descriptions as input to GPT-4 may be more effective in facilitating textual reasoning tasks. Secondly, we explored the utilization of GPT-4 for textual reasoning as opposed to directly concatenating the four answers generated by LLaVA-Med. Our results demonstrate that employing GPT-4 for textual reasoning results in a higher recall improvement compared to directly concatenating the answers from different patches for both fine-tuned and not fine-tuned settings. This highlights the superiority of GPT-4's capabilities for textual reasoning tasks using in-context learning, enhancing the performance in open-ended pathology VQA.

Table 3: Effectiveness of patch retrieval methods in pathology images. Note that we report our method using a fine-tuned model. The first column indicates the form of textual input provided to GPT-4 for reasoning. * Indicating the paired bootstrapping values of Histo and Random patches, by sampling with replacement, calculating the mean estimate and confidence interval (CI) [lower bound, upper bound] over 10,000 iterations.

Form	**Patches**	**Recall (%)**		
		Not H&E pathology	H&E pathology	All
Description	No patches	28.7	30.0	29.1
Description	3 Random patches	28.7	33.9	30.4
Description	3 Histo patches	28.7	37.0	31.5
			*(+3.72, CI [1.95, 5.57])	
Answer	No patches	39.0	36.4	38.1
Answer	3 Random patches	39.0	60.4	46.2
Answer	3 Histo patches	39.0	**64.1**	**47.4**
			*(+3.07, CI [1.13, 5.04])	

Table 4: Effect of different number of patches. More patches yield improved performance. Note that we report results using fine-tuned model & GPT-4 textual reasoning.

Form	**Num Patches**	**H&E-stained**	**All**
Description	0	30.0	29.1
Description	3	37.0	31.5
Description	6	38.3	31.9
Answer	0	36.4	38.1
Answer	3	64.1	47.4
Answer	6	**66.9**	**48.4**

4.4. Effect of different number of patches

In Table 4, we evaluate the impact of employing different numbers of patches. Our findings suggest that incorporating more patches yields improved performance, indicating that employing a greater number of patches is beneficial.

5. Conclusion

Open-ended Pathology VQA is a challenging task that requires an understanding of intricate domain knowledge in pathology. Unlike prior works treating this task as simple VQA, our work for the first time argues to inject domain knowledge as guidance using HistoCartography to select the relevant and information-rich patches from pathology images. Our experiments and ablation studies show consistent performance improvement of this retrieval augmentation across multiple design choices. Our future work includes improving PATH-RAG's current patch selection strategies by considering their relevance with open-ended questions. We also aim to explore the performance benefits of Path-RAG for other datasets using vision-language models (GPT4-V OpenAI and et al. (2024), Qwen2-Vl Wang et al. (2024), LLaVA-Med++ Xie et al. (2024)etc) to capture complex features of the pathology images rather than GPT-4, a text-only model. For a more comprehensive evaluation, we plan to reference recent advancements Sun et al. (2023) that utilize the GPT-4 model as an oracle to analyze and rate responses across multiple aspects.

Acknowledgments

We would like to acknowledge the following funding supports: NIH OT2OD032581, NIH OTA-21-008, NIH 1OT2OD032742-01, NSF 2333703, NSF 2303038.

References

David Ahmedt-Aristizabal, Mohammad Ali Armin, Simon Denman, Clinton Fookes, and Lars Petersson. A survey on graph-based deep learning

for computational histopathology. *Comput. Med. Imaging Graph.*, 95(102027):102027, January 2022.

Thomas W Bauer, Cynthia Behling, Dylan V Miller, Bernard S Chang, Elena Viktorova, Robert Magari, Perry E Jensen, Keith A Wharton, Jr, and Jinsong Qiu. Precise identification of cell and tissue features important for histopathologic diagnosis by a whole slide imaging system. *J. Pathol. Inform.*, 11(1):3, February 2020.

M M Fraz, S A Khurram, S Graham, M Shaban, M Hassan, A Loya, and N M Rajpoot. FABnet: feature attention-based network for simultaneous segmentation of microvessels and nerves in routine histology images of oral cancer. *Neural Comput. Appl.*, 32(14):9915–9928, July 2020.

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

Jevgenij Gamper, Navid Alemi Koohbanani, Ksenija Benet, Ali Khuram, and Nasir Rajpoot. PanNuke: An open Pan-Cancer histology dataset for nuclei instance segmentation and classification. In *Digital Pathology*, Lecture notes in computer science, pages 11–19. Springer International Publishing, Cham, 2019.

