Integrating Human Expertise and Multi-Modal AI in Digital Pathology

Fatemeh Ghezloo

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2024

Reading Committee:
Linda Shapiro, Chair
Ranjay Krishna
Joann Elmore

Program Authorized to Offer Degree:
Computer Science and Engineering

University of Washington

Abstract

Integrating Human Expertise and Multi-Modal AI in Digital Pathology

Fatemeh Ghezloo

Chair of the Supervisory Committee:
Linda Shapiro
Computer Science and Engineering

Whole slide imaging (WSI) has revolutionized digital pathology, yet final diagnosis still relies heavily on pathologists' visual examination, which is often challenging due to the volume of data. Cancer mortality is significantly impacted by diagnostic errors and discordance among pathologists on the same case, which highlights the need for computer-aided diagnosis (CAD) systems to support pathologists in their clinical practices. One essential initial step in designing these systems is to understand the viewing behavior of pathologists. Prior research suggests that accurately identifying and interpreting regions of interest (ROIs) is essential for effective diagnosis. In this dissertation, we investigate the correlation between pathologists' viewing behaviors and diagnostic accuracy using viewport-tracking data for melanocytic skin lesions. Our analysis reveals a significant correlation between time spent viewing ROIs and diagnostic accuracy. Based on these findings, we propose a novel ROI detection method that integrates pathologists' viewing behaviors with deep learning, using an encoder-decoder architecture to predict pixel-level heatmaps of diagnostically relevant areas on WSIs. However, the scarcity and uni-modality of datasets in digital pathology limit the performance and interpretability of deep learning models. To address these issues, we introduce Quilt-1M, the largest vision-language dataset in histopathology, and Quilt-Instruct, an instruction-tuning dataset, both curated from educational YouTube videos. These resources enable the development of advanced multi-modal models: (1) QuiltNet, a

CLIP-based model that excels in zero-shot and few-shot image classification and image-text retrieval, surpassing state-of-the-art performance; (2) QUILT-LLAVA, a multi-modal model with enhanced spatial localization of medical concepts and complex reasoning; and (3) a multi-modal multi-agent diagnosis system that leverages the ROI detection model and QUILT-LLAVA to navigate large WSIs, gather evidence, and make diagnoses in a manner similar to pathologists. This dissertation advances digital pathology by linking viewing behaviors to diagnostic accuracy, presenting a novel ROI detection framework, and providing large-scale datasets and multi-modal models for improved CAD systems and diagnostic workflows.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

It has been a remarkable journey to finish my PhD, and it would not have been possible without the guidance, support, and encouragement of many individuals.

I would like to express my deepest gratitude to my advisor, Prof. Linda Shapiro. Thank you for giving me the opportunity to begin this journey and for believing in me. Your patience and guidance, especially in moments when I needed someone to believe in me, have been a foundation of my PhD experience. I am honored to have had the chance to work with you. Your continuous support, invaluable insights, and expertise have shaped my research and this dissertation. I am grateful for the knowledge and opportunities you have provided me.

I am extremely grateful to Prof. Ranjay Krishna. Your technical guidance has been crucial in advancing my work. Your commitment to excellence has inspired me to strive for the highest standards in my research. I am thankful for your encouragement and the time you spent providing valuable insights and direction. My heartfelt thanks also go to Dr. Joann Elmore. Your mentorship in clinical research has been truly invaluable. I am thankful for the opportunity to learn from you and for the generous amount of time you invested in my development.

To my amazing collaborators, Saygin Seyfioglu, Wisdom Ikezogwo, Dr. Beibin Li, Rustin Soraki, Dr. Oliver Chang, Dr. Stevan Knezevich, Dr. Kathleen Kerr, Dr. Tad Brunyé, Dr. Trafton Drew, Dr. Lisa Reisch, Dr. Kristin Shaw, Pin-Chieh Wang, Kia Gianni Thigpen, Dylan Geva, Fatwir Sheikh Mohammed, and Pavan Kumar Anand: thank you for your support and contributions to this body of work. To my lab mates Dr. Shima Nofallah, Dr. Nicholas Nuechterlein, Mahtab Bigverdi, Kalyani Marathe, Dr. Wenjun Wu, Kechun Liu, Sitong Liu, Yuguang Li, Zucks Liu, Nishat Khan, and Kaitlin Flores: thank you for all the

great conversations, kind words, and the support you offered.

I am thankful to my dear friends who became my second family far from home. Your encouragement and presence during the highs and lows of this journey have brought joy and comfort to my life. I am forever grateful for your friendship.

Finally, my deepest gratitude goes to my family. To my wonderful parents and my amazing sister Farimah, words cannot express how grateful I am for your unconditional love, understanding, and encouragement. Your support has been essential to my success. Thank you for standing by me through every triumph and challenge. I couldn't have achieved this milestone without your endless love and belief in me.

Thank you all for being a part of this journey.

DEDICATION

To my loving parents, Bahman and Susan,
and my dear sister, Farimah.

Chapter 1

INTRODUCTION

The field of pathology has undergone a significant transformation with the advent of Whole Slide Imaging (WSI) and the development of computer-based technologies aimed at improving diagnostic accuracy in recent years. Despite these advancements, the pathological examination remains the gold standard for final diagnosis, and even experienced pathologists sometimes make errors when faced with vast volumes of information on large WSIs [140, 34]. This variability in diagnostic outcomes highlights the need for advanced methods to assist pathologists in their interpretation processes and clinical practices.

A fundamental component of enhancing diagnostic accuracy lies in comprehending pathologists' visual search patterns. As pathologists work with WSIs, they perform intricate operations like panning, zooming, and focusing their attention on specific image characteristics. Traditional eye-tracking techniques, while effective in capturing these behaviors, are costly. As an alternative, viewport-tracking data—which records the pathologists' viewport coordinates—provides a cost-effective and scalable means of analyzing viewing behaviors. We can gain valuable insights into how pathologists' search strategies correlate with diagnostic accuracy by investigating their zooming and panning behaviors [34]. This knowledge is essential for developing tools and techniques that enhance pathologists' performance and reduce diagnostic errors.

Detecting regions of interest (ROIs) on WSIs is another crucial step in the diagnostic process. Pathologists typically focus their analysis on a few regions that contain the most diagnostically relevant information. Incorporating viewing behavior data into computer-aided diagnosis (CAD) systems can significantly enhance the detection of these ROIs, leading to more accurate and efficient diagnoses. Traditional methods for ROI detection often rely on

low-level image features, which may not align well with the visual characteristics observed by pathologists. By combining pathologists' qualitative expertise with deep learning models, we can create more robust and interpretable systems that highlight important areas on WSIs [33]. This approach not only improves the performance of CAD models but also reduces the computational cost and time required for diagnosis.

The complexity of histopathological information necessitates more expressive representations that capture the intricate patterns found in tissue samples. Vision-language models offer a promising solution by connecting visual features with natural language descriptions, providing additional signals beyond the scope of single diagnostic labels. However, the development of these models has been limited by the lack of large-scale, comprehensive datasets specific to histopathology. To address this gap, we present a new dataset comprising more than one million image-text pairs curated from educational videos and other public sources [51]. This dataset, along with an instruction-tuning dataset designed to enhance multi-modal reasoning capabilities [108], forms the foundation for training advanced vision-language models. These models can analyze images in detail, localize medical concepts spatially, and reason beyond individual patches, significantly advancing the field of histopathology.

As the number of cancer cases continues to rise globally, the traditional manual approach to diagnosing diseases through WSI examination is becoming increasingly unsustainable. Recent advancements in artificial intelligence (AI) have shown potential in transforming medical imaging diagnostics, but current models often fall short in capturing the holistic context of WSIs. To overcome these limitations, we propose PathFinder, a multi-modal, multi-agent framework that mimics the decision-making process of expert pathologists. By integrating AI agents specialized in navigation, description, and diagnosis, PathFinder enhances diagnostic efficiency and accuracy, offering a more intuitive and precise diagnostic process.

This dissertation explores the intersection of digital pathology, viewing behaviors, and advanced AI models to enhance diagnostic accuracy and efficiency in histopathology. This research aims to address critical challenges in the field and contribute to the advancement of pathological diagnostics. The findings and contributions presented in this work have

the potential to significantly impact clinical training programs, clinical practices, and the development of CAD systems, ultimately improving patient outcomes. In chapter 2, we describe the dataset of melanocytic skin biopsy images that is used in this work. Melanoma is one of the most aggressive forms of skin cancer and is responsible for the majority of skin cancer-related deaths due to its high potential for metastasis [34, 28]. Chapter 3 outlines our study on pathologists' viewing behaviors and their correlation with diagnostic accuracy. In chapter 4, we describe our novel ROI detection framework. In chapter 5, we summarize two joint works with Wisdom Ikezogwo and Saygin Seyfioglu on creating comprehensive vision-language datasets and multi-modal models for histopathology. Chapater 6 introduces the multi-modal multi-agent diagnosis framework using the ROI detection framework from chapter 4 and multi-modal datasets and models from chapter 5. In chapter 7, the conclusions of all the projects and possible future work are discussed.

Chapter 2

M-PATH DATASET

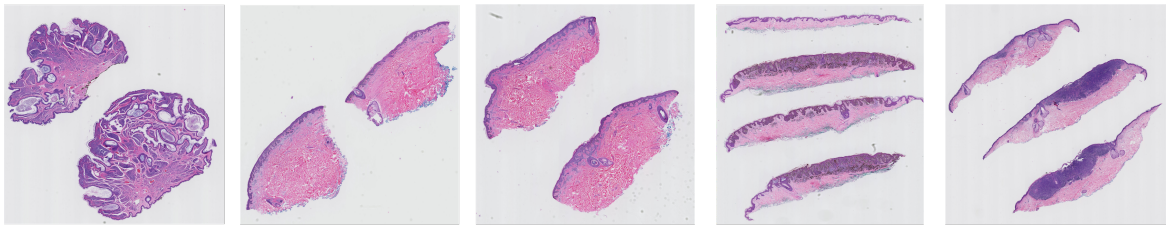
In this chapter, we outline the details of the datasets used in the following chapters. We provide a description on M-Path dataset, including skin biopsy WSIs, pathologists' characteristics, viewing behavior data, and consensus diagnosis and regions of interest.

2.1 M-Path Skin Biopsy WSIs

The skin biopsy WSIs in this dataset are from the prior M-Path study [28, 14] in which skin biopsy specimens of melanocytic lesions (N=240) were randomly selected from available stored specimens at Dermatopathology Northwest in Bellevue, Washington. The hematoxylin and eosin (H&E) stained slides were selected with stratification based on the patient's age and the original diagnosis. Each glass slide was scanned at 40x magnification using a Hamamatsu NanoZoomer 2.0-RS digital slide scanner to generate digital WSIs. These cases were classified into 5 diagnostic classes using the original MPATH-Dx scheme [94]. The number of biopsy cases in each class and example diagnostic terms for each class are as follows: 25 cases in class 1 (nevus/mild atypia), 36 cases in class 2 (moderate atypia/dysplasia), 60 cases in class 3 (severe dysplasia/melanoma in situ), 58 cases in class 4 (stage pT1a invasive melanoma), and 61 cases in class 5 (stage pT1b or higher invasive melanoma). One sample WSI from each class is presented in Figure 2.1.

2.2 Pathologists' characteristics

87 pathologists were recruited from 10 US states (California, Connecticut, Hawaii, Iowa, Kentucky, Louisiana, New Jersey, New Mexico, Utah, and Washington) to participate in the main M-Path study. Pathologists were eligible if they had completed residency and/or



(a) Class 1 (b) Class 2 (c) Class 3 (d) Class 4 (e) Class 5

Figure 2.1: MPath 5 classes H&E digital WSIs

fellowship training, had interpreted skin specimens in their clinical practices in the preceding year, and planned to do so for the next two years. Pathologists were invited to participate in a substudy of interpreting digital WSIs; 41 of the pathologists agreed to participate in the substudy, with 32 completing the study. All participants completed a baseline survey before the study to assess their demographic and clinical practice characteristics. Detailed explanation of these characteristics can be found in Table 2.1. After each interpretation, they were asked to provide their diagnosis using an online histology form using MPATH-Dx classification and to assess the difficulty level of the case and their confidence in their diagnosis.

2.3 Pathologists' Viewport Data

Pathologists' viewport data from the prior M-Path study [89] was collected using an online digital slide viewer that was developed using HD View SL, Microsoft's open-source Silverlight gigapixel picture viewer. The viewer allowed pathologists to pan around the image and zoom in and out up to x60 magnification. The web-based viewer automatically logged the viewport tracking data as pathologists viewed each slide. A viewport is a rectangular area of the image that is visible on the pathologist's computer screen at any time during their interpretation. For each interpretation (pair of pathologist and case), a list of viewport coordinates, magnification (zoom) level, and time stamps were recorded. This de-identified dataset includes viewport tracking data from two groups of pathologists: community pathologists and M-Path consensus

Table 2.1: Characteristics and demographic information of participating pathologists.

Pathologists' characteristics	Categories	Number(%)
Gender	Male	13 (41%)
	Female	19 (59%)
Age (years)	20 - 49	12 (37%)
	50 - 64	20 (63%)
Board certification/ Dermatopathology Fellowship training	Yes	10 (31%)
	No	22 (69%)
Experience with interpreting melanocytic skin lesions (years)	<5	3 (10%)
	5–9	10 (31%)
	10 – 19	9 (28%)
	> 20	10 (31%)
Caseload of melanocytic skin lesions (%)	< 10	14 (44%)
	10 – 24	13 (41%)
	25 – 49	5 (15%)
Ratings on difficulty level of interpreting melanocytic skin lesions	1 (Very easy)	-
	2	1 (3%)
	3	10 (31%)
	4	18 (56%)
	5	3 (10%)
	6 (Very challenging)	-
Ratings on confidence level of interpreting melanocytic skin lesions	1 (Not at all confident)	-
	2	2 (6%)
	3	4 (12%)
	4	7 (22%)
	5	16 (50%)
	6 (Extremely confident)	3 (10%)

reference panel. Community pathologists who were recruited for the M-Path Study had completed residency and/or dermatopathology post-doctoral training, had interpreted skin specimens in their clinical practices in the preceding year, and planned to do so for the next

two years. Three dermatopathologists participated in this study as members of the M-Path consensus panel, each with expertise in cutaneous melanocytic lesions (see section 2.4). Each of the pathologists from these two groups viewed and diagnosed these cases independently, and their viewport logs are available. Each case in our dataset was interpreted by one consensus reference panel dermatopathologist and an average of five community pathologists.

2.4 Consensus Reference Diagnosis and Relationship to Diagnostic Accuracy

The consensus reference panel of three dermatopathologists with internationally recognized expertise independently interpreted the full set of 240 cases in glass slide format, and then participated in a series of six full-day review meetings as part of the earlier M-Path study [14]. Utilizing a multi-headed microscope during the review meetings, they agreed on a consensus diagnosis for each case using a modified Delphi approach [14] and wrote case guidelines together for each of the 240 cases. Cases were then digitized, and an additional dermatopathologist from the M-Path research team joined the panel to determine a consensus rectangular region as the ROI for each case. ROIs were selected by the expert dermatopathologists as the area that best supported their diagnosis and best represented the critical features on the slide, as described in the aforementioned case guidelines. These variable-sized ROIs provide valuable, diagnostically important information, and can be extracted using their coordinates. We evaluate the diagnostic accuracy by assessing the agreement between the diagnosis provided by community pathologists and the consensus diagnosis determined by our panel of three internationally recognized dermatopathologists. Diagnostic error is a metric used to measure the divergence between a pathologist's diagnosis and the consensus diagnosis. For instance, if the consensus diagnosis is class 3 and the pathologist's diagnosis is class 2 or class 4, it would be considered a 1-class error. Note that these are the diagnostic accuracy and error of the pathologists and are unrelated to the accuracy of the proposed methods in the following chapters.

Chapter 3

INSIGHTS FROM PATHOLOGISTS VIEWING BEHAVIOR

3.1 Introduction and Motivation

As the use of WSIs in pathology has increased, new computer-based technologies have been developed to improve diagnostic accuracy. However, the pathological examination remains the gold standard for ultimate diagnosis. Even expert pathologists make mistakes when confronted with vast volumes of information on large slides [140]. According to studies, pathologists' diagnosis differ even when they notice similar features on a biopsy sample slide. One of the leading causes of death is an error in cancer diagnosis [28]. Diagnostic accuracy varies according to information processing frameworks due to the combination of case data, pathologist features, and the visual search process that characterizes interpretation [13]. With the advancement of digital imaging and the use of WSI in pathology, it is now possible to monitor pathologists' viewing behaviors and investigate how different search patterns affect diagnosis accuracy.

As a pathologist interacts with a case, the visual search process involves zooming and panning patterns, and allocating visual attention to image characteristics. The former is assessed using the viewport coordinates of a computer screen, whilst the latter is measured using eye-tracking devices. Although eye-tracking techniques are the most common and direct method of recording viewing behaviors, they are costly and complex, necessitating in-person data collection. On the contrary, viewport-tracking data can be gathered at a lower cost and a large scale. Data from viewport tracking reveals pathologists' movement, zooming behavior, and total interpretation time, providing insight into pathologists' attention.

Various studies have been conducted using eye-tracking and mouse-tracking analysis in the fields of radiology and pathology. These studies have shown significant predictive factors

during the visual search process and diagnostic accuracy. Previous research in radiology [24] and breast pathology [80] demonstrates that one of two search strategies is used by physicians when evaluating medical images: drilling or scanning. A pathologist with a drilling strategy focuses on a specific area and uses magnification settings to zoom in and out at various locations, while a pathologist with a scanning strategy utilizes a fixed zoom level while searching and panning over a wide area of interest [24, 80]. While the obtained strategy was not a predictor of diagnostic accuracy in breast pathology, some association among pathologists' characteristics and their search strategy was found [80]. Also, better performance in searching volumetric images in radiology using a drilling strategy was detected [24]. No previous study has investigated the use of these two search strategies in dermatopathology.

In this chapter we outline the types of data points that can be gathered to describe pathologists' viewing behavior using viewport data. We investigated zooming and panning behaviors of 32 pathologists viewing slidesets of 36 digital melanocytic skin cases. Pathologists were instructed to view and diagnose a set of cases assigned to them. The recording of these viewing sessions resulted in a total of 1073 interpretations. For each interpretation, we used the recorded information such as coordinates of the viewports on their screen, the zoom levels used to view a scene and the timestamps to define several viewing variables. These variables measure and quantify pathologists' interactions with the digital slide such as zooming and panning patterns, total interpretation time, and attending to the experts' consensus ROI. We then used these to address the association of the pathologists' search strategy and diagnostic accuracy in melanocytic skin lesions on whole slide images. Our main questions are:

- How are pathologists' characteristics associated with specific viewing patterns?

- How do viewing patterns change as pathologists gain more expertise in diagnosing melanocytic lesions?

- How do specific viewing behaviors and search patterns contribute to diagnostic accuracy?

3.2 Related work

3.2.1 Viewing Behavior Analysis

Diagnosis of pathology slides is a complex task, and pathologists go through years of training to be able to make a diagnosis. Traditionally, a pathologist views a skin biopsy on a glass slide using a microscope, trying to identify critical regions and visual features in the biopsy. These features at the region and cellular levels can be subtle and perplexing. Even experienced pathologists are prone to errors when confronted with massive amounts of information in large slides [140]. Studies show that there is discordance among pathologists' diagnoses even when they observe similar features on a biopsy sample slide. Errors made in cancer diagnosis are one of the main reasons that cause death [28]. Considering the significant impact of diagnostic errors on patients, it is important to understand the underlying reasons for these errors. Studying pathologists' viewing behaviors and search strategies and how these behaviors contribute to their diagnostic accuracy could be beneficial for both educational and clinical purposes.

Digital pathology has enabled researchers to investigate pathologists viewing behaviors. Instead of utilizing a microscope, pathologists can now do their examinations using a virtual slide viewer on their computer screens. The interpretation procedure when using WSI is different from the traditional use of a microscope, which can significantly change the pathologists' interpretive behavior [58]. The visual search process conveys how a pathologist interacts with a case using zooming and panning patterns, as well as allocating visual attention to image characteristics. The former is measured by computer screen's viewport coordinates whereas eye-tracking devices are used to measure the latter. Although eye-tracking techniques are the most common and direct method of recording viewing behaviors, they are expensive and complex. On the contrary, mouse-tracking data can be obtained on a large scale at a lower cost. Mouse-tracking data provides information on pathologists' movement, zooming behavior, and total interpretation time, offering insight into pathologists' attention. [100] showed that mouse-tracking data might be used to investigate pathologists' diagnostic accuracy and

efficiency when using WSI.

Previous research studied pathologists' viewing behavior while viewing skin biopsy images in various ways. In an eye-tracking study, residents and experts spent a similar amount of time viewing slides at low magnification. In contrast, at high magnifications, residents spent significantly more time [121]. This suggests that the residents may have experienced difficulty integrating their findings at high magnifications into the overall image context. To investigate this suggestion, [77] used "search maps" to compare the search characteristics of pathology experts and residents at low, medium, and high magnification levels while exploring inflammatory skin disease virtual slides. A search map is composed of three components based on the magnification level used by the pathologist. Areas were labeled as low, medium, and high magnification, referring to a magnification level of less than 4x, between 4x and 10x, and more than 10x, respectively. The study reveals that residents and experts saw roughly the same slides at low magnification, whereas experts saw much less at high magnification. Their results suggest that the expert pathologists knew where to search for information confirming or disproving a specific diagnostic theory. Therefore, it can be assumed that initial diagnostic hypotheses are formed at low magnification but are resolved at high magnification.

Research has demonstrated that pathologists focus their visual attention on certain image regions that are more salient, and their experience level affects whether these viewed regions can be appropriately interpreted and used to arrive at an accurate diagnosis [13]. Previous studies in this area can be summarized in two main outcomes. First, accurate interpretation by a pathologist is dependent on being able to recognize significant histopathological characteristics. Each case contains Regions of Interest (ROIs) that are visually appealing to all viewers and provide diagnostic information. Studies suggest that pathologists are more likely to agree with an expert reference diagnosis when they locate an overlapping diagnostic image region [82]. Second, visual diagnosis expertise is derived from prior experience and knowledge of exemplars and pathologists' search strategies [63]. Experts are better at searching, processing, and interpreting larger perceptual units because they can perceive these units more rapidly and effectively as configurations or chunks of information instead of individual parts. Visual

search patterns have been studied for a better understanding of how expertise develops both in radiology and pathology and they suggest that improved pattern recognition and better allocation of attention and visual processing resources are also significant aspects of expertise development [62, 77, 45, 118].

3.2.2 Histopathology Image Analysis

Melanoma diagnosis and early detection have changed dramatically during the previous few decades. Digitizing entire glass slides became possible with the advent of slide scanners. The process of scanning histopathology, immunohistochemistry, or cytology slides with whole-slide scanners, as well as the interpretation, management, and analysis of these digitized whole-slide pictures using computational methodologies, is referred to as digital pathology [20]. The US Food and Drug Administration (FDA) has approved several slide scanning devices for whole-slide imaging for use in clinical settings.

The traditional pathological analysis requires specially trained pathologists to look for areas of interest under the microscope one by one, and then analyze and diagnose based on professional knowledge. Traditional analysis of pathological images has many drawbacks. There are no quantitative indicators, so the qualitative analysis results cannot be reproduced [140]. Moreover, most doctors have tight working conditions, heavy workload, and time pressure. In this case, the human cognitive process is easily disturbed, leading to incomplete diagnosis and misdiagnosis [28]. Although traditional slide analysis is accurate, it can be deeply personal. It is possible for the same pathologist to evaluate a slide one day and to get different conclusions the following week. Besides, the procedure is a challenging and time-consuming task [128]. CAD systems are now rapidly developing and can help pathologists improve diagnosis accuracy and detection rate and reduce the overall misdiagnosis rate. Moreover, the computer is not affected by fatigue and human error and provides better assistance to doctors [28, 13]. It is also a valuable tool to reduce the workload of clinicians [14].

Recent literature in computer-aided melanoma diagnosis systems has witnessed significant

advancements. CAD systems can provide prognostic and diagnostic information on skin biopsy images, including the detection of cellular level entities, segmentation of clinically important tissue structures, and other important factors toward the accurate diagnosis of skin biopsy images [87]. Histopathological image analysis research addresses a range of issues associated with disease diagnosis. These include tasks such as detecting nuclei within images [130, 83], identifying regions of interest [78, 68], predicting clinical variables such as diagnosis [96], grade [85], and survival time [136].

Researchers have explored various techniques, including deep learning models, to improve the accuracy and efficiency of melanoma detection. Convolutional neural networks (CNNs) have been widely employed, demonstrating their capability to learn discriminative features from skin lesion images. Transfer learning, utilizing pre-trained models such as VGGNet [111] and ResNet [43], has also gained attention, enabling effective feature extraction and improving generalization. Additionally, advanced techniques like attention mechanisms in [129], generative adversarial networks (GANs) in [141], and ensemble models in [2] have been investigated to further refine the classification results. The emerging trend of explainable AI has also been incorporated, aiming to provide interpretable and transparent decision support. Overall, the recent literature reflects continuous progress in developing computer-aided melanoma diagnosis systems, emphasizing the potential for enhanced accuracy, efficiency, and clinical utility in melanoma detection and diagnosis.

3.3 Methods

3.3.1 Viewing Behavior Quantification

M-Path study data was used to perform this study. The description of M-Path skin biopsy specimens of melanocytic lesions, pathologists' characteristics and viewport data, and consensus reference panel are explained in sections 2.1, 2.2, 2.3, and 2.4 respectively. A subset of 180 cases were selected from M-Path dataset for this study. The distribution of these 180 cases among the 5 classes is as follows: 8.3% class 1, 16.7% class 2, 25.0% class 3, 25.0%

class 4, and 25.0% class 5. These cases were grouped into 5 sets of 36 digital melanocytic skin cases. A group of 32 pathologists participated and completed this phase of the study and were randomly assigned to one of the 36 slidesets. Pathologists were instructed to view and diagnose their assigned cases. The recording of these viewing sessions resulted in a total of 1073 interpretations [1]. Each interpretation consists of a series of viewport information including viewport coordinates and dimensions (x and y of the top left corner and height and width of the viewport), magnification level used at that viewport and the timestamp. We use this information to define variables that quantify pathologists' viewing behaviors. Entries associated with a duration of more than 1 minute (frozen at one location without any activity for more than 1 minute) were excluded due to the assumption that the pathologist was not actively interpreting during that time. The proposed variables are briefly described below:

Total interpretation time: Using the time stamp (TS) of each viewport (v_i), we calculated the duration (d) of each viewport being viewed. Total interpretation time (T) is calculated by summing the duration of all the viewports.

$$\begin{aligned} d(v_i) &= TS(v_{i+1}) - TS(v_i) \\ T &= \sum_{v_i=1}^{v_n} d(v_i) \end{aligned} \tag{3.1}$$

Average zoom level, maximum zoom level, and zoom level variance: The web-based viewer allowed pathologists to zoom from $\times 1$ to $\times 60$. Viewport tracking logs provided a variable number of zoom level values for each interpretation based on pathologists' interpretive behavior; thus, summary statistics were used to describe zoom level behavior during each interpretation. Average (avg), maximum (mx), and variance (var) of zoom levels were calculated for each interpretation.

---

[1] A group of 41 pathologists participated in the digital phase of the M-Path study, each interpreting 36 cases. A small number of interpretations (N=79) were not recorded due to a glitch in the system. After pre-processing the data, only 1073 interpretations are usable for this study.

$$
\begin{aligned}
avg &= \frac{\sum_{v_i=1}^{v_n} zoom(v_i)}{n} \\
mx &= max\{zoom(v_i) : i = 1..n\} \\
var &= \frac{\sum_{v_i=1}^{v_n} (zoom(v_i) - avg)^2}{n-1}
\end{aligned}
\tag{3.2}
$$

Scanning Percentage: We defined scanning percentage (SP) similar to the Mercan et al. [80] study for digital breast pathology. Scanning percentage quantifies the panning behavior where scanning refers to the behavior of fixing the zoom level and panning around the image. Higher scanning percentage is achieved when a pathologist pans across the image using a constant zoom level whereas a lower scanning percentage is achieved when a pathologist zooms in and out at various locations.

$$
\begin{aligned}
s(v_i) &= \begin{cases} 1, & zoom(v_{i+1}) == zoom(v_i) \\ 0, & otherwise \end{cases} \\
SP &= \frac{\sum_{v_i=1}^{v_n} s(v_i)}{n}
\end{aligned}
\tag{3.3}
$$

ROI Time Percentage: ROI time percentage (RTP) measures the amount of time a pathologist spends viewing regions that experts marked as ROI. When a pathologist's viewport intersects with the consensus ROI by 40% or more, we consider that the pathologist is viewing the consensus ROI (r). However, to ensure that the pathologist is actually attending to the ROI, we apply a size constraint. We exclude cases where the ratio of the ROI area to the viewport area is smaller than 10%. This way we make sure that a viewport intersects with a large area of the ROI and this intersection covers the most parts of the viewport. ROI time percentage calculates the proportion of interpretation time spent viewing such regions. Pathologists were not informed about the consensus ROI at the time of interpretation, spending more time on such regions means that they independently identified the region as important.

$$s(v_i) = \begin{cases} 1, & \frac{area(intersect(v_i, ROI))}{area(ROI)} > 0.4 \quad and \quad \frac{area(ROI)}{area(v_i)} > 0.1 \\ 0, & otherwise \end{cases} \tag{3.4}$$
$$RTP = \frac{\sum_{v_i=1}^{v_n} r(v_i)}{n}$$

Magnification percentage: Magnification percentage (MP) is calculated based on the number of times a pathologist zooms in consecutively. This variable captures how deeply and frequently a pathologist zooms while interpreting a case. We count the number of viewports that are associated with consecutive zoom-in behavior (m). A consecutive zoom-in is a sequence of viewports where the zoom level of each viewport is greater than its previous viewport in the sequence. Magnification percentage calculates the proportion of viewports associated with this behavior.

$$m(v_i) = \begin{cases} 1, & zoom(v_{i+1}) \geq zoom(v_i) \\ 0, & otherwise \end{cases} \tag{3.5}$$
$$MP = \frac{\sum_{v_i=1}^{v_n} m(v_i)}{n}$$

3.3.2 Statistical Analysis

Both case and pathologist contribute to the variation of the outcome of our models. Due to this crossed-level structure of cases and pathologists in our dataset, we used the cross-classified multilevel model [10] to address our study's questions. To investigate possible associations between pathologists' demographics and clinical characteristics and their viewing behaviors, we used a cross-classified multilevel model. For each model, we used one of the pathologists' characteristics shown in Table 2.1 as the explanatory variable and each of the viewing behaviors defined in section 3.3.1 as the outcome. The notation of the model is defined in Expression 3.6; y_i denotes the viewing behavior variable of interpretation i, x_i

denotes the pathologist's demographic/clinical characteristic, $u_{pathologist(i)}$ and $u_{case(i)}$ indicate the pathologist and case random effects, and e_i denotes the interpretation-level residual error.

$$y_i = \beta_0 + \beta_1 x_i + u_{pathologist(i)} + u_{case(i)} + e_i \tag{3.6}$$

To investigate associations between pathologists' viewing behavior and diagnostic accuracy, we used an analogous generalized linear mixed model with logit link. We define diagnostic accuracy as the binary agreement of a pathologist's diagnosis with the consensus reference diagnosis. For each univariate model, we used one of the viewing behaviors defined in section 3.3.1 as the explanatory variable of interest and diagnostic accuracy as the outcome. To control for pathologist experience or expertise, all models also included pathologists' years of experience with melanocytic skin lesions (categorical covariate with 4 levels) and having board certification and/or fellowship training (binary variable). In Expression 3.7; P_i denotes the probability of an accurate diagnosis for interpretation i, $logit(p) = log(\frac{p}{1-p})$, and x_i denotes the viewing behavior, E_2, E_3, and E_4 are indicators of the second, third, and fourth levels of pathologists' years of experience, and F indicates board certification and/or fellowship training.

$$logit(P_i) = \beta_0 + \beta_1 x_i + \beta_2 E_2 + \beta_3 E_3 + \beta_4 E_4 + \beta_5 F + u_{pathologist(i)} + u_{case(i)} + e_i \tag{3.7}$$

To further analyze the association of these viewing behaviors with diagnostic accuracy in the presence of each other, a subset of variables was chosen as explanatory variables to study using a multivariate model. Scanning percentage was chosen for the panning behavior, zoom variance for the zooming behavior, ROI time percentage for the interaction with consensus ROI behavior, and total interpretation time to address the diagnostic efficiency. The multivariate model included the same covariates as the univariate models to study accuracy. SAS version 9.4 (SAS Institute, NC) was used to perform all the statistical analyses in this study.

3.4 Results

Pathologists' viewing behaviors and characteristics: We hypothesized that pathologists with different characteristics might demonstrate different viewing behavior. To investigate this, we modeled the association between pathologist characteristics (Table 2.1) and viewing behaviors variables (section 3.3.1). In this section, we present results with a P-value < 0.1 and interpret the type of association between the predictor and outcome variables based on the Contrast value. The results (Table 3.1) suggest that pathologists with a board certification and/or fellowship training and those with a higher caseload of melanocytic skin lesions have lower average, maximum, and variance of zoom levels (higher caseload and having board certification are higher order and Contrast < 0). In addition, pathologists with lower confidence level in interpreting melanocytic skin lesions have higher average, maximum, and variance of zoom levels (lower confidence level is higher value and Contrast > 0). Lastly, pathologists in a higher age range have higher maximum and variance of zoom levels (higher age range is higher order and Contrast > 0). No significant associations were found among other viewing behaviors and pathologists' characteristics.

Diagnostic accuracy: To study associations among viewing behaviors and diagnostic accuracy, we used a series of cross-classified multilevel models. Seven separate models were generated for each of the defined viewing behavior variables in section 3.3.1. The Odds Ratio (OR) and the P-value of each model are shown in Table 3.2. The odds ratio is a measure of association between a predictor variable and an outcome. It quantifies the likelihood of an outcome occurring with the exposure of the predictor variable compared to the odds of the outcome occurring without that exposure. All viewing behaviors show a statistically significant association with diagnostic accuracy (P-value < 0.05), except for scanning percentage which was marginally significant ($0.05 <$ P-value < 0.1). Except for magnification percentage, each viewing behavior was positively associated with accuracy (adjusted OR >1), meaning that interpretations exhibiting more of the behavior were more likely to yield an accurate diagnosis (Table 3.2). Interpretations with a larger magnification percentage were less likely to yield an

Table 3.1: Pathologist's characteristic, clinical experience and ratings of difficulty and confidence on melanocytic skin lesions as predictor variables and Average zoom, maximum zoom, and zoom variance as outcome variables. Contrast specifies the mean difference of the outcome among a predictor variable's categories. * Ordinal variable, summarized in Table 2.1.

Variables	Average zoom		Maximum zoom		Zoom variance	
	Contrast	P-value	Contrast	P-value	Contrast	P-value
Pathologists' demographics						
Gender (Female vs. Male)	0.03	0.878	0.05	0.825	0.22	0.290
Age (50–64 vs. 20–49)	0.29	0.192	0.39	0.068	**0.41**	**0.038**
Clinical Experience Level						
Board certification/ Fellowship training (Yes vs. No)	**-0.62**	**0.003**	**-0.45**	**0.037**	**-0.45**	**0.030**
Experience with melanocytic skin lesions* (Higher order is more years of experience)	0.12	0.286	0.12	0.266	0.16	0.100
Caseload of melanocytic skin lesions* (Higher order is higher caseload)	**-0.35**	**0.015**	**-0.29**	**0.039**	**-0.31**	**0.017**
Ratings on melanocytic skin lesions						
Difficulty level* (Higher order is higher difficulty)	-0.05	0.765	-0.03	0.849	-0.09	0.534
Confidence level* (Higher order is lower confidence)	**0.21**	**0.044**	**0.19**	**0.059**	**0.17**	**0.075**

accurate diagnosis (adjusted OR <1).

To further investigate the associations between viewing behavior and accuracy in the presence of other confounding factors, we modeled our data using a multivariate cross-classified multilevel model. We selected a subset of predictor variables, including one variable for each of the zooming (zoom variance), panning (scanning percentage), interacting with ROI behaviors (ROI time percentage), and interpretation efficiency (total time), based on the relative strength of odds ratios shown in Table 3.2. The results from the multivariate model are shown in Table 3.3. Total interpretation time and ROI time percentage are significantly

Table 3.2: Each row represents one model with a viewing behavior as the predictor variable and diagnostic accuracy as the outcome. Each model was adjusted for pathologists' years of experience in interpreting melanocytic skin lesions and having board certification and/or fellowship training as covariates. OR stands for Odds Ratio.

Predictor Variable	Adjusted OR (95% CI)	P-value
Total interpretationtime	1.33 (1.09, 1.62)	0.005
Average zoom	1.26 (1.03, 1.54)	0.023
Maximum zoom	1.24 (1.03, 1.50)	0.026
Zoom variance	1.37 (1.11, 1.68)	0.003
Magnification percentage	0.76 (0.63, 0.92)	0.006
ROI time percentage	1.35 (1.07, 1.69)	0.011
Scanning percentage	1.21 (1.00, 1.47)	0.054

associated with diagnostic accuracy (P-value < 0.05), whereas zoom variance and scanning percentage are marginally significant ($0.05 <$ P-value < 0.1).

Table 3.3: Each row represents one model with a viewing behavior as the predictor variable and diagnostic accuracy as the outcome. Each model was adjusted for pathologists' years of experience in interpreting melanocytic skin lesions and having board certification and/or fellowship training as covariates. OR stands for Odds Ratio.

Predictor Variable	Adjusted OR (95% CI)	P-value
Total interpretation time	1.25 (1.01, 1.54)	0.0360
Zoom variance	1.22 (0.98, 1.53)	0.0786
ROI time percentage	1.38 (1.10, 1.73)	0.0058
Scanning percentage	1.20 (0.98, 1.47)	0.0716

3.5 Limitations, impact, and conclusion

This study leveraged WSI viewing behavior data to reveal associations between viewing behavior and pathologist characteristics and diagnostic accuracy. When exploring the former association, we showed that average, maximum and variance of zoom level were negatively associated with pathologists' caseload of melanocytic skin lesions, having board certification and/or fellowship training in dermatopathology, and their confidence level in interpreting melanocytic skin lesions. This means pathologists with these characteristics on average used a lower and limited range of zoom levels. In addition, we found a positive association between pathologists' age and maximum and variance of zoom level. When investigating the associations among viewing behaviors and diagnostic accuracy, we showed average, maximum, and variance of zoom levels, total interpretation time, and the proportion of interpretation time spent viewing consensus ROIs have positive associations with diagnostic accuracy. Magnification percentage, which measures consecutive zoom-in behaviors, was seen to have a negative association with diagnostic accuracy. In other words, pathologists who performed many consecutive zoom-ins on various image locations were less likely to reach a correct diagnosis. Scanning percentage, which measures the proportion of time spent panning with a fixed zoom level, has a marginally significant positive association with accuracy.

The association between time spent viewing the consensus ROI and diagnostic accuracy highlights the importance of detecting critical image regions, deeming them worthy of interrogation, and gaining high-power views of histopathological features in these regions. As digital WSI and computer-aided diagnostic (CAD) tools continue to pervade training and clinical practice, we believe this result can be leveraged in future research and development. For example, given the relatively strong association between time spent examining the ROI and diagnostic accuracy, these specific regions can be used to train CAD and artificial intelligence (AI) algorithms on the histopathological features critical to enabling accurate diagnoses. Future adaptive tutoring systems can also monitor trainee viewing behavior and guide novice pathologists towards these features, helping them discover the most critical

image regions for deriving an accurate diagnosis. With FDA approval, using digital pathology is becoming an essential part of daily practice of pathology. As a result, digital whole slide imaging has the potential to alter practically every area of the clinical workflow, teaching and education, and research. We hypothesize that the rich depth of new data becoming available since the advent of WSI will open up future studies that might use this information in teaching and evaluation. In this paper we outlined methods of collecting data remotely on pathologists' viewing behaviors. Given the wide range of pathologists' interpretations and diagnoses of complex melanocytic lesions, studying pathologists' viewing behaviors and interpretive strategies might be beneficial in many areas, and specifically in education.

Despite the fact that using WSIs in digital pathology has many benefits, there are various challenges in obtaining such technology and procedures. Each step of high-quality pathology slide preparation, including embedding, cutting, staining, and scanning, is critical to the successful adaptation of whole-slide images in digital pathology [20]. The methods and operational quality controls must be standardized to reduce system mistakes and random errors, because a single noise in huge data might cause misclassification of the case. Therefore, acquisition of high-quality scanners and staff to manage the complete WSI system is costly. High-capacity servers are needed for storage and distribution purposes. Moreover, numerous technological and ethical issues must be resolved when allowing clinical teams to share and analyze imaging data and patient information across a larger platform. Besides the technical challenges, the small experimental sample sizes in pathology studies may limit generalizability and introduce challenges to the statistical analysis. Although these challenges remain true, computational pathology will continue to improve clinical processes and communication among pathologists with the introduction of digital pathology technologies, statistic algorithms, and the growing medical data and whole slide images.

Chapter 4

ATTENTION-GUIDED ROI DETECTION

4.1 Introduction and Motivation

Detecting regions of interest (ROIs) on a whole slide image (WSI) involves a visual assessment of an image to locate regions with the most relevant and representative pathology. An eye-tracking study highlights the crucial role of fixating on a consensus-defined ROI, as failure to do so can lead to the pathologist overlooking these critical areas [12]. Previous studies show a connection between pathologists' viewing behaviors and diagnostic accuracy [34, 80]. The study presented in this chapter hypothesizes that computer-aided diagnosis (CAD) systems might benefit from incorporating viewing behavior data. Hence, automatic ROI recognition is a reasonable first step to developing an automated diagnosis system. Marzahl et al. [76] shows that automatic annotations on microscopy slides increased consensus among experts and increased accuracy in deep learning classifiers more than manual annotations, ensuring more consistent and repeatable results which is highly desirable in the medical field.