Simon Graham, Hao Chen, Jevgenij Gamper, Qi Dou, Pheng-Ann Heng, David Snead, Yee Wah Tsang, and Nasir Rajpoot. MILD-Net: Minimal information loss dilated network for gland instance segmentation in colon histology images. *Med. Image Anal.*, 52:199–211, February 2019.

Cigdem Gunduz, Bülent Yener, and S Humayun Gultekin. The cell graphs of cancer. *Bioinformatics*, 20 Suppl 1(suppl_1):i145–51, August 2004.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

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

Guillaume Jaume, Pushpak Pati, Valentin Anklin, Antonio Foncubierta, and Maria Gabrani. Histocartography: A toolkit for graph analytics in digital pathology. In Manfredo Atzori, Nikolay Burlutskiy, Francesco Ciompi, Zhang Li, Fayyaz Minhas, Henning Müller, Tingying Peng, Nasir Rajpoot, Ben Torben-Nielsen, Jeroen van der Laak, Mitko Veta, Yinyin Yuan, and Inti Zlobec, editors, *Proceedings of the MICCAI Workshop on Computational Pathology*, volume 156 of *Proceedings of Machine Learning Research*, pages 117–128. PMLR, 27 Sep 2021.

Sajid Javed, Arif Mahmood, Muhammad Moazam Fraz, Navid Alemi Koohbanani, Ksenija Benes, Yee-Wah Tsang, Katherine Hewitt, David Epstein, David Snead, and Nasir Rajpoot. Cellular community detection for tissue phenotyping in colorectal cancer histology images. *Med. Image Anal.*, 63 (101696):101696, July 2020.

Jakob Nikolas Kather, Cleo-Aron Weis, Francesco Bianconi, Susanne M Melchers, Lothar R Schad, Timo Gaiser, Alexander Marx, and Frank Gerrit Zöllner. Multi-class texture analysis in colorectal cancer histology. *Sci. Rep.*, 6(1), September 2016.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024.

Xin Luo, Xiao Zang, Lin Yang, Junzhou Huang, Faming Liang, Jaime Rodriguez-Canales, Ignacio I Wistuba, Adi Gazdar, Yang Xie, and Guanghua Xiao. Comprehensive computational pathological image analysis predicts lung cancer prognosis. *J. Thorac. Oncol.*, 12(3):501–509, March 2017.

OpenAI. Gpt-4 technical report, 2024.

OpenAI and Josh Achiam et al. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Mehmet Saygin Seyfioglu, Wisdom O Ikezogwo, Fatemeh Ghezloo, Ranjay Krishna, and Linda Shapiro. Quilt-llava: Visual instruction tuning by extracting localized narratives from open-source histopathology videos. *arXiv e-prints*, pages arXiv–2312, 2023.

Muhammad Shaban, Ruqayya Awan, Muhammad Moazam Fraz, Ayesha Azam, Yee-Wah Tsang, David Snead, and Nasir M Rajpoot. Context-aware convolutional neural network for grading of colorectal cancer histology images. *IEEE Trans. Med. Imaging*, 39(7):2395–2405, July 2020.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented rlhf, 2023. URL https://arxiv.org/abs/2309.14525.

Minjie Wang, Da Zheng, Zihao Ye, Quan Gan, Mufei Li, Xiang Song, Jinjing Zhou, Chao Ma, Lingfan Yu, Yu Gai, Tianjun Xiao, Tong He, George Karypis, Jinyang Li, and Zheng Zhang. Deep graph library: A graph-centric, highly-performant package for graph neural networks. *arXiv preprint arXiv:1909.01315*, 2019a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024. URL https://arxiv.org/abs/2409.12191.

Shidan Wang, Donghan M Yang, Ruichen Rong, Xiaowei Zhan, Junya Fujimoto, Hongyu Liu, John Minna, Ignacio Ivan Wistuba, Yang Xie, and Guanghua Xiao. Artificial intelligence in lung cancer pathology image analysis. *Cancers (Basel)*, 11 (11):1673, October 2019b.

Yunfei Xie, Ce Zhou, Lang Gao, Juncheng Wu, Xianhang Li, Hong-Yu Zhou, Sheng Liu, Lei Xing, James Zou, Cihang Xie, and Yuyin Zhou. Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine, 2024. URL https://arxiv.org/abs/2408.02900.