The most common methods for detecting ROIs include supervised machine learning techniques on a low-level image features. Pathologists, on the other hand, find this difficult to understand and often unrelated to the visual characteristics they observe. The semantic gap between pathological observation and low-level representation of visual properties of the image is a significant barrier to the correct translation of expert knowledge into CAD models. Encoding pathologists' qualitative expertise into low-level features is a complicated but promising task [32]. With the introduction of digital pathology and whole slide imaging, pathologists' interpretive behaviors during the medical decision-making process can now be recorded and studied. This enables us with a greater integration of expert knowledge into CAD systems. Typically, a pathologist does not examine the entire slide, but rather

focuses their assessment on a few visual areas or ROIs [39]. As a result, recognizing ROIs in histopathology images may be a potential source of knowledge in a variety of diagnostic tasks. Such ROIs could establish new learning frameworks that would be used in medical education and training, and diagnosis assistance. In addition, an accurate detection of such regions can highly reduce the computational cost and diagnosis time.

Previous ROI detection systems have been developed in different frameworks including object detection [60, 74, 88, 134, 73, 6], tissue segmentation [91, 93, 46, 67, 52], classification [79], CNN-based feature extraction [55, 126, 72], and content-based histopathology image retrieval [143, 144, 90]. These methods mostly rely on pathologists' manual ROI annotations, which are costly, time-intensive, and require domain expertise. However, pathologists' viewing behavior data collected during their routine diagnosis sessions on digital viewers, offers a rich and efficient source of information for ROI detection [117]. While Mercan et al. [79] employed pathologists' viewport-tracking for breast biopsy images and Zou et al. [147] used ophthalmologists' eye-tracking for retinal images to localize diabetic macular edema ROIs, these models are restricted by their reliance on basic image attributes like color and texture. These models face challenges in generalization and performance across varied conditions such as different scanners, color distributions, and image types. Moreover, research in computer vision has demonstrated that deep learning algorithms can outperform algorithms that use hand-crafted features [19].

In this chapter, we propose an innovative method combining information on pathologists' viewing behavior and deep image features to generate heatmaps indicating diagnostically relevant areas on WSI. A heatmap is a visual representation of data where varying colors highlight the significance or frequency of pathologists' attention on specific regions. These heatmaps guide our model, enabling the reconstruction of heatmaps for input images. Our approach integrates pathologists' domain knowledge with deep image features, enabling robust ROI detection. The model's effectiveness is demonstrated by evaluating its performance on WSIs of skin biopsies of melanocytic lesions. The proposed model excels by utilizing pathologists' viewing behaviors, offering the potential to assist pathologists in clinical training

programs, clinical practices, and the development of CAD systems. The key contributions of our study include:

- a novel system that emphasizes viewing behaviors for ROI detection,
- broad applicability to varied pathology types,
- high recall in ROI identification,
- performance improvement of computer-aided diagnosis models by incorporating ROI detection result as supplementary signals.

4.2 Related work

Mercan et al. [79] proposes a bag-of-words (BoW) model [113] to represent diagnostically relevant regions of interest (ROIs) in a digital breast biopsy dataset. The bag-of-words model represents images as collections of visual words, which are 120 x 120 pixel image patches extracted from the original WSI.The bag-of-words model is constructed by clustering the visual words into a visual vocabulary using k-means clustering. Each visual word is represented by a feature vector that combines local binary pattern (LBP) histograms for texture and Lab histograms for color. The sliding window approach is used to extract visual bags (3600 x 3600 pixel image windows) that overlap with ROIs as positive samples and everything else as negative samples for classification using logistic regression. Experiments were conducted to compare different dictionary sizes, visual word definitions, and training data. The proposed method achieved 75% accuracy in detecting ROIs from unseen images.

Elmes et al. [27] proposes a supervised deep learning framework and autoencoder based network to get image representations from the encoder. They use 30 images (segmented into 128 $\times$ 128 pixel patches) of small bowel biopsies and 5-fold cross-validation in their training and validation phase. They cluster their patches into seven clusters representing villi, crypt, stroma, etc. using k-means. They used an autoencoder with 5 stages convolution as encoder

and 5 layers of transposed convolution as their decoder. They assign the feature vector for each patch to one of seven clusters they defines based on the least squared Euclidean distance between the patch's feature vector and the cluster centers. The recorded heatmaps were integrated over the segments in each cluster to give the cluster's weightings. Then, kernel density estimation is used to generate the heatmap based on the cluster weightings. They evaluate their results by a pathologist and also empirically using a deep network for coeliac disease classification. They do not provide details on how they obtain viewing behavior heatmaps and how to integrate the heatmaps into clusters.

4.3 Dataset

M-Path dataset was used to perform this study. The description of M-Path skin biopsy specimens of melanocytic lesions, pathologists' characteristics and viewport data, and consensus reference panel are explained in sections 2.1, 2.2, 2.3, and 2.4 respectively. To be consistent with the latest revision of the MPATH-Dx classification scheme [5], we combined classes 1 and 2 in the original dataset. This leaves us with a more balanced data distribution among 4 different classes. Table 1 summarizes our dataset distribution among the four MPATH-Dx classes. From the M-Path dataset, which contained 240 patients' digital WSIs of their skin biopsies, we narrowed down our selection to 172 cases. This selection was based on the availability of viewport tracking data for a case and the inclusion of interpretations (pathologist, case) with a maximum of 1-class error, as defined in section 2.4.

As a consequence of this criterion, a total of 856 interpretations (an average of 5 pathologists independently interpreted each case) were retained out of the initial 1036 interpretations. We analyze our WSIs at 10x magnification as they provide enough clinical information to allow diagnostic classification by the pathologists for most cases, yet are of reasonable size for processing. To address the challenges posed by the large size and variability of WSIs, various processing techniques can be applied. While one approach involves down-sampling and resizing the WSIs to a fixed size, this can lead to a loss of valuable information. Instead, we employ a cropping strategy, dividing the WSIs into non-overlapping patches of size 256x256

and 512x512. By processing each patch individually, we can retain important details while effectively managing the computational requirements associated with the analysis of WSIs. Our dataset was split and stratified based on the consensus MPATH-Dx class of each case to train (60%), validation (20%), and test (20%) sets. This ensures that each subset contains a representative distribution of the four different MPATH-Dx classes. In Table 4.1 we provide a summary of the size of the train, validation, and test subsets.

Table 4.1: Dataset summary.

MPATH-Dx Class	# of Cases (Train 60%)	# of Cases (Validation 20%)	# of Cases (Test 20%)	Total
1 and 2	26	9	9	44
3	26	9	9	44
4	24	8	8	40
5	26	9	9	44
Total cases	102	35	35	172
Total interpretations	507	180	169	856
Total patches (256 x 256)	96614	15812	23440	135866
Total patches (512 x 512)	26699	4691	6604	37994

4.4 Methods

4.4.1 Extracting Viewing ROIs

The viewport tracking procedure recorded the coordinates of the windows corresponding to the parts of the WSI visible on the screen. A viewport log is a sequence of screen coordinates and zoom levels with timestamps that indicate the pathologists' screen location in the digital WSI. The sequence of viewports from a particular pathologist's interpretation of a particular slide is denoted as $l_t, t = 1, 2, ..., T$ in the notations below. We employed the method proposed by Mercan et al. [79] to extract diagnostically important areas from WSIs based on pathologists' viewing behavior. This method involves three behaviors: zoom peaks, slow pannings, and

fixations. We describe these three behaviors below and more details about their methodology can be found elsewhere [79].

- **Zoom peaks (l_i)** are the log entries where the zoom level is higher than the previous and the next entries. A zoom peak identifies a region where the pathologist intentionally zoomed to look at a higher magnification. During the diagnostic process, low magnification views are also very important in terms of planning the search strategy and seeing the big picture. In low magnification, the pathologists determine the areas of importance to zoom into.

$$zoom(l_{i-1}) < zoom(l_i)$$
$$zoom(l_{i+1}) < zoom(l_i)$$

- **Slow pannings ($l_{i..j}$)** are the log entries where the zoom level is the same as the previous entry, and the displacement (distance between the center of two viewports) is small. We used a 100 pixel displacement threshold on the screen level to define the slow pannings. The quick pans intended for moving the viewport to a distant region result in high displacement values (more than 100 pixels). In comparison, slow pannings are intended for investigating a slightly larger and closer area without completely moving the viewport.

$$zoom(l_k) = zoom(l_{k+1}) \quad \forall k \in i, ..., j-1$$
$$displacement(l_k, l_{k+1}) < 100 \quad \forall k \in i, ..., j-1$$

- **Fixations (l_i)** are the log entries where the duration is longer than 2 seconds. Fixations identify the areas to which a pathologist focused extra attention by looking at them longer. Entries associated with a duration of more than 1 minute were excluded due to the assumption that the pathologist was not actively interpreting during that time.

$$fixation(l_i) > 2sec$$

In histopathologic diagnosis, the field of view holds significance for pathologists, as they can explore digital cases by zooming in and out. Lower magnification viewports encompass a larger area of the WSI. To maintain control over the size of extracted viewports using this methodology and to identify more precise regions within the images, we exclusively consider view-ports with a magnification greater than 5x. In the following sections, we refer to these regions as viewing ROIs. Note that these regions are not necessarily related to the final diagnosis given to a case by the expert and may include distracting regions as well as diagnostic regions. Figure 1 shows how viewing behaviors of different pathologists differ while viewing the same case which results in different viewing ROIs.

4.4.2 Generating Viewing Heatmaps

Each skin biopsy case in our dataset is independently viewed by an average of 5 community pathologists. We generated a single heatmap for each case by merging all the viewing ROIs extracted from pathologists' interpretations as described in section 4.4.1 and shown in Figure 4.2. However, to reduce the distracting areas viewed by pathologists, we only consider interpretations with a maximum of 1-class diagnosis error as defined in section 2.4. We define an accurate diagnosis as a diagnosis in agreement with the consensus reference classification and diagnosis error as a difference between the pathologist's diagnosis and the consensus diagnosis.

We generated a pixel-level heatmap based on the duration that each pixel was viewed. The total viewing time for each pixel across all viewports was accumulated to determine its heatmap value. These heatmaps were then normalized to values between 0 and 1. This means regions with a lower value (less bright regions) are of lower importance and pixels with higher values (brighter regions) are more important in the diagnosis as they have been viewed more during diagnosis. These heatmaps are used as the ground truth in this study.

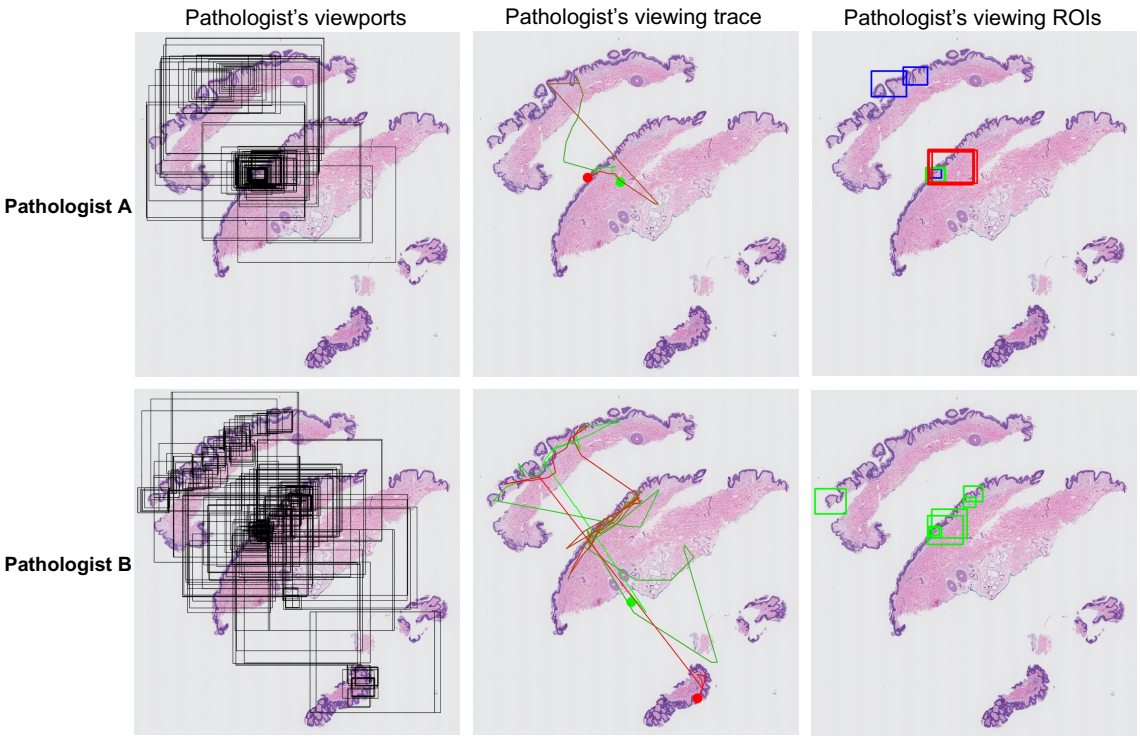


Figure 4.1: Each row visualizes a different pathologist's viewing patterns and behaviors. Left: All viewports are shown in rectangular regions with black borders. Middle: Traces of the viewports by connecting the center of rectangles shown on the left, starting the viewing process from the green circle, and ending viewing of the case with the red circle. Right: Viewing ROIs extracted from all viewports on the left using zoom peaks (blue), slow pannings (red), and fixations (green).

4.4.3 Method and Experiment Setup

Autoencoders, as initially conceptualized [105], are designed to reconstruct their input. They are mainly composed of an encoder network that maps input data into a low-dimensional latent space and a decoder network that reconstructs the input from this latent space representation. The objective is to ensure the reconstructed version closely resembles the original. Encoder-decoder models are optimized by minimizing the disparity between the

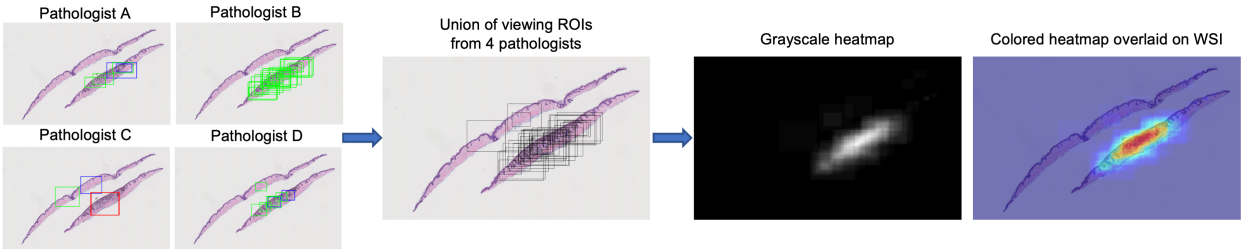


Figure 4.2: Left: Extracted viewports from four different pathologists independently viewing the same case. Middle: Merger of all the viewports shown on the left column. Right: Generated grayscale heatmap of the middle column viewports based on the viewports' duration and the colored version overlaid on top of the WSI, highlighting the important regions.

input and output images, typically by using mean squared error (MSE) as a loss function. This equips them with the proficiency to reconstruct images from condensed representations with high fidelity.

Deep learning has shown considerable potential in medical image analysis applications in recent years [123, 18, 35]. However, translating research breakthroughs into clinical tools remains a challenging process [114]. One of the primary barriers is the scarcity of high-quality labeled data required for developing accurate models [122]. Transfer learning [135] offers a solution by leveraging a model pre-trained on a different task, like ImageNet [21], as a foundation for a novel task. In the context of encoder-decoder architectures, transfer learning can be used to fine-tune a pre-trained model as the encoder to extract features for a new task.

In this study, we used three model architectures to reconstruct input images as illustrated in Figure 3: a convolutional autoencoder (ConvAE), a U-Net, and an Attention U-Net.

- ConvAE: We initialized the encoder with the ResNet-18 [43] model pre-trained on ImageNet ImageNet [21]. Our decoder consisted of 5 deconvolution layers with ReLU activation, except for the final layer, which used Sigmoid activation.

- U-Net and Attention U-Net: We used the implementation of U-Net [102] by Yakubovskiy [132]. Both models were initialized with ResNet-34 [64] pre-trained on the ImageNet dataset as the encoder and a standard U-Net decoder. Figure 4.3 demon-strates the pipeline of our system. The Attention U-Net incorporated spatial Squeeze and channel Excitation (scSE) attention modules [104].

For our experiments, we used the Adam optimizer with a learning rate of 0.001. For the 256x256 patch size experiments, we used 2 GPUs with a 64 batch size. For the 512x512 patch size experiments, we used 4 GPUs with a 32 batch size. Models were trained on the training set and validated using the validation set to stop training when the model's performance started to degrade and avoid overfitting. All experiments were done on NVIDIA GeForce GTX 1080 GPUs with 8GB memory each.

For image pre-processing, we used the ImageNet standard normalization, setting the mean to (0.485, 0.456, 0.406) and the standard deviation to (0.229, 0.224, 0.225). Additionally, we employed a diverse set of image augmentations, including horizontal and vertical flips, random cropping, sharpening, embossing, brightness adjustments, hue and saturation modifications, grayscale conversion, and contrast adjustments. These augmentations were applied in a randomized sequence to enhance the robustness and variability of our dataset.

In addition to our approach, we also re-implemented the method by Mercan et al. [79]. Originally designed for ROI detection in breast biopsy images, we adapted, trained, and tested this model using our M-Path dataset. The method follows a bag-of-words approach [113] for feature construction. By using a sliding window, the WSI is divided into 1024x1024 bags, overlapping by 512 pixels in both dimensions. Each bag is further divided into 128x128 non-overlapping words (8x8 words per bag). Using the K-means clustering algorithm, words are grouped into 40 clusters based on their color (Lab) and texture (LBP) features extracted earlier. For each bag, a frequency histogram is calculated, representing the distribution of the 8x8 patches across the 40 clusters. Next, viewing ROIs are extracted as described in section 2.2.1, and bags are labeled as either positive (ROI) or negative (non-ROI) based on

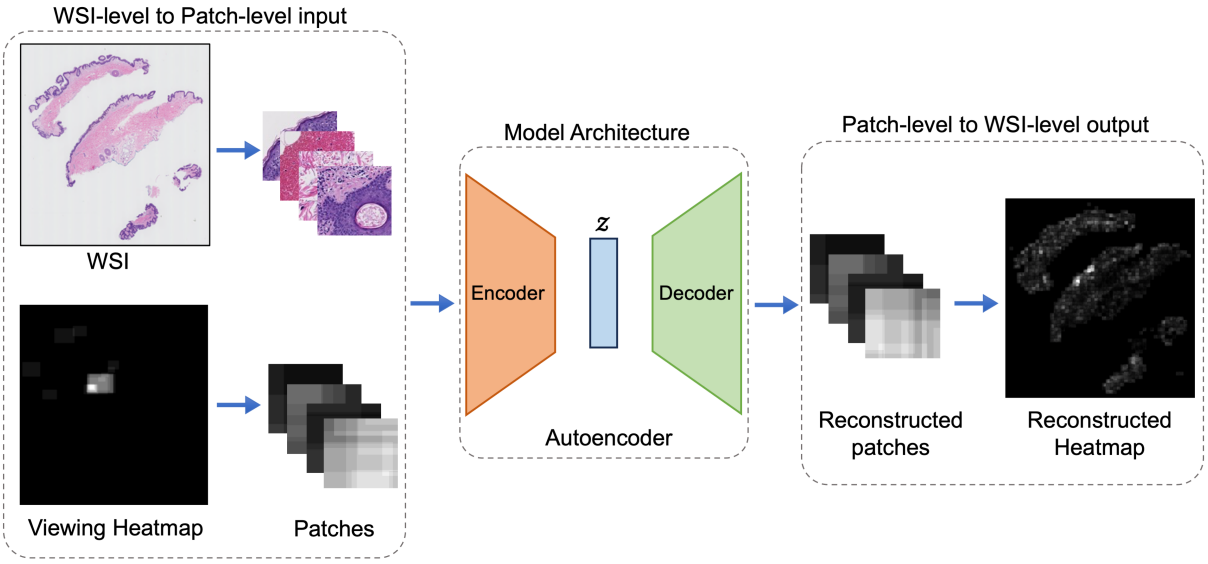


Figure 4.3: Pipeline of the ROI detection model. The encoder transforms input patches into a latent representation z, while the decoder then reconstructs these inputs from the latent space back into the original pixel space. See section 4.4.3 for details of the encoder and decoder architectures of the model.

their intersection with the extracted viewing ROIs. Finally, we employed a Random Forest classifier to distinguish between ROI and non-ROI. More details of this method can be found in [79].

4.4.4 Evaluation

Quantitative Assessment. To evaluate our results at an individual patient skin biopsy WSI level, we stitched patches generated by our model together to generate the WSI-level heatmap. We used Mean Squared Error (MSE) and Structural Similarity Index (SSIM) to measure the similarity between the reconstructed heatmaps and the ground truth. Additionally, we employed standard pixel-level segmentation metrics, including Intersection over Union (IoU), precision, recall, and F1-score to assess each model's performance. Collectively, these metrics

offer a comprehensive assessment of the model's capability.

- MSE: Measures the average squared differences between the predicted and actual values, commonly used to assess an autoencoder's performance. In our context, the predicted value corresponds to the model-generated heatmap, while the actual value is the ground truth from pathologists' viewing behavior. MSE is defined below in equation 4.1 where m and n are the dimensions of the image, $y_{i,j}$ and $\hat{y}_{i,j}$ are (i, j) pixel values at input and output images respectively.

$$MSE = \frac{\sum_{i=0}^{m} \sum_{j=0}^{n} (y_{i,j} - \hat{y}_{i,j})^2}{m * n} \tag{4.1}$$

- SSIM: Measures the similarity between two images by comparing their structural information, including luminance, contrast, and structure. It provides a score ranging from 0 to 1, with 1 denoting identical images. In our study, we calculated the SSIM score between the model's reconstructed heatmap and the ground truth heatmap. The SSIM score was used as an objective measure of the similarity between the two images, with a higher score indicating a better match. The formula for calculating SSIM is provided in equation 4.2 where l, c, and s represent the luminance, contrast, and structure components. The parameters α, β, and γ are used to weight each component, with typical values of 0.01, 0.03, and 0.03, respectively. y and $\hat{y}$ are the input and output images respectively.

$$SSIM = l(y, \hat{y})^{\alpha} * c(y, \hat{y})^{\beta} * s(y, \hat{y})^{\gamma} \tag{4.2}$$

- IoU, Precision, Recall, and F1-score: Measure the overlap between the generated heatmap and the ground truth, revealing how much of the ground truth is identified by the model. First, we apply a binary thresholding for each heatmap with a threshold of 0.5, categorizing pixels with values above this threshold as '1' (ROI) and below

as '0' (non-ROI). We conducted experiments with several threshold values—0.4, 0.45, 0.5, 0.6, and 0.7—and found 0.5 to be the best threshold for this task. Based on this binary thresholding, the definitions of True Positive (TP), False Positive (FP), and False Negative (FN) are given below:

- TP refers to the number of pixels correctly predicted as ROI,
- FP denotes the pixels incorrectly predicted as ROI,
- FN represents the ROI pixels that were missed by the model.

$$\begin{aligned} Precision &= TP/(TP+FP) \\ Recall &= TP/(TP+FN) \\ F1 &= 2 * Precision * Recall/(Precision + Recall) \\ IoU &= TP/(TP+FP+FN) \end{aligned} \tag{4.3}$$

Clinical Evaluation by Dermatopathologists. We asked three dermatopathologists to review the model-generated heatmaps on the test set containing 35 WSIs and grade the model's performance using discrete scoring. It's crucial to note that these dermatopathologists are different from the community pathologists whose viewing behavior was used to train our model. Their task was to evaluate the segmentation of the whole slide images. Each dermatopathologist received an individual Google Forms survey. Each of the 35 WSIs was presented at 10x magnification alongside the grayscale model-generated heatmaps. An overlay of the heatmap on the corresponding WSI was also available for better clarity. The dermatopathologists addressed two questions aimed at discerning whether the model was over-detecting or under-detecting essential regions:

- Q1: Does the heatmap closely correlate with your viewing behavior? Rate yes, somewhat, or no.

- Q2: Does the most intense region of the heatmap include the region most representative of your diagnostic impression? Rate yes or no.

It's essential to underscore that human analysis, particularly within medical evaluations, embodies a degree of inherent subjectivity. Recognizing this, our dermatopathologists convened in a collaborative session before their individual case analyses to develop standardized definitions to follow for each of the two clinical questions. This meeting enabled them to arrive at a mutual understanding of the interpretation of the cases. This consensus-building initiative was strategically implemented to instill a level of uniformity in the evaluation process, aiming to reduce individual biases. We analyzed the feedback from all three surveys, considering each one individually and collectively. We categorized the responses for Q1 and Q2 into distinct labels. Specifically, for Q1, the responses were categorized as "No," "Somewhat," and "Yes." For Q2, the responses were categorized as "No" and "Yes."

Computer-aided Diagnosis. The proposed ROI detection framework generates heatmaps that can be used as supplementary signals to train Diagnostic model. We utilize the architecture presented in [86] for this purpose. In this architecture, multiple masks can be appended as additional channels to the input image. Using a MobileNetV2 backbone [106], we extract features from the images at three scales of 7.5x, 10x, and 12.5x. These feature vectors are subsequently fed into ScATNet [129] which aggregates information of the three scales to perform the diagnostic task using Transformer blocks. Specific details regarding the model architecture can be found in [86, 129]. We trained our models for 200 epochs on a single NVIDIA RTX A4000 GPU with 16 GB GPU memory. All the training details and hyperparameters are the same as those in [129].

We train two models for comparison: one using only WSIs and the other incorporating the heatmaps generated by our ROI detection model as a fourth channel added to the WSIs. We evaluate the models using F1-score (equation 4.3), as well as sensitivity (recall) and specificity as shown in equation 4.4. Given that this is a multi-class classification problem, TP, FP, FN , and TN are calculated by summing across all classes.

$$\begin{aligned} Sensitivity(Recall) &= TP/(TP+FN) \\ Specificity &= TN/(TN+FP)) \end{aligned} \tag{4.4}$$

4.5 Results

Quantitative Evaluation. In this section, we provide the results of our experiments. We present the results of our experiments and their improvement over the method by Mercan et al. [79] in Table 4.2. Experiments v1-v3 and experiments v4-v6 use patch sizes of 256 and 512 respectively. In order to validate the consistency of our model's performance, we conducted multiple runs with three distinct random seeds and reported the average values for each metric. Our best model outperforms Mercan et al. [79] by 20% in precision, 11% in recall, 22% in F1-score, and 12% in Intersection over Union (IoU). Figure 4.4 shows heatmaps generated by our model on an unseen test set, alongside their ground truth viewing heatmaps. Additionally, we conducted experiments to investigate the effects of patch size and types of pathologists' viewing behavior on the model's performance. The results of these experiments are discussed in the subsequent sections.

Table 4.2: Results of Experiments evaluated using the M-Path dataset (see chapter 2.)

Model Architecture	Patch size	Avg. MSE	Avg. SSIM	Precision	Recall	F1 score	IoU
v1: ConvAE	256	0.0146	0.876	**0.28**	0.49	0.36	0.18
v2: U-Net	256	0.0149	0.709	0.27	**0.53**	**0.36**	**0.20**
v3: Attention U-Net	256	0.0147	0.712	0.26	0.45	0.33	0.18
v4: ConvAE	512	0.0147	0.692	0.20	0.48	0.28	0.15
v5: U-Net	512	0.0155	0.682	0.19	0.44	0.27	0.14
v6: Attention U-Net	512	0.0157	0.677	0.18	0.53	0.27	0.15
Mercan et al. [79]	-	-	-	0.08	0.42	0.14	0.08

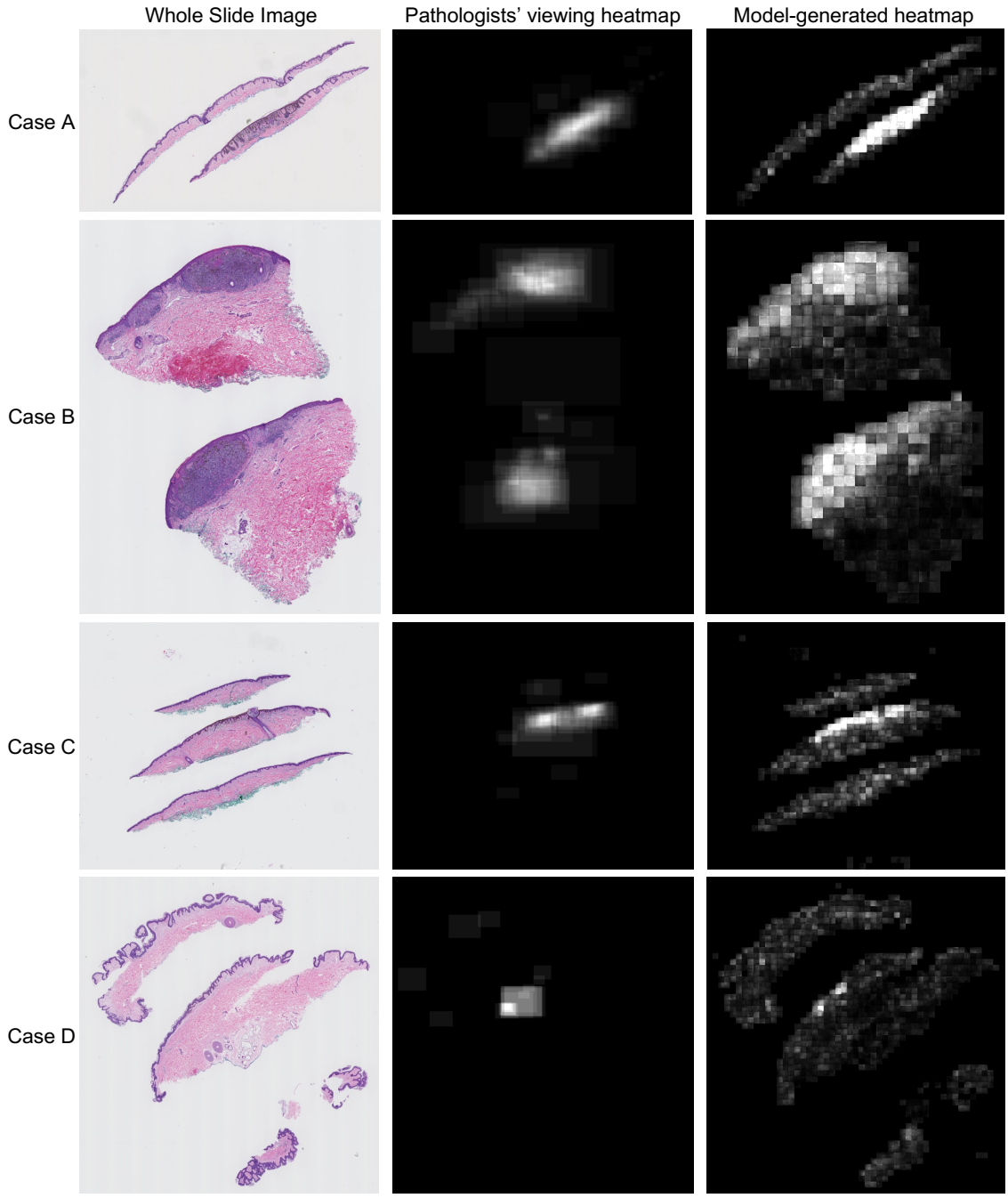


Figure 4.4: Visualized result for 4 example WSIs. Left: WSIs. Middle: Ground truth heatmaps from pathologists' viewing ROIs (see section 4.4.1). Right: Model-generated heatmap on unseen data.

Patch size. To investigate the affect of patch size on the performance of our model, we setup our experiments with two different patch sizes: 256 x 256 and 512 x 512. A summary of the number of training and testing samples is provided in Table 4.1. By reducing the size of patches, the model loses insight on the location of these patches and their neighbor patches. On the other hand, increasing the size of patches would require more computing resources and higher training time. The results of this experiments show that a smaller patch size would result in a more fine-grained heatmap which is more similar to the original heatmaps. Figure 4.5 shows the results from 256x256 patches and 512x512 patches compared to the ground truth heatmap. Models trained with a smaller patch size have higher Precision, Recall, and F1 score.

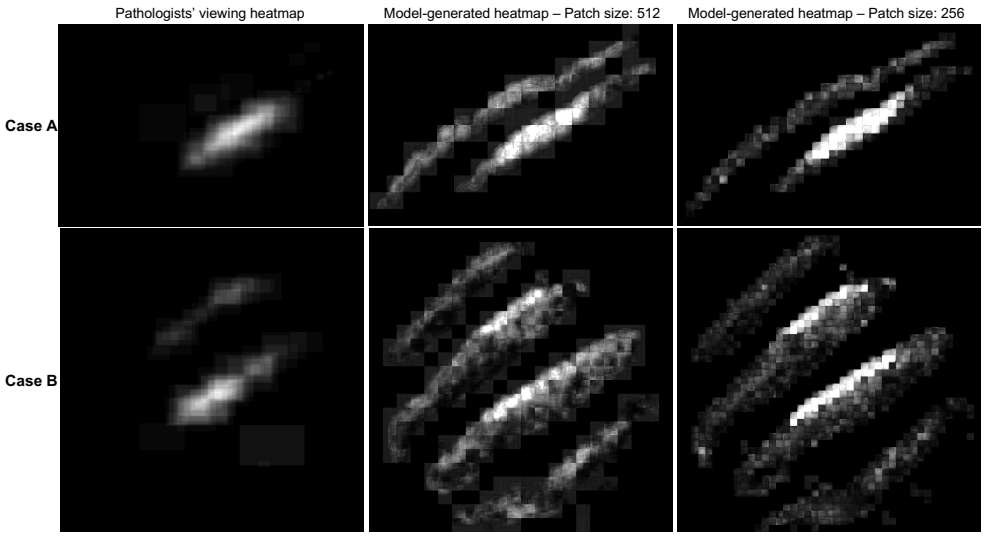


Figure 4.5: Left: Heatmap generated using pathologists' viewing ROIs (see section 2.2.2). Middle and Right: Heatmaps generated by the model on unseen data with 512x512 and 256x256 patch sizes respectively.

Expert vs community pathologists viewport data. We investigated how viewing behavior from two groups of pathologists, community and M-Path consensus reference panel pathologists, would impact the performance of the model in detecting more precise ROIs.

Hence, we used viewing behavior heatmaps generated from viewports of these two groups as input for training our model. Figure 4.6 shows a comparison of the consensus reference panel and community pathologists' viewing behavior heatmaps and the corresponding results generated using these heatmaps during training. Heatmaps of the consensus reference panel are less cluttered and focused on a few smaller regions whereas community pathologists perform a more comprehensive scan of the slide.

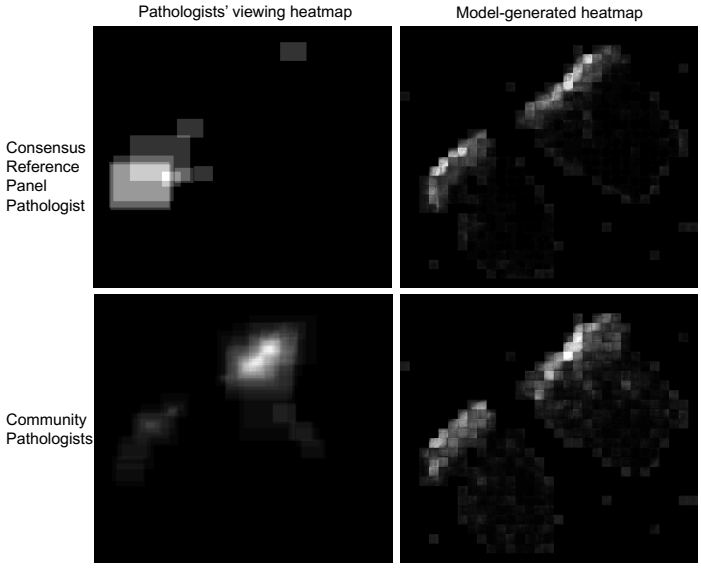


Figure 4.6: Top: The consensus reference panel pathologist ground truth heatmap and its model-generated heatmap. Bottom: Community pathologists ground truth heatmap and its model-generated heatmap.

Clinical Evaluation. WSIs often contain multiple important regions. However, the ground truth heatmap, generated from pathologists' viewing behavior (see section 6.3.1), might not encompass all of these important regions. We observed that our model identified certain areas with characteristics akin to these critical regions, leading to a high false positives rate. Consequently, the conventional pixel-level segmentation metrics do not entirely reflect the model's efficacy. To address this, we performed a clinical evaluation, involving three

dermatopathologists. This evaluation comprised two questions, measuring the resemblances between the pathologists' assessment of the critical regions of the WSI and the model-generated heatmaps. To provide a clear representation of the feedback, we used spineplots to display the proportion of responses within each category for each pathologist, as well as the average proportion across all pathologists. Figure 4.7 visualizes the distribution of responses, providing insights into the consensus among pathologists and highlighting any variations in their evaluations. The outcomes from this assessment demonstrate the capability of our model to generate a heatmap that replicates the regions that a pathologist would view and also to highlight the regions most representative of the final diagnosis.

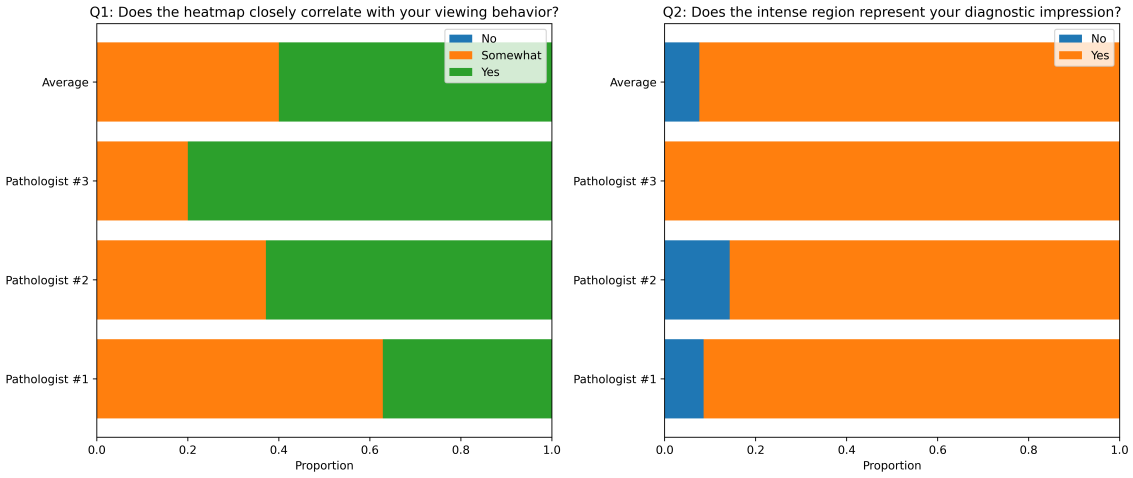


Figure 4.7: Proportion of responses from individual pathologists and the average of all three pathologists for (a) Q1: Does the heatmap closely correlate with your viewing behavior? and (b) Q2: Does the most intense region of the heatmap include the region most representative of your diagnostic impression?

Computer-aided diagnosis. In Table 4.3, we present the results of our diagnostic experiments. Each model was trained using 5 different random seeds and we are reporting the average scores of each experiment. The results indicate an improvement in the diagnostic performance of the model when the heatmap is included as an additional channel in the input.

Additionally, saliency analysis using gradients helps identify relevant areas in an input image that contributed to the prediction. We compare the heatmaps generated by our model with the saliency maps of the ScATNet [129] model trained only on WSIs. Figure 4.8 shows that our model's heatmaps are more aligned with pathologists' viewing heatmaps.

Table 4.3: Results of WSI diagnosis. All numbers are average scores over 5 random seeds per experiments.

Model Input	Micro F1-score	Specificity	Sensitivity
WSI	0.59	0.86	0.59
WSI + Heatmap	**0.63**	**0.88**	**0.63**

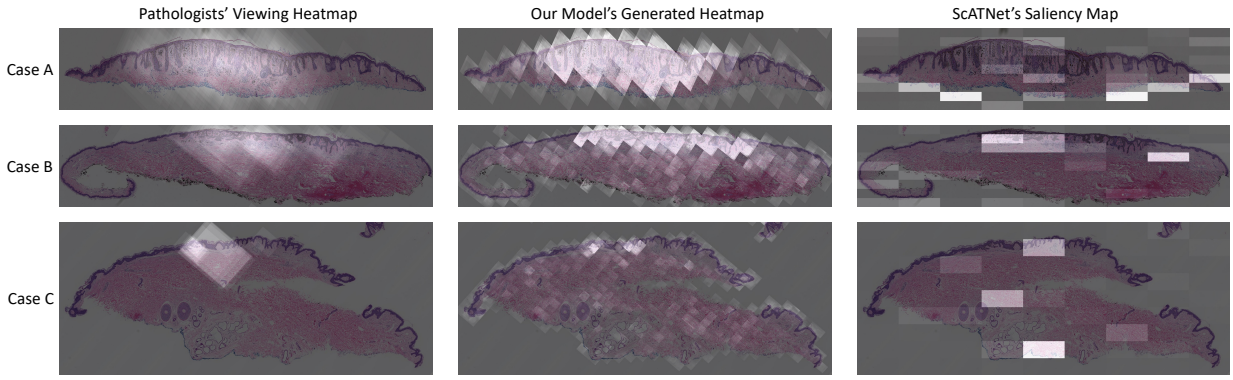


Figure 4.8: Comparison of the heatmaps generated by our ROI prediction model (middle) and the saliency maps of ScATNet [129] trained for diagnosis using WSIs (right). Ground truth heatmaps, based on pathologists' viewing behavior, are shown on the left.