Appendix A. Path-RAG Prompts

Prompts

Generate Answer from LLaVa-Med:

⟨question⟩ + ⟨image⟩

Get Response from GPT-4: You are a professional pathologist. Please generate a comprehensive answer that integrates multiple perspectives for a pathology image and provides a balanced conclusion based on the information provided.

- Perspective 1: *Answer from LLaVa-Med with full image input*
- Perspective 2: *Answer from LLaVa-Med with Patch 1 input*
- Perspective 3: *Answer from LLaVa-Med with Patch 2 input*
- Perspective 4: *Answer from LLaVa-Med with Patch 3 input*
- Question: *Question*

Generate Description from LLaVa-Med:

"Describe the following image in detail." + ⟨image⟩

Get Response from GPT-4: You are a professional pathologist. Please answer the question in several sentences based on the description of an image and three important patches.

- Description of image: *Description from LLaVa-Med with full image input*
- Description of patch 1: *Description from LLaVa-Med with Patch 1 input*
- Description of patch 2: *Description from LLaVa-Med with Patch 2 input*
- Description of patch 3: *Description from LLaVa-Med with Patch 3 input*
- Question: *Question*

Appendix B. Path-RAG Example

Path-RAG GPT-4 Prompt and Response

Prompt Design: You are a professional pathologist. Please generate a comprehensive answer that integrates multiple perspectives for a pathology image, providing a balanced conclusion.

- Perspective 1: "other tubules"
- Perspective 2: "other molecules"
- Perspective 3: "small intestine"
- Perspective 4: "red cells in vessels well shown"

Model Response: In the pathology image under consideration, we see a complex interplay of various elements requiring a detailed analysis from multiple perspectives to reach a comprehensive understanding.

From the first perspective, focusing on "other tubules," we observe structures likely part of the nephron system in the kidney. These tubules, when cut in cross-section, reveal essential architecture for filtration and absorption, suggesting a renal context.
The second perspective, "other molecules," hints at the biochemical aspect. This could involve identifying molecules within the tubules or tissues, indicating metabolic or pathological states.
The third perspective, "small intestine," introduces a gastrointestinal context with villi, essential for nutrient absorption. The structures could represent small intestine tissue if viewed cross-sectionally.
The fourth perspective, "red cells in vessels well shown," emphasizes vascularization, highlighting oxygen and nutrient delivery within the tissue.
In conclusion, the cross-sections likely represent the small intestine with a focus on its internal structure and vascularization. The tubules could represent intestinal glands, with visible molecules indicating enzymatic activity, and clear red cells in vessels underscoring vascular support.

Appendix C. Prompt for ARCH-Open Dataset Construction

ARCH-Open Dataset Prompt

Prompt Design: You are provided with a text description (figure caption) of a pathology image. Unfortunately, you don't have access to the original image. Your job is to generate a total of 5 open-ended question/answer pairs from this figure caption starting with "What" or "Where". Below are the requirements to generate the question/answer pairs:

- Requirement 1: Avoid quoting or referring to specific facts, terms, abbreviations, dates, numbers or names, as these may reveal the conversation is based on the text information, rather than image itself
- Requirement 2: Focus on the visual aspects of the image that can be inferred without the text information
- Requirement 3: Do not use phrases like "mentioned", "caption", "context", "without the image" in the question/answer pairs. Instead, refer to the information as being "in the image" or preferably don't mention anything
- Requirement 4: Ensure that question/anwer pairs are diverse and cover a range of visual aspects of the image
- Requirement 5: Answer responsibly, avoiding overconfidence, and do not provide medical advice or diagnostic information

Caption: {caption}
Question:
Answer:

Model Response:
Question: What type of cell morphology can be observed in the image?
Answer: The image likely displays spindle-shaped cells, which are a characteristic of certain types of tumors.

Question: What is the visual appearance of the cells in the image?
Answer: The cells in the image likely have a prominent eosinophilic appearance, meaning they are stained bright pink or red.

Question: What distinctive feature might some cells in the image show?
Answer: Some cells in the image might show prominent paranuclear vacuolisation, which is a clear or empty space around the nucleus of the cell.

Question: Where can we observe the fascicles of eosinophilic spindle cells in the image?
Answer: The fascicles of eosinophilic spindle cells are likely to be spread throughout the image, arranged in bundles or clusters.

Question: What similar characteristic could be inferred from the image when compared to leiomyosarcoma?
Answer: Just like in images of leiomyosarcoma, we might see prominent paranuclear vacuolisation in some of the cells in this image.