4.6 Limitations, impact, and conclusion

Whole slide imaging has provided the opportunity to study the diagnostic viewing process of pathologists, yielding valuable insights that can be utilized to develop innovative training and

evaluation programs as well as possibly using the data to improve computer-aided diagnosis systems. We have introduced an ROI detection system as the first step of the diagnosis process, aimed at assisting pathologists in quickly identifying relevant regions. The ROIs, identified using pathologists' viewing behaviors such as zoom peaks, slow pannings, and fixations, were utilized to generate a grayscale heatmap which guides our model to focus on crucial image regions. We employed three deep learning architectures for reconstructing the heatmaps. These regions may not necessarily represent the definitive ROIs of the digital slide but replicate a pathologist's viewing patterns that can include distracting or misleading regions, providing a more realistic depiction of the diagnostic process.

Our model outperformed the Mercan et al. method [79], with an emphasis on high recall, capturing all relevant regions to reduce the chance of missing crucial information, despite potentially including some false positives. The use of viewport-extracted ROIs and square-shaped patches allowed our model to align closely with the ground truth in terms of shape and structure. In additional experiments, we analyzed the impact of patch size and type of pathologists' viewing behavior on our model's performance. Larger patch size had little effect on performance but required more computing resources. Models trained using the consensus reference panel pathologists' viewing heatmaps produced fewer false positive samples since these heatmaps highlight smaller image regions as these pathologists did not require a lot of scanning to find the ROIs. Consequently, the final output of the model generated from the viewing data of the consensus reference panel pathologists consisted of smaller and fewer ROIs. The intrinsic complexity of ROI detection can lead the model to detect regions as ROIs that are not present in the ground truth set. However, this does not imply that these regions are insignificant. These regions can be ignored if found irrelevant by pathologists. The findings from our clinical evaluation demonstrate the effective performance of our model, despite its low precision. Moreover, the tracking software records visible regions in a rectangular shape, introducing unimportant surrounding regions and white space background, especially at lower zoom levels. Despite our efforts to minimize non-tissue patches during WSI pre-processing, the complete exclusion of unwanted regions was not achievable. Furthermore, the absence

of eye-tracking data restricts our ability to accurately determine the specific focus points of pathologists within these full viewports. Despite these limitations and challenges, our model demonstrated efficiency by simplifying and accelerating ROI annotation, thereby reducing costs.

We integrated the results of the ROI detection model into a computer-aided diagnosis system as supplementary signals and demonstrated that the performance of the diagnosis model improved with this addition. Moreover, we visualized the saliency maps of the diagnosis model trained solely on WSIs (without the heatmaps). Upon comparison, our model's generated heatmaps showed greater alignment with pathologists' viewing heatmaps than the saliency maps of the diagnosis model.

In the field of ROI detection in histopathological images, our approach distinguishes itself by integrating pathologists' viewing behavior data from their clinical review and interpretation of each case into the model's training; this viewing behavior data is quite distinctive from the many methods that predominantly rely on manually labeled ROIs. While numerous studies have focused on an object detection approach, our analysis suggests that this might not be the optimal paradigm for such a nuanced task. ROIs in histopathological images differ from standard objects found in natural images, challenging exact bounding box comparisons. Instead, our model uses behavior-driven heatmaps to effectively highlight diagnostically relevant regions. This unique methodology, grounded in real-world clinical insights, positions our approach a notch above most state-of-the-art techniques, which often overlook the importance of replicating the intricate clinical viewing behavior of pathologists. Moreover, the lack of available public datasets that capture viewing behavior in histopathological images is a recognized challenge. This restricts external validation of our methodology on diverse datasets and poses a barrier to direct comparisons with other existing techniques.

As the future direction, the addition of precise eye-tracking data would help determine the exact focus of pathologists within the full rectangular viewports, potentially refining the output of the model. The proposed ROI detection model can be used for developing automated diagnosis systems by locating crucial regions rather than processing the entire

slide. Additionally, it would be beneficial to explore the optimal integration of these models into practical, clinical settings and understand how this technology can be more tailored to individual needs for pathologists at varying experience levels. This is because integrating CAD models into healthcare practice requires strict regulatory standards, exhaustive validation, and certification to ensure patient safety and compliance with medical protocols. Moreover, scalability is a pivotal concern, as models proven in controlled experimental settings must be adeptly tailored to accommodate the heterogeneity of data encountered in practical clinical environments. This type of algorithm to identify important image ROIs could be quite helpful as a resource for training and educating the next generation of pathologists.

Chapter 5

VISION-LANGUAGE MODELS IN HISTOPATHOLOGY

5.1 Introduction and Motivation

For the microscopic study of tissue in histopathology, a thorough analysis of whole slide images (WSIs) is crucial. The analysis of these information-rich images requires more than just a patch-level examination for an accurate diagnosis. Previous approaches frequently summarize WSI data into a single diagnostic label, failing to account for the complexity of histopathological information [112]. The necessity for more expressive, interconnected representations that can capture the complex patterns found in histopathological patches is shown by this oversimplification. Natural language descriptions offer a promising solution by providing additional signals beyond the scope of single labels by connecting the features within these patch structures [30, 50].

Although vision-language models have great promise for improving histopathology analysis, the lack of large-scale, comprehensive datasets poses a major obstacle for the field. Existing open-source contributions like OpenPath [50] and ARCH [30], while notable, are constrained in size and scope. Approaches such as PMC-15M [137], which curate large volumes of biomedical image-text pairs, are not publicly available and their specificity to histopathology is yet unclear. This gap emphasizes that a large-scale vision-language dataset specific to histology is needed.

To address this need, we present QUILT, a dataset including $437,878$ images aligned with $802,144$ text pairs at different microscopic magnification scales ranging from 10x to 40x. This dataset was curated from publicly available educational YouTube videos on histology which contain pathologists' expertise. QUILT is extracted from $1,087$ hours of educational content using a variety of models, including large language models, handcrafted algorithms, human

knowledge databases, and automatic speech recognition. This procedure ensures that the dataset provides distinct contributions without duplicating information from other sources. Additionally, we combine QUILT with additional publicly available image-text histopathology datasets from sources like as X (formerly Twitter), PubMed research articles, and the Internet to build QUILT-1M, the largest image-text dataset of its kind to date.

Furthermore, accurate diagnosis depends on an understanding of the holistic nature of WSIs. Using critical reasoning and navigation of these WSIs, pathologists make diagnoses based on morphological concepts and spatial relationships within the tissue [12]. The diagnostic value of current histopathology models is limited since they are unable to reason beyond isolated patches, which is how they are generally analyzed. The success of multi-modal models like the Large Language and Vision Assistant (LLaVA) [65] in natural image domains suggests the potential for comparable approaches in the histopathology domain. Nevertheless, existing multi-modal models for histopathology [65, 81], which rely on image-caption pairs extracted from PubMed articles, fall short due to the absence of visually grounded captions and more comprehensive contextual information from WSIs.

To fill this void, we present QUILT-INSTRUCT, an instruction-tuning dataset of 107,131 histopathology-specific question/answer pairs. This dataset, like QUILT, is extracted from educational YouTube videos but goes further by grounding histopathology concepts using spatio-temporal clustering of narrators' mouse cursors and proposing novel instruction-tuning QA prompting techniques. These techniques enable complex reasoning and iterative abductive reasoning, incorporating global WSI diagnosis and supporting facts with image captions to ground factual information and prevent hallucinations. Using QUILT-INSTRUCT and QUILT, we train QUILT-LLAVA, a multi-modal model for histopathology capable of analyzing images in detail, localizing medical concepts spatially, and reasoning beyond individual patches.

This chapter presents research conducted as a collaborative effort with Wisdom Ikezogwo and Mehmet Saygin Seyfioglu, my colleagues at University of Washington. The contributions presented in this chapter aim to facilitate the development of effective histopathology models and advancing the field through improved educational and diagnostic tools for pathologists.

The key contributions of our study include:

- Introduction of QUILT and QUILT-1M, the largest public vision-language histopathology datasets, providing a rich, diverse resource for developing advanced histopathology models.

- Development of QUILT-INSTRUCT, an instruction-tuning dataset that enhances multi-modal reasoning capabilities in histopathology, grounded in educational video content.

- Training and evaluation of QUILT-LLAVA, a multi-modal model demonstrating significant advancements in detailed image analysis, spatial localization of medical concepts, and reasoning, thus bridging a critical gap in the field of histopathology.

5.2 Related work

Medical vision-language datasets. Learning vision-language representations necessitates extensive datasets of images paired with descriptive text, a resource that is particularly scarce in histopathology. For instance, the MIMIC-CXR-JPG v2.0.0 dataset [56] comprises de-identified hospital-sourced chest radiographs and reports. In histopathology, The Cancer Genome Atlas[1] offers de-identified PDF reports for a limited number of WSIs. However, the enormous size of this data (up to $120,000^2$ pixels) poses significant processing challenges, limiting its application to a few focused studies [75].

Most medical vision-language datasets are concentrated in the radiology sub-domain, as collecting validated multimodal data in this field is relatively straightforward [56]. Many models are trained on subsets of PubMed [101] or comparable radiology datasets [139, 49, 29, 92]. PMC-15M [137], a recent subset of PubMed not specific to histopathology, was used to train multiple models. While these models are public, PMC-15M itself is not, making it difficult to determine the histopathology-relevance of its content. One of the earliest histopathology vision-language datasets, ARCH, contains only $7,614$ accessible image-text

[1] https://www.cancer.gov/tcga

pairs [30, 44]. Later, OpenPath [50], a dataset of 200K image-text pairs extracted from Twitter, became the largest histopathology dataset until QUILT-1M.

Visual instruction-tuning. The curation of instruction-tuning datasets from image captions using the open-source LLMs [120, 17, 54, 116] has accelerated advances in training multi-modal models. Significant capabilities have been demonstrated by prior work [70, 146, 31], where LLaVA-1.5 matched GPT-4's performance in some multi-modal tasks [69]. In the medical domain, visual Med-Alpaca [41] created 54K question-answer pairs for instruction-tuning using GPT-3.5. Although PMC-VQA [138] curated a larger multiple-choice answer dataset from general medical fields using PubMed, its coverage of histopathology remains restricted. For instance, LLaVA-Med [65] uses image captions from PubMed articles, adding sentences from the article to short captions that might not always relate to the figure that is being referred. A subset of 17K images in LLaVA-Med are relevant to histology, resulting in 49K question-answer pairs. These methods rely on isolated image-text pairs which restricts GPT-4's ability to think beyond its local context. The possibility of hallucinations increases with GPT-4's attempts to extrapolate context, indicating a major obstacle in creating useful medical visual instruction-tuning datasets.

5.3 Dataset

5.3.1 QUILT: Collecting medical image and text pairs from YouTube

Creating a vision-language dataset from videos presents several challenges including inappropriate content and limitations of Automatic Speech Recognition (ASR). Many videos are not in English, lack audio, have irrelevant medical information, or include static, poorly detailed histopathology images. Additionally, conventional ASR systems struggle with specialized histopathology terminologies, necessitating sophisticated solutions. Furthermore, the de-noising of text and images from conversational videos, in which instructors regularly pan and zoom, adds to the complexity of this task. We designed an extensive pipeline to curate the dataset in order to overcome these difficulties. The pipeline includes the following phases,

which are also shown in Figure 5.1:

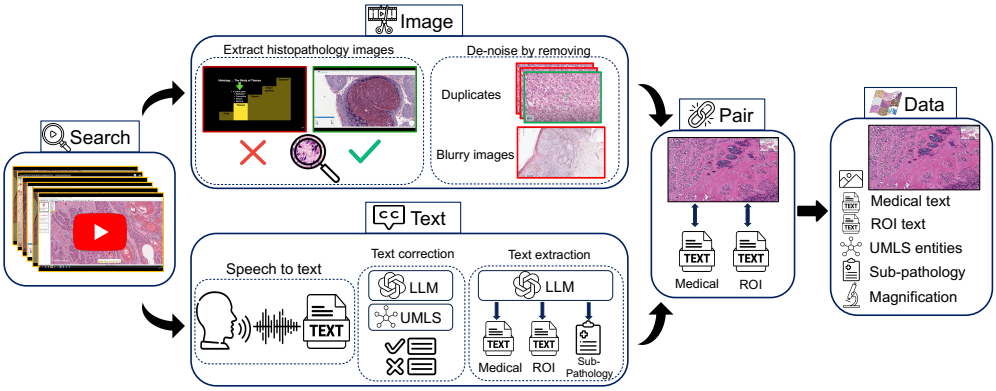


Figure 5.1: **Overview of Quilt curation pipeline.** We identify relevant histopathology YouTube videos in **Search**. For **Image** extraction, we find and de-noise histopathology frames using trained models. In **Text** section, we rely on a conventional Automatic Speech Recognition (ASR) model and leverage Unified Medical Language System (UMLS) and large language models (LLMs) for post-processing and ASR error correction. Relevant sub-pathology, medical and region-of-interest (ROI) text are extracted using an LLM. Finally, domain-specific algorithms are used to **Pair** images and text, eliminating duplicates to yield Quilt, a richly annotated image-text dataset for histopathology.

(1) **Channel and Video Data Collection:** We identify relevant YouTube channels and videos using keywords from 18 sub-pathology fields, avoiding channels with over 300K subscribers to exclude broad science content. We download low-resolution versions of all identified videos.

(2) **Filtering Narrative-Style Medical Videos:** We exclude videos that are shorter than one minute, non-voiced, or have non-English audio. For the remaining videos, we assess:

   (a) Medical Content: Using FFmpeg, we extract keyframes and employ an ensemble

of histopathology image classifiers to identify videos with histopathology images.

(b) Narrative Style: We analyze selected keyframes and their subsequent frames for cosine similarity, identifying videos where presenters spend significant time describing WSIs in detail. We then download these videos at high resolution.

(3) **Text Extraction and De-noising:** We use Whisper [98], a conventional ASR model, to transcribe the videos. Given its limitations with medical terms, we implement a four-step text de-noising pipeline. This process ensures the extraction of relevant medical text and region-of-interest (ROI) text from the ASR output:

(a) Extract keywords using the Rake [103] algorithm.

(b) Cross-check keywords against UMLS [7] using SciSpacy [84] and correct misspelled words.

(c) Use LLMs to correct and identify additional errors within the context.

(d) Validate corrections against UMLS and a curated list of histopathology terms.

(4) **Image Frame Extraction and De-noising:** We extract representative image frames from the videos by analyzing histopathology keyframes. This involves breaking videos into time-intervals (chunks) and extracting the median image of stable frames within each chunk.

(5) **Aligning Modalities:** For each narrative-style video, we align text and image modalities by:

(a) Mapping ASR-extracted text to corresponding video chunks.

(b) Extracting and aligning medical and ROI captions to the representative images within these chunks.

This alignment process ensures that each image is paired with accurate and relevant textual descriptions. Figure 5.2 presents a few samples from our dataset with corresponding metadata.

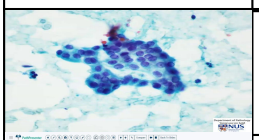
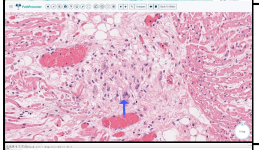
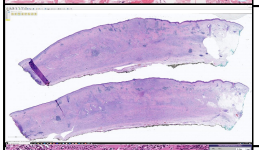
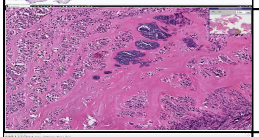
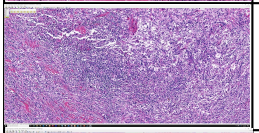
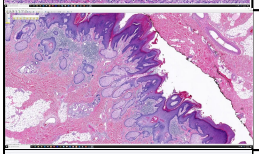
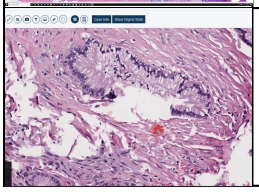
Image	Medical TEXT	ROI Text	Sub-pathology Classification
	['There are clusters of cells with micro-follicular formations.','Nuclear pseudo-inclusions, oval nuclei, nuclear grooves, and small nucleoli are present in some cells.']	['clusters of cells', 'micro-follicular formations', 'nuclear pseudo-inclusions', 'oval nuclei', 'nuclear grooves', 'small nucleoli']	['Endocrine', 'Cytopathology', 'Head and Neck']
	['Cluster of macrophages and T cells is characteristic of acute rheumatic fever.', 'Aschoff body is a characteristic feature of acute rheumatic fever.', 'Macrophages with elongated chromatin are called Anitchkow cells and are commonly seen in Aschoff bodies.', 'Pancarditis with Aschoff bodies is present.']	['Cluster of macrophages and T cells', 'Aschoff body', 'Macrophages with elongated chromatin', 'Anitchkow cells', 'Pancarditis']	['Cardiac', 'Hematopathology', 'Endocrine']
	['An 80-year-old man has a scar-like plaque on the scalp that has been called malignant on a biopsy.', 'The tissue affected by the plaque extends from the epidermis to the galea aponeurotica, near the periosteum of the skull.', 'The skin, dermis, and subcutis are all affected by the process.']	['scar-like plaque on the scalp', 'malignant on a biopsy', 'skin, dermis, and subcutis affected by the process']	['Dermatopathology', 'Soft tissue', 'Hematopathology']
	['Inflammatory cells surrounding cartilage can indicate acute chondritis, with neutrophils being the principal cell type.', 'Chronic chondritis may be diagnosed if lymphocytes are the predominant inflammatory cell type.']	['cartilage', 'inflammatory cells']	['Hematopathology', 'Bone', 'Dermatopathology']
	['Large histiocytes with abundant cytoplasm identified as Rosai-Dorfman histiocytes.', 'S100 stain showed perivascular cuffing.', 'Initial diagnosis of inflammatory pseudotumor of the orbit.', 'Rosai-Dorfman disease may burn out and leave behind fibrotic pockets.']	['Large histiocytes', 'perivascular cuffing', 'fibrotic pockets']	['Dermatopathology', 'Soft tissue', 'Hematopathology']
	['Epidermal acanthosis and papillomatosis resembling a wart or seborrheic keratosis.', 'Presence of large sebaceous glands that drain directly through their duct out to the skin surface, which is abnormal.', 'Presence of a demodex mite.']	['Epidermal acanthosis and papillomatosis', 'large sebaceous glands', 'demodex mite']	['Dermatopathology', 'Soft tissue', 'Hematopathology']
	['Histological description of glandular tissue with little atypia but located in a place where it does not belong can be a helpful criteria to discern the presence of malignancy.', 'Glands located on the periphery and infiltrating into adventitia and peripancreatic tissue may be malignant.']	['glandular tissue', 'pancreas',]	['Gastrointestinal', 'Pancreatic', 'Hematopathology']

Figure 5.2: A collection of sample images from our dataset, accompanied by corresponding medical text, ROI text, and the top three sub-pathology classifications derived from the ASR text using the LLM.

5.3.2 QUILT-1M: Combining QUILT with other histopathology data sources

To create QUILT-1M, we expanded QUILT by adding other histopathology image-text open-access sources: LAION, Twitter, and PubMed.

(1) **PubMed Open Access Articles.** We searched the PubMed open-access from 2010-2022, extracting 59,371 histopathology image-text pairs, using our histopathology classifier and multi-plane figure cropping algorithm. The images are categorized into (1) images that are fully histopathology, (2) multi-plane images that contain histopathology sub-figures, and (3) histopathology sub-figures cropped from (1) and (2).

(2) **Histopathology Image Retrieval from LAION.** The Large-scale Artificial Intelligence Open Network (LAION-5B) [107] curated over 5 billion pairs of images and text from across the Internet, including a substantial volume of histopathology-related data. We tapped into this resource by retrieving 22,682 image and text pairs.

(3) **Twitter Data from OpenPath.** We utilized a list of tweets curated by [50], which totaled up to 55,000 unique tweets and made up 133,511 unique image-text pairs. This exhibits a one-to-many relationship where many images were matched with multiple captions; this differentiated our work from the OpenPath approach. To maintain comparability, we followed their text pre-processing pipeline [50].

5.3.3 Curating QUILT-INSTRUCT

To create our visually-grounded instruction dataset, QUILT-INSTRUCT, we utilized educational videos where narrators often pause while exploring WSIs and indicate salient areas with their cursor [57, 125, 53, 95]. Our process involves three main steps:

(1) **Cursor Localization:** We isolate segments in videos where the background remains mostly static, detecting these "stable chunks" by computing the absolute difference between consecutive frames. We apply a Gaussian filter for adaptive thresholding to

pinpoint frames with minimal changes, further verified by Structural Similarity Index Measure (SSIM). In these stable chunks, we capture the cursor location by identifying the coordinates of the maximum pixel value. To eliminate distractions such as the narrator's facial expressions, we use a face detection model to mask such movements.

(2) **Spatio-Temporal Clustering:** We cluster the cursor points across time, transforming inputs to include spatial and temporal coordinates. An exponential decay function prioritizes points closer in time. Clusters are dynamically formed based on word counts, and each cluster's temporal midpoint is used to map words to clusters, represented by bounding boxes [x1, y1, x2, y2].

(3) **Generating QUILT-INSTRUCT:** Using the grounded captions, we prompt a large language model (LLM) to create instruction-tuning datasets. We generate two sets of question-answer pairs:

   (a) Independent Prompts: These prompts generate Q/A pairs from single image captions, constrained by the context within each image patch [70]. These include conversation based and detailed description Q/A types.

   (b) Reasoning-based Prompts: Leveraging the contextual continuity of the entire WSI, these prompts guide the LLM to reason beyond immediate contexts, using global WSI supporting facts for diagnosis to enhance the depth of generated Q/A pairs. These include complex reasoning and iterative abductive reasoning Q/A types.

By employing these steps, we convert 4149 educational YouTube videos into QUILT-INSTRUCT, ensuring the dataset provides rich, spatially-grounded instruction data suitable for advancing histopathological analysis. Figure 5.3 demonstrates steps 1 and 2 from the pipeline above and Figure 5.4 presents sample Q/A pairs for each prompt type.

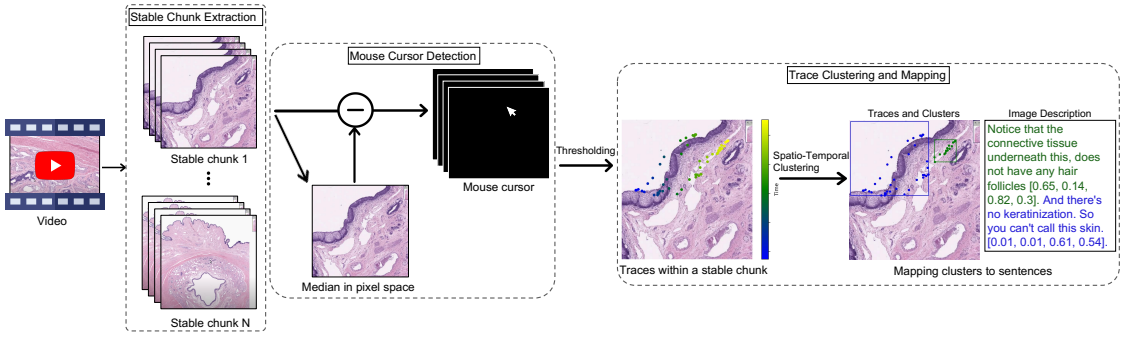


Figure 5.3: To create QUILT-INSTRUCT, we first identify stable chunks within the video. For each chunk, we compute a median frame in the pixel domain and subtract it from every frame within the chunk. We then take the maximum value to capture the mouse cursor points. These cursor points are then clustered to localize medical content in image captions.

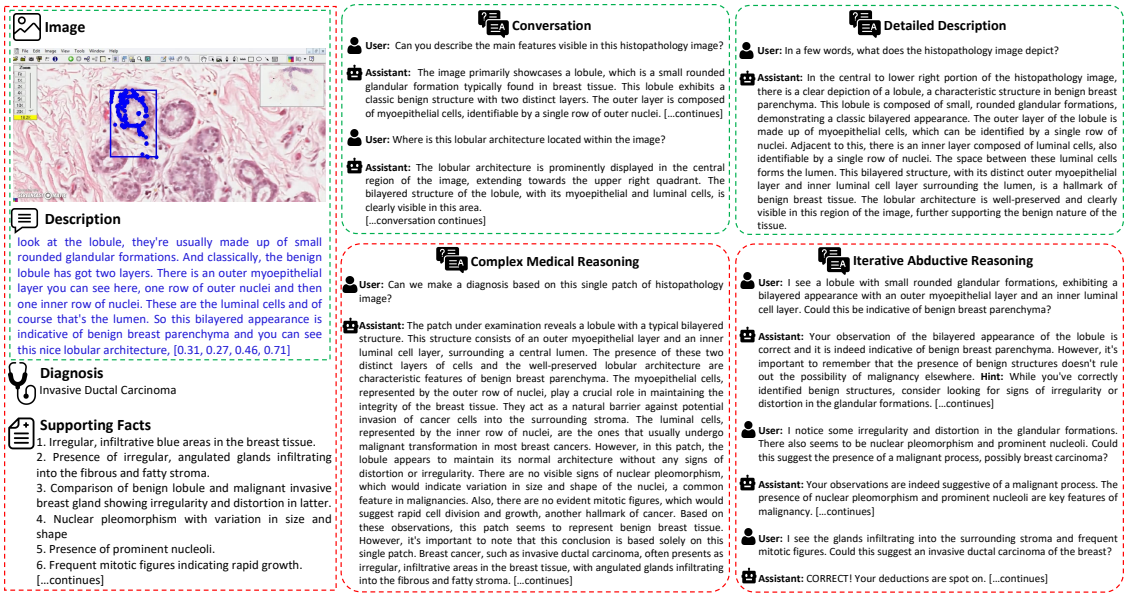


Figure 5.4: A sample from QUILT-INSTRUCT. Only Complex Medical Reasoning and Iterative Abductive Reasoning prompts implicitly utilize Diagnosis and Supporting Facts. Note that some of the content is truncated due to space limitations. Please see Figures A.1-A.14-A.15 for a more complete overview of QUILT-INSTRUCT in the Appendix.

5.4 Methods

5.4.1 QuiltNet: Experiments training with Quilt-1M

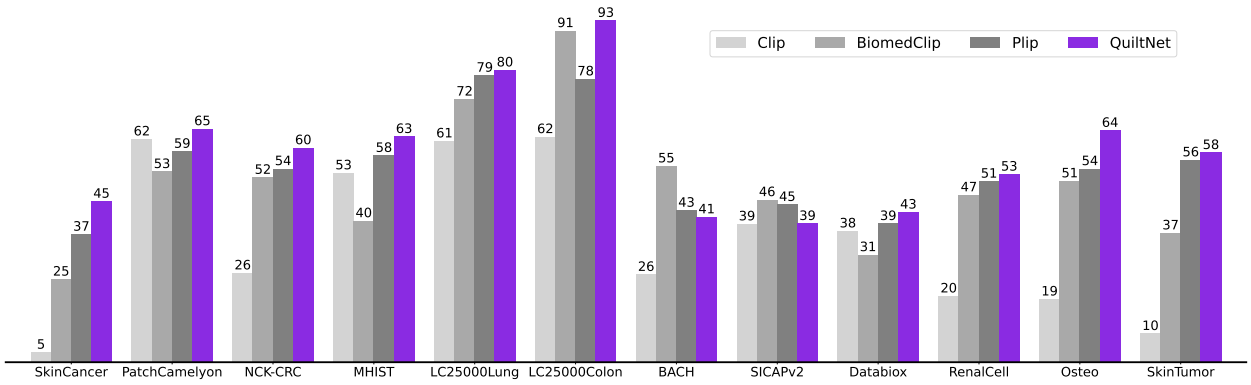


Figure 5.5: QuiltNet, outperforms out-of-domain CLIP baseline and state-of-the-art histopathology models across 12 zero-shot tasks, covering 8 different sub-pathologies (accuracy percentage provided). Details of downstream histopathology datasets can be found in Appendix A.1.

We use the Contrastive Language-Image Pre-training (CLIP) objective [97] to pretrain QuiltNet using Quilt-1M. CLIP takes a batch of N (image, text) pairs and optimizes a contrastive objective to create a joint embedding space. The optimization process involves concurrent training of both image and text encoders to increase the cosine similarity of embeddings from aligned pairs, while decreasing it for unaligned pairs. The objective is minimized via the InfoNCE loss, expressed as:

$$\mathcal{L} = -\frac{1}{2N}\left(\sum_{i=1}^{N}\log\frac{e^{\cos(I_i,\boldsymbol{T}_i)}}{\sum_{j=1}^{N}e^{\cos(I_i,\boldsymbol{T}_j)}} + \sum_{i=1}^{N}\log\frac{e^{\cos(\boldsymbol{I}_i,\boldsymbol{T}_i)}}{\sum_{j=1}^{N}e^{\cos(I_j,T_i)}}\right)$$

where I_i and T_i are the embeddings for the aligned i-th image and text, respectively. For the image encoder, we use both ViT-B/32 and ViT-B/16 architectures [23]. For the text encoder,

we use GPT-2 [99] with a context length of 77, and PubmedBert [137]. We train QUILTNET by finetuning an OpenAI pre-trained CLIP model [97] on QUILT-1M to enhance its performance in histopathology. Once finetuned, we conduct zero-shot image classification experiments. Figure 5.5 demonstrates the performance QUILTNET compared against the CLIP baseline and state-of-the-art histopathology models across 12 tasks. Details of downstream histopathology datasets can be found in Appendix A.1.

5.4.2 Training QUILT-LLAVA

We adopted the LLAVA autoregressive model architecture due to its efficiency and to ensure consistent evaluation against our baselines LLAVA [70] and LLaVA-MED [65]. The LLAVA architecture integrates a vision module, a large language model (LLM), and a multi-layer perceptron (MLP) connector, allowing the LLM to process visual information. Initially, we trained the MLP (projector) while keeping the LLM and vision module frozen. After convergence, we fine-tuned both the MLP and the LLM with instruction-following data to align the model with human pathologists. Our architecture is shown in Figure 5.6. We used a pre-trained CLIP image encoder, adapted for our domain using public histopathology datasets such as QUILTNET [51] and PLIP [50].

1. **Histopathology Domain Alignment:** We first aligned our vision and language models within the histopathology domain. We extracted 723K image-text pairs from QUILT, converting the captions into Q/A format by randomly selecting an instruction (Question) and prepending it to the caption (Answer). The instructions, drawn from a predefined list, variably described the visual information in the images. At this stage, we froze the vision and language models, training only the MLP layer to project the image encoder embeddings to enable the language model to predict the image captions given the questions. This step aligned the histology image embeddings with their corresponding text embeddings.

2. **Histopathology Instruction-Tuning:** Finally, we fine-tuned our model with QUILT-

INSTRUCT, keeping the visual encoder weights frozen and continuing to train the MLP layer and the language module. This stage ensured that the model could effectively follow complex instructions within the histopathology domain.

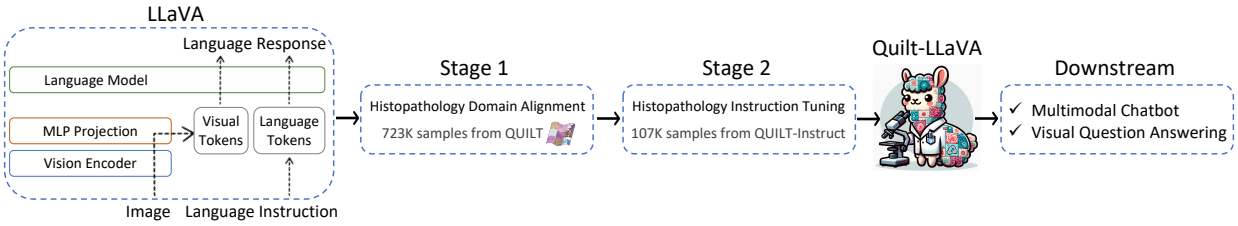


Figure 5.6: QUILT-LLAVA was initialized with the general-domain LLaVA and trained for two stages: Histopathology Domain Alignment on QUILT and instruction-tuning on QUILT-INSTRUCT. We evaluated QUILT-LLAVA on visual conversation and question answering tasks.

5.5 Evaluation and Results

5.5.1 Evaluation data generation

QUILT-VQA. Traditional assessment datasets in histopathology, such as PathVQA [44] and PMC-VQA [138], have issues with contradicting answers and repetitious questions. Rather, the interactive Q/A style that narrators use in educational videos makes them a useful tool for evaluation. In the video, for example, narrators may pose and respond to questions, offering rich Q/A datasets. In order to make use of this, video transcript "?" are identified, and GPT-4 is used to extract question-answer pairs from these texts, if the answer to the question is provided by tha narrator within a 45 second timeframe of the question. Accuracy and medical relevance of these generated Q/A pairs are ensured by manual verification. To fully test the model's image analysis and medical knowledge, the dataset is divided into two categories: image-dependent (1055 Q/A pairs) and general-knowledge (228 Q/A pairs). Figure 5.7 demonstrates samples of this evaluation set.

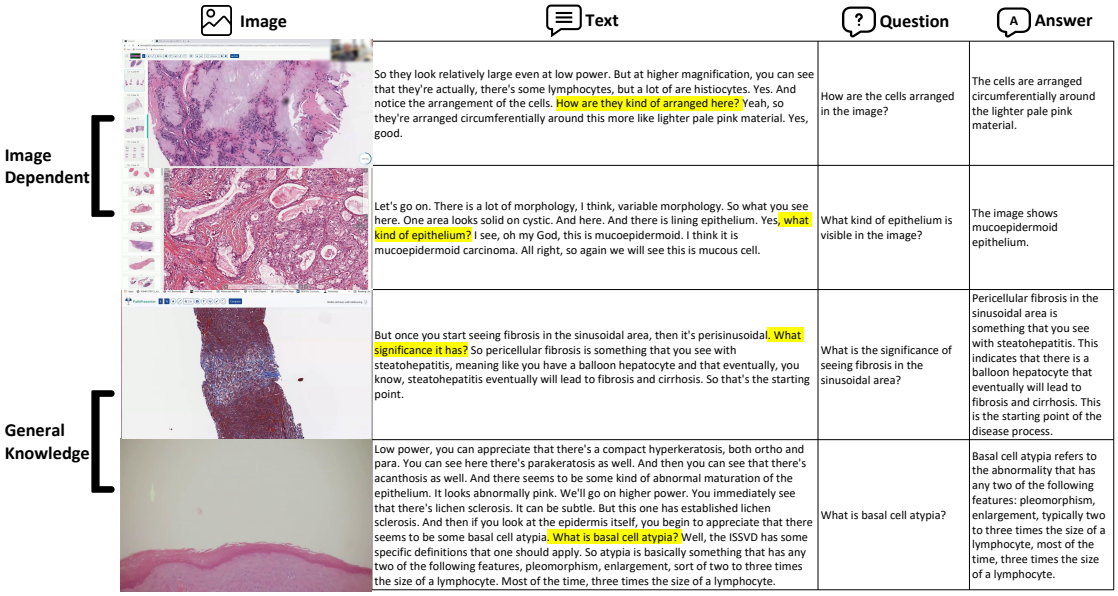


Figure 5.7: Some examples of QUILT-VQA, our evaluation dataset. Top two rows show image-dependent Q/A pairs and bottom two rows show general-knowledge Q/A pairs. The original question posed by the narrator of the video is highlighted in yellow.

Instruction Following Test Set. In addition to QUILT-VQA, which is used to measure QUILT-LLAVA's medical knowledge, we also aimed to analyze the model's capacity to follow instructions in the context of multi-modal chat. In order to do so, we created a collection of 326 questions, comprising 256 conversational and 70 detailed description questions. These questions were all created using image-text pairs that were taken from unseen videos in QUILT-VQA. We utilized the same Conversation and Detailed Description based prompts that we used to build QUILT-INSTRUCT to generate this evaluation set.

5.5.2 Oracle (GPT-4) Alignment Evaluation

Using the Instruction Following test-set and GPT-4 (a language-only model), we evaluated the *helpfulness, relevance, accuracy, and level of detail* of the responses from the candidate model and GPT-4 itself in order to assess the performance of QUILT-LLAVA in multi-modal

conversations. A total score of 1 to 10 was assigned to each response; a higher number denoted superior performance. Furthermore, GPT-4 improved our comprehension of the models by offering thorough justifications for its assessments. As demonstrated in Table 5.1a across 14 sub-pathologies and 2 QA categories, we next computed the relative score by normalizing it against the GPT-4 reference score [65, 70].

With QUILT-LLAVA trained on a single epoch in stage-1 and a balanced selection of 40K instruction-tuning pairs in stage-2 fine-tuning, surpassing LLAVA and LLAVA-MED by more than 16% and 7%, respectively, all QUILT-LLAVA variations performed better than the baselines. Furthermore, we obtained an even higher improvement of 10.8% over LLAVA-MED by augmenting the instruction-tuning data and pre-training for three epochs.

5.5.3 Visual Question Answering Evaluation

Table A.1 in the Appendix contains the details of the three histopathological VQA datasets that we used to evaluate QUILT-LLAVA. Both closed-ended and open-ended Q/A pairs are included in these datasets. Recall was used to determine how frequently model responses included ground-truth tokens in open-set questions, whereas accuracy measured the percentage of accurate answers for closed-set questions [65]. We compared QUILT-LLAVA variations against the general domain LLAVA and the medical domain LLAVA-MED in Table **??**. Every QUILT-LLAVA variation performed better than LLAVA. When combined with pre-trained open-sourced QUILTNET models, the text encoder initialized from Vicuna performed exceptionally well on open-set questions, showing an average 4% improvement over the state-of-the-art (SOTA). On closed-set questions, however, the text encoder initialized from LLAVA fared better, averaging 9% above SOTA.

In all binary and multi-choice QA forms, QUILT-LLAVA demonstrated superior instruction-following ability, surpassing both baselines on PathVQA and QUILT-VQA. Nonetheless, Table A.2 of the Appendix discusses dataset-specific issues that hindered its performance on the PMC-VQA-Subset. In open-set evaluations on PathVQA and QuiltVQA, QUILT-LLAVA showed notable gains over both baselines.

Table 5.1: Results with varying training epochs at different stages and models alongside baselines. 107K indicates the size of instruct data used in Stage-2.

	Question Types		**Sub-Domains**													**Overall**
	Conv	Desc	Bone	Breast	Cyto	Derm	Endo	Gastro	Bone	Geni	Gyne	H&N	Neuro	Pulm	Renal	Soft	
(Question Count)	(256)	(70)	(25)	(23)	(23)	(21)	(23)	(23)	(23)	(22)	(24)	(22)	(24)	(25)	(23)	(25)	(326)
LLaVA [69]	61.4	36.5	54.5	62.0	49.2	48.0	60.1	49.5	62.5	62.2	61.9	49.7	59.7	44.8	53.9	62.7	55.7
LLaVA-MED [65]	70.1	46.9	62.1	69.3	54.1	64.0	61.0	60.7	71.2	68.1	70.3	66.9	66.0	58.9	62.7	73.4	64.8
Quilt-LLaVA @ 40K	76.3	58.7	83.4	73.3	69.2	66.7	71.7	67.2	84.5	81.1	78.4	63.2	68.9	55.2	63.5	87.7	72.3
Quilt-LLaVA @ 107K	78.4	66.0	82.5	84.4	75.0	79.0	76.2	72.8	75.3	82.1	79.1	69.1	68.7	58.1	67.8	89.0	75.6

(a) Performance comparison of multi-modal chat instruction-following abilities, measured by the relative score via language GPT-4 evaluation. Our best model Quilt-LLaVA with ViT-B-32 Vision Encoder [51], 7B Language Model (trained for Stage1: 3 epochs, Stage2:1 epoch) outperforms the baselines.

Quilt-LLaVA Model Variants			**PathVQA**		**PMC-VQA-Subset**	**Quilt-VQA**	
Instruct	Stage 1	Stage 2	Open	Closed	Closed	Open	Closed
QuiltNet ViT-B-32 Vision Encoder[51], 7B Language Model							
107K	1	0	14.34	53.78	27.05	47.69	56.56
107K	1	1	14.24	58.42	19.63	59.82	64.43
107K	1	3	12.79	56.30	17.21	57.62	63.55
107K	3	1	15.30	54.93	16.01	60.97	60.64
LLAVA [69] checkpoint, 7B Language Model							
107K	1	0	11.65	54.03	33.91	55.80	58.02
107K	1	1	15.06	58.68	28.56	55.39	68.81
Baselines							
LLaVA-Med [65] 7B	0	0	11.97	56.15	1.34	54.81	61.22
LLaVA [69] 7B	0	0	11.65	54.02	33.91	55.81	57.73

Table 5.2: Quantitative results on histopathology VQA datasets. For open-set questions, we report recall for our free-form text generation method in column *Open*. For closed-set questions, we report the accuracy in column *Closed*. Red indicates the best-performing model.

5.6 Limitations, Impact, and Conclusion

Even with the promising outcomes obtained using QUILT and QUILT-INSTRUCT, a number of limitations still exist. We leveraged Large language models (LLMs) and handcrafted algorithms in our data curation pipelines. This can introduce biases and inaccuracies to the data. For instance, our histopathology classifier occasionally produced false positive results (about 5%), manually verified by human evaluation. Furthermore, automatic speech recognition (ASR) may misinterprets medical terminology in some instances, such as transcribing 'serous carcinoma' as "serious carcinoma," which cannot be corrected using the proposed pipeline. Additionally, fine-tuning a pre-trained CLIP outperformed training a new CLIP model, indicating that a dataset containing one million image-text pairs might not be sufficient. Further research could examine different self-supervised goals.

Our data which is generated from raw video footage, inherently contains noise which causes errors in our mouse cursor detection and clustering pipeline. Additionally, when generating QUILT-INSTRUCT, GPT-4 occasionally experiences hallucinations in spite of efforts to restrict its context, which causes the model to produce false information. In some cases, GPT-4 may ignore spatially grounded captions when collecting information from the image and instead refer to the captions, producing data that lacks spatial context.

Another critical limitation, is the legal and societal biases. QUILT was derived from public videos, and while steps were taken to limit privacy and consent harms, societal biases remain. Since many narrators of the videos that form QUILT come from Western universities, the emphasis on English-only content may cause the model to perform better on data related to these demographics and under-perform on data from other linguistic or cultural groups. Furthermore, only English is included in the dataset, which limits the QUILTNET and QUILT-LLAVA to a single language.

The introduction of QUILT-1M represents a significant development in the domain of histopathology. With this dataset—the largest open-source image-text histopathology dataset to date—researchers can develop and fine-tune models that perform better than current state-

of-the-art models for a variety of sub-pathology types and tasks, such as cross-modal retrieval, zero-shot, few-shot, and full-shot. Our findings set new standards for the performance of histopathology models and confirm the benefits of pre-training large models with QUILT.

QUILT-INSTRUCT introduces new reasoning-based prompts and spatially grounds histopathological concepts by tracking the narrators' mouse movements in videos taken from QUILT. This innovative approach allowed for the training of QUILT-LLAVA, a multi-modal model that is outperforming baselines in both closed-ended and open-ended histopathological question answering. We introduce and employ QUILT-VQA, an evaluation dataset consisting of human-generated Q/A pairs, to further test QUILT-LLAVA's reasoning capacity. Future study will concentrate on expanding the application to additional medical modalities outside of histopathology and collaborating with pathologists to assess our models in greater detail. In order to enhance the practicality and effectiveness of these models, it will be crucial to look at their multilingual capabilities and fix the issues with the current data curation methods.

Chapter 6

MULTI-AGENT NAVIGATION AND DIAGNOSIS

6.1 Introduction and Motivation

The diagnosis of diseases through the examination of histopathology whole slide images (WSIs) is a cornerstone of modern pathology. These WSIs are high-resolution, digitally scanned histopathology cases, providing an extensive view of tissue architecture and cellular detail. Pathologists navigate these gigapixel-scale images to identify morphological features and spatial relationships critical for accurate diagnoses. This process is not only labor-intensive but also requires significant expertise to interpret the complex visual information effectively. The traditional manual approach, while the gold standard, is becoming increasingly unsustainable due to the rising number of cancer cases globally. This necessitates a shift towards more efficient diagnostic methods without compromising accuracy.

Recent advancements in artificial intelligence (AI) have shown promising potential in transforming medical imaging diagnostics. AI systems, particularly those based on deep learning, have achieved expert-level performance in various medical tasks, offering a pathway to enhance the efficiency and scalability of pathological assessments [119]. However, multiple instance learning (MIL) solutions [64, 133, 145], a popular approach in the field, typically segment WSIs into smaller patches analyzed independently. This approach fails to capture the holistic context of the slides and falls short in replicating the comprehensive and iterative diagnostic approach of pathologists. Pathologists rely on a multi-scale evaluation, starting with low magnification to identify suspicious regions and then zooming in for detailed examination [34, 71]. This multi-turn reasoning process is essential for accurate diagnosis but is not effectively captured by existing AI models. Transformer-based model architectures [109, 129, 38, 142, 16] try to capture both local patterns in individual patches and global

patterns across WSIs by learning long-range interactions between entities. However, they often fall short in efficiency and scalability when dealing with the extremely high resolution and large size of WSIs, making them less practical for real-world clinical use.

Until recently, AI systems for medical imaging have predominantly been developed using vision-only models, such as convolutional neural networks (CNNs) and vision transformers, which require large quantities of expertly annotated data. In contrast, healthcare workflows are typically multi-modal in nature. The emergence of large language models (LLMs) and large multi-modal models (LMMs) has created an entirely new field of multi-modal and generative AI systems. Although current histopathology multi-modal models [115, 1, 131, 51, 108, 37] can analyze isolated image patches effectively, they lack the capability to navigate beyond that patch to identify other patches in the WSI that might contain additional evidence. This issue raises concerns regarding the interpretability and reliability of these models, and whether they accurately align with the diagnostic approach used by pathologists.

With the advancement of large models and the development of specialized models for various tasks, recent research has explored the use of multi-agent collaboration. This approach allows these models to work together, achieving tasks that are beyond the capabilities of any single model alone. This research proposes PathFinder, a multi-modal, multi-agent framework designed to mimic the decision-making process of expert pathologists. PathFinder aims to enhance diagnostic efficiency and accuracy by integrating three AI agents: the Navigation Agent, the Description Agent, and the Diagnosis Agent. These agents work collaboratively to navigate WSI patches, gather evidence, and provide holistic diagnoses with natural language descriptions, thereby offering a more intuitive and precise diagnostic process.

6.2 Related Work

Multi-modal Histopathology Models. The integration of multi-modal data in histopathology has seen significant advancements with the emergence of large language models and vision-language models. Recent studies have explored the potential of multi-modal models to enhance clinical tasks. For instance, studies like Quilt-1M [51], PathGen-1.6M [115]

aim to release large image-text pairs of histopathology and train CLIP-based models to learn joint vision-language embeddings, significantly enhancing clinical downstream tasks. PathAlign [1] is another notable effort that aligns WSIs with corresponding diagnostic texts from pathology reports, facilitating applications such as automatic report generation and case-level visual question answering, moving towards a more integrated and holistic diagnostic process. GigaPath [131] introduces a WSI pathology foundation model pretrained on a large dataset of pathology image patches, leveraging a novel vision transformer architecture to handle the computational challenges of processing gigapixel pathology slides. This approach enables the model to capture both local and global patterns across entire slides, significantly improving its ability to learn from large-scale pathology data. However, these models still face challenges in effectively navigating WSIs towards a diagnosis.

Multi-agent Systems. The concept of multi-agent systems has gained traction in AI research, particularly for tasks requiring dynamic behavior and contextual understanding. Recent research has demonstrated the potential of large foundation models in creating interactive agent-based AI systems including interactions between robots, environments, and humans in the field of robotics [26, 40]. These systems can perform complex tasks by leveraging the strengths of individual agents utilizing collaboration and coordination. The potential of multi-agent systems in handling real-world scenarios has been demonstrated in recent studies including but not limited to role-playing [66], reasoning [25], gaming [48] and software engineering [42]. In medical imaging, multi-agent systems can simulate the collaborative nature of human decision-making processes.

6.3 Methods

The multi-agent multi-modal framework proposed in this study includes three agents: 1) Navigation Agent ; 2) Description Agent ; and 3) Diagnosis Agent . The details of training data and model architectures are described below. Figure 6.1 demonstrates how the three agents interact with each other towards the final goal which is diagnosing a WSI.

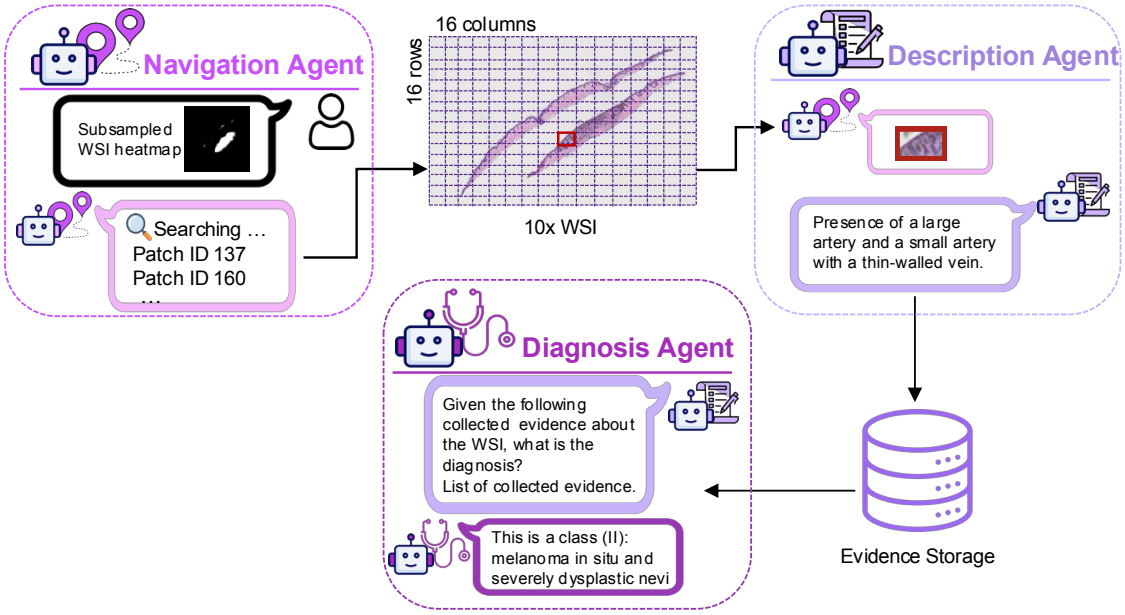


Figure 6.1: **Overview of PathFinder pipeline.** The Navigation Agent starts the process by finding important patches from the given WSI heatmap from our U-Net model. Then the Description Agent generates a textual description of the corresponding high-resolution patch. Lastly, all the collected evidence is passed to the Diagnosis agent to classify the WSI.

6.3.1 Navigation Agent

The Navigation Agent is designed to mimic a pathologist's methodical approach to identifying regions of interest (ROIs) in WSIs. The most critical component of the Navigation Agent is the U-Net model that analyzes the entire WSI to predict a WSI-level heatmap. Following the results of the study presented in chapter 4.1, we train a lightweight U-Net model of 4 layers [102] using our previously curated pathologist viewing behavior dataset from human pathologists (see section for more details). Due to the vast size of WSIs, processing them at full resolution is computationally infeasible. To be computationally efficient, we sub-sample the WSI to a smaller resolution (512x512). The viewing behavior heatmaps dataset reflects the areas that human pathologists focused on before making a diagnosis which is marked as

a binary segmentation mask. Given a low-resolution version of the WSI, the U-Net tries to learn this segmentation mask, which encodes the most important regions according to human pathologists.

The primary role of the Navigation Agent is to select the most relevant ROI from the given WSI heatmaps. To streamline this process, the low-resolution WSI is divided into a 16x16 grid with numbered grid cells in the range of 0 to 255. The Navigation Agent identifies the most pertinent grid number based on heatmap values, which corresponds to the ROI. This grid cell number then serves as a reference in a lookup table, enabling the retrieval of the exact image patch at its original high resolution (10x). We extract 10 patches per WSI and are subsequently forwarded to the Description Agent for further analysis.

6.3.2 Description Agent

We utilize Quilt-LLaVA, a multi-modal large language model capable of describing individual histopathology patches, as our Description Agent. The original Quilt-LLaVA generates findings in a very detailed manner. In this work, to reduce compute cost, we further instruction-tuned the model to generate concise findings for a given histopathology image. We prompted GPT-4 to generate a list of findings as concise as possible, given captions from the Quilt-LLaVA dataset, in the format of "finding": "X", "significance": "Y" where Y could be low, medium, or high. Using this approach, generated 102k samples. Figure 6.2 presents our prompt and sample data generated for instruction-tuning the Description Agent.

6.3.3 Disgnosis Agent

The Disgnosis Agent agent will be a language-only model that analyzes all the gathered natural text descriptions produced by the Description Agent over all the patches identified by the Navigation Agent. It will produce a diagnosis, performing a classification task. The Diagnosis Agent consists of an LLM and a classification head on top of the LLM. The vocabulary size of the LLM and the number of classes (four in our case) are input and output sizes, respectively, for the classification head. The output of the LLM is forwarded to the

GPT-4 Prompt: Description Agent Instruction Tuning Data Generation

System Message: You are an expert in histopathology. Given a histopathology image, I want you to extract medical findings concisely. Even though you will only see the observation in a text format, act like you are seeing the image. Keep your findings very short and to the point. Do not explain the situation, just write the findings. Also return findings in a json format.

Example:

User: of those toxins in the epithelial cells that line these tubules, like literally just one of these epithelial cells, that's what we're talking about here. One of those epithelial cells really, really concentrated with something like a bad antibiotic that you want to get out of your body. And that's how those cells die. That's how you get tubular cell death. Langerhans cells and a child going into the epidermis with a lot of crust on the surface. Very common to see that in this condition. What disease do we think about, obviously? Yeah, we think about Langerhansel histiocytosis in the touch prep. And then you can see the large polygonal cells over here. They can have prominent granulomas. It's very often seen in seminomas. You can have prominent granulomas. You can have signet ring cell-like features. So this is a nice case to demonstrate this concept and also it's a good reminder of just how much things can change from section to section and why we order deeper levels so often in dermatopathology because oftentimes the deeper level really saves the day and shows us the finding of interest.

Assistant:

{"findings": "Concentration of toxic substances within epithelial cells lining tubules.", "significance": medium},
{"findings": "Presence of Langerhans cells infiltrating the epidermis with significant surface crusting.", "significance": medium},
{"findings":"Large polygonal cells with prominent granulomas and signet ring cell-like features observed.", "significance": large}

Figure 6.2: GPT-4 prompt to generate instruction-tuning dataset for the Description Agent.

classification head to get the final classification probabilities. The details of training data and model training are provided in the following sections.

To collect the data for training and evaluation of the Diagnosis Agent, we perform the following tasks: 1) we randomly select patches from each WSI and collect the text descriptions using the Description Agent and 2) we run the Navigation Agent for 5 rounds and at each round, collect 10 patches. Next, since each grid cell can be mapped to a larger image region relevant to the original 10x WSI size, two types of crops are given to the Description Agent: 1) the grid cell cropped from the 10x WSI and 2) a list of 256x256 crops from the grid cell cropped from the 10x WSI. With this approach, we collect both a high-level description of the patch and more detailed descriptions from the smaller crops that form the patch. Finally we combine the training data collected from approaches 1 and 2 to train the model. After we collect all the findings from the Description Agent, we use the following prompt

for the diagnosis task. "The image descriptions below are extracted from different patches from the same whole slide image (WSI); please tell me which class the image belongs to: {descriptions}" with *{description}* being a list of all findings collected from the WSI, each finding in a separate line.

We used QLoRA [22] to train the LLM inside the Diagnosis Agent efficiently. We used $\alpha = 16$, $dropout = 0.1$, and $r = 64$ for the LoRA [47] config. For the QLoRA config, we used the settings recommended in the original paper. We used Cross Entropy Loss with a learning rate of $2e^{-5}$, weight decay of 0.01, and batch size of 4. We use GPT-2 [99] and OLMo-1B [36] for our experiments.

6.4 Evaluation and Results

For the evaluation of our model, we collect evaluation data using two approaches similar to the training data of the Diagnoser Agent described in the previous section: 1) random patch selection and 2) Navigation-based patch selection. We run each experiment 10 times and perform a majority voting evaluation across different runs. We show that the Navigation Agent selected patches improve the classification results compared to the random patch selection approach. We also compare our results with three image-only baseline models and report accuracy, Micro F1-score, Micro Precision and Micro Recall for each experiment in Table 6.1. We train and evaluate the baseline models on the M-Path dataset (see chapter 2 for details) with 190 training cases and 48 test cases. The evaluation of the baseline models are similarly done using the majority voting over 10 runs.

The three baseline models use ScAtNet [129] architecture. This study uses a MobileNetV2 backbone [106] to extract features from the images at three scales of 7.5x, 10x, and 12.5x. These feature vectors are subsequently fed into ScATNet [129] which aggregates information of the three scales to perform the diagnostic task using Transformer blocks. The second approach [33] appends the ROI heatmaps generated by the U-Net model presented in chapter 4 to the WSI as the fourth channel and uses ScAtNet as backbone to perform the classification task. The third baseline model, SAG [71], converts diagnostically relevant entities into

attention signals and integrates these signals with ScAtNet. It utilizes an attention-guiding loss that leverages heuristic guidance (HG) and tissue guidance (TG) which are obtained from disease-specific prior knowledge, such as structures, tissues, and cells.

We hypothesized that employing the Navigation Agent to strategically select ROI patches for diagnosis would yield better results than a naive random patch selection approach. The results from both experiments, using GPT-2 and OLMo-1B, presented in Table 6.1, support this hypothesis. Specifically, the GPT-2 experiments show a 19% improvement and the OLMo-1B experiments demonstrate a 17% enhancement compared to the baseline versions. Our best-performing model, PathFinder-OLMo-1B + Navigation Agent, achieves 4% improvement compared to the image-only baseline approach, ScAtNet (MC) + ROI Heatmap [33]. Considering that GPT-2 and OLMo-1B are relatively small LLMs compared to the current state-of-the-art, we believe that utilizing larger LLMs could further improve diagnostic outcomes.

Table 6.1: Majority voting results of WSI diagnosis on M-Path dataset (see chapter 2.PathFinder baseline experiments are on random patch selection. (MC) is multi-scale.

Methods	Accuracy	Micro F-1 score	Micro Precision	Micro Recall
PathFinder-GPT2 baseline	0.46	0.46	0.46	0.46
PathFinder-GPT2 + Navigation Agent	**0.65**	**0.65**	**0.65**	**0.65**
PathFinder-OLMo-1B baseline	0.50	0.50	0.50	0.50
PathFinder-OLMo-1B + Navigation Agent	**0.67**	**0.67**	**0.67**	**0.67**
Baselines				
ScAtNet (MC) [129]	0.62	0.62	0.62	0.62
ScAtNet (MC) + ROI Heatmap [33]	0.63	0.63	0.63	0.63
ScAtNet (MC) + SAG [71]	0.60	0.60	0.60	0.60

6.5 Limitations, Impact, and Conclusion

This study presents PathFinder, a novel multi-agent, multi-modal AI framework designed to enhance the diagnostic process of histopathology whole slide images. By mimicking the

multi-scale, iterative reasoning approach of expert pathologists, PathFinder aims to improve diagnostic accuracy and efficiency. The framework leverages the strengths of Navigation, Description, and Diagnosis Agents to collaboratively navigate WSIs, gather evidence, and provide holistic diagnoses with natural language descriptions.

PathFinder can replicate the nuanced, iterative diagnostic approach of expert pathologists, thereby reducing the burden of manual examination. This can lead to faster diagnostic turnaround times, which is crucial in clinical settings where timely diagnosis can significantly affect patient outcomes. Additionally, PathFinder's ability to provide natural language descriptions of pathological findings can improve the interpretability of AI-driven diagnostics, making it easier for pathologists to understand and validate the AI's conclusions.

Furthermore, the framework's use of large language models (LLMs) and vision-language models (VLMs) represents a significant advancement in the integration of multi-modal AI systems in healthcare. This approach not only improves diagnostic accuracy but also paves the way for more comprehensive AI applications in medicine. By enhancing the scalability of pathological assessments, PathFinder could address the growing demand for pathology services, particularly in regions with limited access to specialized healthcare professionals.

Despite the promising advancements presented in this study, several limitations should be acknowledged. First, the PathFinder framework relies heavily on pre-existing datasets for training like the viewing behavior data from pathologists. Additionally, the computational demands of processing WSIs at high resolutions pose significant challenges. While the framework employs efficient methods like sub-sampling and multi-agent collaboration, the high resource requirements may hinder its deployment in resource-constrained clinical settings.

Another limitation is the interpretability of the multi-agent system. Although the Navigation, Description, and Diagnosis Agents are designed to mimic the diagnostic reasoning of pathologists, the decision-making process within each agent, particularly the Navigation Agent, may not be fully transparent. This lack of transparency can affect the trust and acceptance of AI-assisted diagnostics among healthcare professionals. Moreover, the model's performance heavily depends on the quality and accuracy of the textual descriptions generated

by the Description Agent. Any inaccuracies in these descriptions could propagate through the system, potentially affecting the final diagnosis.

Future research should focus on expanding the dataset diversity, optimizing computational efficiency, and enhancing the navigation agent patch selection strategy. Despite these challenges, PathFinder represents a substantial step forward in the integration of AI in digital pathology, offering a promising tool to support pathologists and improve patient care outcomes.

Chapter 7

CONCLUSION

In this dissertation, we have explored the applications of digital whole slide imaging (WSI) and artificial intelligence (AI) in enhancing the field of histopathology. Each chapter has contributed to understanding and improving diagnostic processes through innovative AI methodologies and analysis of pathologists' viewing behaviors.

In chapter 3, we focused on leveraging viewing behavior data to investigate associations between viewing behaviors, pathologist characteristics, and diagnostic accuracy. We defined variables to quantify viewing behaviors. Key findings include the negative association between zoom level usage and pathologists' experience and confidence, and the positive correlation between specific viewing behaviors including amount of time pathologists spend on viewing regions of interest (ROIs) and diagnostic accuracy. These insights suggest that digital pathology and AI tools can significantly impact training and clinical practice by identifying critical image regions and enhancing diagnostic accuracy.

In chapter 4, we introduced an ROI detection system based on pathologists' viewing behaviors, resulting in improved diagnostic performance. By utilizing deep learning encoder-decoder architectures to reconstruct heatmaps of important regions, the study demonstrated that integrating pathologists' viewing data can enhance computer-aided diagnostic systems. This chapter highlighted the potential of behavior-driven models in replicating the diagnostic process more realistically than traditional object detection approaches.

In chapter 5, we presented the curation and implications of the QUILT-1M and QUILT-INSTRUCT datasets from YouTube's educational videos and development of vision-language models (QUILTNET and QUILT-LLAVA) for clinical downstream tasks including zero-shot and few-shot image classification, image-text retrieval, and visual question answering. Despite

the limitations of large language models, the innovative approach of spatially grounding histopathological concepts through video annotations has shown promising improvements in multi-modal model performance for histopathological question answering.

In chapter 6, we introduced PathFinder, a multi-agent, multi-modal AI framework designed to mimic the diagnostic approach of expert pathologists. This framework demonstrated advancements in diagnostic accuracy and improving the interpretability of CAD systems outcomes by integrating natural language descriptions. However, it also highlighted challenges related to computational demands and the need for transparent decision-making processes within AI systems.

The research presented in this dissertation underscores the transformative potential of digital pathology and AI in histopathology. By capturing and analyzing pathologists' viewing behaviors, we can develop more effective training programs, improve computer-aided diagnostic systems, and enhance overall diagnostic accuracy and efficiency. The integration of behavior-driven models and multi-modal, multi-agent AI frameworks represents a significant step forward in replicating the nuanced diagnostic processes of expert pathologists. Furthermore, the advancements in dataset development, such as Quilt and Quilt-Instruct, provide valuable resources for training and evaluating AI models in histopathology. These datasets enable the creation of more accurate and interpretable models, ultimately improving the quality of AI-assisted diagnostics.

It's important to recognize that the analysis of different tissue types varies significantly, with some being easier to interpret and others presenting more complex challenges. This variability in tissue analysis underscores the need for tailored AI approaches that can adapt to the specific complexities of each tissue type, enhancing diagnostic accuracy across a broader spectrum of cases.

Moreover, it is essential to acknowledge the inherent uncertainty in histopathological diagnostics. Even experienced pathologists may sometimes be unable to diagnose a case with complete certainty. This uncertainty should be considered in computational analysis, emphasizing the importance of incorporating confidence scores in AI models. By accounting

for this uncertainty, we can ensure that AI systems are not only accurate but also appropriately cautious in cases where the diagnosis is less certain.

Additionally, while digital imaging is gaining popularity, the field is still far from becoming entirely digital. The process of acquiring tissue samples, scanning them, and preparing them for computational analysis is intricate, time-consuming, and not yet streamlined. This reality presents a significant barrier to the widespread adoption of fully digital workflows in histopathology, highlighting the need for further innovation in both technology and processes to achieve a more seamless integration of digital pathology into clinical practice.

Here we list several avenues for future research and development emerging from this dissertation:

- Integration of eye-tracking Data: Incorporating precise eye-tracking data could further refine the models by accurately determining the specific focus points of pathologists within full viewports.

- Scalability and Deployment: Addressing the computational demands and optimizing the efficiency of AI frameworks like PathFinder is crucial for their deployment in resource-constrained clinical settings.

- Multilingual Capabilities: Expanding the datasets and models to include multiple languages will enhance the applicability and effectiveness of AI systems across diverse linguistic and cultural contexts.

- Transparency and Trust: Enhancing the transparency of decision-making processes within AI systems will be vital for gaining trust and acceptance among healthcare professionals.

In conclusion, the research presented in this dissertation highlights the significant potential of digital pathology and AI to revolutionize histopathological diagnostics. By continuing

to address the challenges and exploring new frontiers, we can further enhance the accuracy, efficiency, and accessibility of pathology services, ultimately improving patient care outcomes.

BIBLIOGRAPHY

[1] Faruk Ahmed, Andrew Sellergren, Lin Yang, Shawn Xu, Boris Babenko, Abbi Ward, Niels Olson, Arash Mohtashamian, Yossi Matias, Greg S Corrado, et al. Pathalign: A vision-language model for whole slide images in histopathology. *arXiv preprint arXiv:2406.19578*, 2024.

[2] Iftiaz A. Alfi, Md. Mahfuzur Rahman, Mohammad Shorfuzzaman, and Amril Nazir. A non-invasive interpretable diagnosis of melanoma skin cancer using deep learning and ensemble stacking of machine learning models. *Diagnostics*, 12, 2022.

[3] Guilherme Aresta, Teresa Araújo, Scotty Kwok, Sai Saketh Chennamsetty, Mohammed Safwan, Varghese Alex, Bahram Marami, Marcel Prastawa, Monica Chan, Michael Donovan, et al. Bach: Grand challenge on breast cancer histology images. *Medical image analysis*, 56:122–139, 2019.

[4] Harish Babu Arunachalam, Rashika Mishra, Ovidiu Daescu, Kevin Cederberg, Dinesh Rakheja, Anita Sengupta, David Leonard, Rami Hallac, and Patrick Leavey. Viable and necrotic tumor assessment from whole slide images of osteosarcoma using machine-learning and deep-learning models. *PloS one*, 14(4):e0210706, 2019.

[5] Raymond L Barnhill, David E Elder, Michael W Piepkorn, Stevan R Knezevich, Lisa M Reisch, Megan M Eguchi, Boris C Bastian, Willeke Blokx, Marcus Bosenberg, Klaus J Busam, et al. Revision of the melanocytic pathology assessment tool and hierarchy for diagnosis classification schema for melanocytic lesions: A consensus statement. *JAMA Network Open*, 6(1):e2250613–e2250613, 2023.

[6] Tetiana Biloborodova, Semen Lomakin, Inna Skarga-Bandurova, and Yana Krytska. Region of interest identification in the cervical digital histology images. In *EPIA Conference on Artificial Intelligence*, pages 133–145. Springer, 2022.

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

[8] Hamidreza Bolhasani, Elham Amjadi, Maryam Tabatabaeian, and Somayyeh Jafarali Jassbi. A histopathological image dataset for grading breast invasive ductal carcinomas. *Informatics in Medicine Unlocked*, 19:100341, 2020.

[9] Andrew A Borkowski, Marilyn M Bui, L Brannon Thomas, Catherine P Wilson, Lauren A DeLand, and Stephen M Mastorides. Lung and colon cancer histopathological image dataset (lc25000). *arXiv preprint arXiv:1912.12142*, 2019.

[10] William J Browne, Harvey Goldstein, and Jon Rasbash. Multiple membership multiple classification (mmmc) models. *Statistical Modelling*, 1(2):103–124, 2001.

[11] Otso Brummer, Petri Polonen, Satu Mustjoki, and Oscar Bruck. Integrative analysis of histological textures and lymphocyte infiltration in renal cell carcinoma using deep learning. *bioRxiv*, pages 2022–08, 2022.

[12] Tad T Brunyé, Agnes Balla, Trafton Drew, Joann G Elmore, Kathleen F Kerr, Hannah Shucard, and Donald L Weaver. From image to diagnosis: Characterizing sources of error in histopathologic interpretation. *Modern Pathology*, page 100162, 2023.

[13] Tad T Brunyé, Ezgi Mercan, Donald L Weaver, and Joann G Elmore. Accuracy is in the eyes of the pathologist: the visual interpretive process and diagnostic accuracy with digital whole slide images. *Journal of biomedical informatics*, 66:171–179, 2017.

[14] Patricia A Carney, Lisa M Reisch, Michael W Piepkorn, Raymond L Barnhill, David E Elder, Stevan Knezevich, Berta M Geller, Gary Longton, and Joann G Elmore. Achieving consensus for the histopathologic diagnosis of melanocytic lesions: use of the modified delphi method. *Journal of cutaneous pathology*, 43(10):830–837, 2016.

[15] Hyaline Change. Cellular responses to stress and toxic insults: Adaptation, injury, and death. *Robbins and Cotran Pathologic Basis of Disease, Professional Edition E-Book*, page 1, 2009.

[16] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16144–16155, 2022.

[17] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023.

[18] Emily L Clarke, Ryckie G Wade, Derek Magee, Julia Newton-Bishop, and Darren Treanor. Image analysis of cutaneous melanoma histology: a systematic review and meta-analysis. *Scientific Reports*, 13(1):4774, 2023.

[19] Angel Cruz-Roa, Hannah Gilmore, Ajay Basavanhally, Michael Feldman, Shridar Ganesan, Natalie NC Shih, John Tomaszewski, Fabio A González, and Anant Madabhushi. Accurate and reproducible invasive breast cancer detection in whole-slide images: A deep learning approach for quantifying tumor extent. *Scientific reports*, 7(1):1–14, 2017.

[20] Miao Cui and David Y Zhang. Artificial intelligence and computational pathology. *Laboratory Investigation*, 101(4):412–422, 2021.

[21] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[22] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*, 2023.

[23] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[24] Trafton Drew, Melissa Le-Hoa Vo, Alex Olwal, Francine Jacobson, Steven E Seltzer, and Jeremy M Wolfe. Scanners and drillers: characterizing expert visual search through volumetric images. *Journal of vision*, 13(10):3–3, 2013.

[25] Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.

[26] Zane Durante, Bidipta Sarkar, Ran Gong, Rohan Taori, Yusuke Noda, Paul Tang, Ehsan Adeli, Shrinidhi Kowshika Lakshmikanth, Kevin Schulman, Arnold Milstein, et al. An interactive agent foundation model. *arXiv preprint arXiv:2402.05929*, 2024.

[27] Sebastian Elmes, Tapabrata Chakraborti, Mengran Fan, Holm Uhlig, and Jens Rittscher. Automated annotator: Capturing expert knowledge for free. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 2664–2667. IEEE, 2021.

[28] Joann G Elmore, Raymond L Barnhill, David E Elder, Gary M Longton, Margaret S Pepe, Lisa M Reisch, Patricia A Carney, Linda J Titus, Heidi D Nelson, Tracy Onega, et al. Pathologists' diagnosis of invasive melanoma and melanocytic proliferations: observer accuracy and reproducibility study. *bmj*, 357, 2017.

[29] Sedigheh Eslami, Gerard de Melo, and Christoph Meinel. Does clip benefit visual question answering in the medical domain as much as it does in the general domain? *arXiv preprint arXiv:2112.13906*, 2021.

[30] Jevgenij Gamper and Nasir Rajpoot. 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, 2021.

[31] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.

[32] Alessandro Gherardi, Sara Bravaccini, and Alessandro Bevilacqua. Computer assisted detection of regions of interest in histopathology using a hybrid supervised and unsupervised approach. In *Medical Imaging 2013: Digital Pathology*, volume 8676, pages 288–295. SPIE, 2013.

[33] Fatemeh Ghezloo, Oliver H. Chang, Stevan R. Knezevich, Kristin C. Shaw, Kia Gianni Thigpen, Lisa M. Reisch, Linda G. Shapiro, and Joann G. Elmore. Robust roi detection in whole slide images guided by pathologists' viewing patterns. *Journal of Imaging Informatics in Medicine*, 2024.

[34] Fatemeh Ghezloo, Pin-Chieh Wang, Kathleen F Kerr, Tad T Brunyé, Trafton Drew, Oliver H Chang, Lisa M Reisch, Linda G Shapiro, and Joann G Elmore. An analysis of pathologists' viewing processes as they diagnose whole slide digital images. *Journal of Pathology Informatics*, 13:100104, 2022.

[35] Sydney R Grant, Tom W Andrew, Eileen V Alvarez, Wendy J Huss, and Gyorgy Paragh. Diagnostic and prognostic deep learning applications for histological assessment of cutaneous melanoma. *Cancers*, 14(24):6231, 2022.

[36] Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*, 2024.

[37] Hongyan Gu, Chunxu Yang, Mohammad Haeri, Jing Wang, Shirley Tang, Wenzhong Yan, Shujin He, Christopher Kazu Williams, Shino Magaki, and Xiang'Anthony' Chen. Augmenting pathologists with navipath: design and evaluation of a human-ai collaborative navigation system. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2023.

[38] Ziyu Guo, Weiqin Zhao, Shujun Wang, and Lequan Yu. Higt: Hierarchical interaction graph-transformer for whole slide image analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 755–764. Springer, 2023.

[39] Ricardo Gutiérrez and Eduardo Romero. A visual model approach to extract regions of interest in microscopical images of basal cell carcinoma. In *Diagnostic Pathology*, volume 8, pages 1–4. BioMed Central, 2013.

[40] Shanshan Han, Qifan Zhang, Yuhang Yao, Weizhao Jin, Zhaozhuo Xu, and Chaoyang He. Llm multi-agent systems: Challenges and open problems. *arXiv preprint arXiv:2402.03578*, 2024.

[41] Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*, 2023.

[42] Junda He, Christoph Treude, and David Lo. Llm-based multi-agent systems for software engineering: Vision and the road ahead. *arXiv preprint arXiv:2404.04834*, 2024.

[43] 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.

[44] 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.

[45] Peter Johan Heiberg Engel. Tacit knowledge and visual expertise in medical diagnostic reasoning: implications for medical education. *Medical teacher*, 30(7):e184–e188, 2008.

[46] Md Shakhawat Hossain, Galib Muhammad Shahriar, MM Mahbubul Syeed, Mohammad Faisal Uddin, Mahady Hasan, Shingla Shivam, and Suresh Advani. Region of interest (roi) selection using vision transformer for automatic analysis using whole slide images. *Scientific Reports*, 13(1):11314, 2023.

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

[48] Jen-tse Huang, Eric John Li, Man Ho Lam, Tian Liang, Wenxuan Wang, Youliang Yuan, Wenxiang Jiao, Xing Wang, Zhaopeng Tu, and Michael R Lyu. How far are we on the decision-making of llms? evaluating llms' gaming ability in multi-agent environments. *arXiv preprint arXiv:2403.11807*, 2024.

[49] Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3942–3951, 2021.

[50] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas Montine, and James Zou. Leveraging medical twitter to build a visual–language foundation model for pathology ai. *bioRxiv*, pages 2023–03, 2023.

[51] Wisdom Oluchi Ikezogwo, Mehmet Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Stefan Chan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1m: One million image-text pairs for histopathology. *arXiv preprint arXiv:2306.11207*, 2023.

[52] Kobiljon Ikromjanov, Subrata Bhattacharjee, Yeong-Byn Hwang, Rashadul Islam Sumon, Hee-Cheol Kim, and Heung-Kook Choi. Whole slide image analysis and detection of prostate cancer using vision transformers. In *2022 international conference on artificial intelligence in information and communication (ICAIIC)*, pages 399–402. IEEE, 2022.

[53] Julie A Jacko. Human computer interaction handbook: Fundamentals, evolving technologies, and emerging applications. 2012.

[54] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[55] Shancheng Jiang, Huichuan Li, and Zhi Jin. A visually interpretable deep learning framework for histopathological image-based skin cancer diagnosis. *IEEE Journal of Biomedical and Health Informatics*, 25(5):1483–1494, 2021.

[56] 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. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

[57] Daniel Kahneman. *Attention and effort*, volume 1063. Citeseer, 1973.

[58] Thomas Kalinski, Saadettin Sel, Harald Hofmann, Ralf Zwönitzer, Johannes Bernarding, and Albert Roessner. Digital workflow management for quality assessment in pathology. *Pathology-Research and Practice*, 204(1):17–21, 2008.

[59] Jakob Nikolas Kather, Niels Halama, and Alexander Marx. 100,000 histological images of human colorectal cancer and healthy tissue. *Zenodo10*, 5281, 2018.

[60] Pavel Kisilev, Eli Sason, Ella Barkan, and Sharbell Hashoul. Medical image description using multi-task-loss cnn. In *Deep Learning and Data Labeling for Medical Applications: First International Workshop, LABELS 2016, and Second International Workshop, DLMIA 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 21, 2016, Proceedings 1*, pages 121–129. Springer, 2016.

[61] Katharina Kriegsmann, Frithjof Lobers, Christiane Zgorzelski, Jörg Kriegsmann, Charlotte Janßen, Rolf Rüdinger Meliß, Thomas Muley, Ulrich Sack, Georg Steinbuss, and Mark Kriegsmann. Deep learning for the detection of anatomical tissue structures and neoplasms of the skin on scanned histopathological tissue sections. *Frontiers in Oncology*, 12, 2022.

[62] Elizabeth A Krupinski, Anna R Graham, and Ronald S Weinstein. Characterizing the development of visual search expertise in pathology residents viewing whole slide images. *Human pathology*, 44(3):357–364, 2013.

[63] Elizabeth A Krupinski and Ronald S Weinstein. Changes in visual search patterns of pathology residents as they gain experience. In *Medical Imaging 2011: Image Perception, Observer Performance, and Technology Assessment*, volume 7966, pages 206–211. SPIE, 2011.

[64] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2021.

[65] 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. *arXiv preprint arXiv:2306.00890*, 2023.

[66] G Li, HAAK Hammoud, H Itani, D Khizbullin, and B Ghanem. Camel: communicative agents for" mind" exploration of large scale language model society.(2023). *arXiv preprint arXiv:2303.17760*, 2023.

[67] Jiayun Li, Wenyuan Li, Anthony Sisk, Huihui Ye, W Dean Wallace, William Speier, and Corey W Arnold. A multi-resolution model for histopathology image classification and localization with multiple instance learning. *Computers in biology and medicine*, 131:104253, 2021.

[68] Ruoyu Li and Junzhou Huang. Fast regions-of-interest detection in whole slide histopathology images. In *Patch-MI@MICCAI*, 2015.

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

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

[71] Kechun Liu, Wenjun Wu, Joann G Elmore, and Linda G Shapiro. Semantics-aware attention guidance for diagnosing whole slide images. *arXiv preprint arXiv:2404.10894*, 2024.

[72] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021.

[73] Tahir Mahmood, Muhammad Arsalan, Muhammad Owais, Min Beom Lee, and Kang Ryoung Park. Artificial intelligence-based mitosis detection in breast cancer histopathology images using faster r-cnn and deep cnns. *Journal of clinical medicine*, 9(3):749, 2020.

[74] Komal Mariam, Osama Mohammed Afzal, Wajahat Hussain, Muhammad Umar Javed, Amber Kiyani, Nasir Rajpoot, Syed Ali Khurram, and Hassan Aqeel Khan. On smart gaze based annotation of histopathology images for training of deep convolutional neural networks. *IEEE Journal of Biomedical and Health Informatics*, 26(7):3025–3036, 2022.

[75] Niccolò Marini, Stefano Marchesin, Sebastian Otálora, Marek Wodzinski, Alessandro Caputo, Mart Van Rijthoven, Witali Aswolinskiy, John-Melle Bokhorst, Damian Podareanu, Edyta Petters, et al. Unleashing the potential of digital pathology data by training computer-aided diagnosis models without human annotations. *NPJ digital medicine*, 5(1):102, 2022.

[76] Christian Marzahl, Christof A Bertram, Marc Aubreville, Anne Petrick, Kristina Weiler, Agnes C Gläsel, Marco Fragoso, Sophie Merz, Florian Bartenschlager, Judith Hoppe, et al. Are fast labeling methods reliable? a case study of computer-aided expert annotations on microscopy slides. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, pages 24–32. Springer, 2020.

[77] Claudia R Mello-Thoms, Carlos A Mello, Olga Medvedeva, Eugene Tseytlin, and Rebecca Crowley. Characterizing virtual slide exploration through the use of'search maps'. In *Medical Imaging 2011: Image Perception, Observer Performance, and Technology Assessment*, volume 7966, pages 212–221. SPIE, 2011.

[78] Caner Mercan, Selim Aksoy, Ezgi Mercan, Linda G. Shapiro, Donald L. Weaver, and Joann G. Elmore. Multi-instance multi-label learning for multi-class classification of whole slide breast histopathology images. *IEEE Transactions on Medical Imaging*, 37:316–325, 2018.

[79] Ezgi Mercan, Selim Aksoy, Linda G Shapiro, Donald L Weaver, Tad T Brunyé, and Joann G Elmore. Localization of diagnostically relevant regions of interest in whole slide images: a comparative study. *Journal of digital imaging*, 29:496–506, 2016.

[80] Ezgi Mercan, Linda G Shapiro, Tad T Brunyé, Donald L Weaver, and Joann G Elmore. Characterizing diagnostic search patterns in digital breast pathology: scanners and drillers. *Journal of digital imaging*, 31:32–41, 2018.

[81] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yash Dalmia, Eduardo Pontes Reis, Pranav Rajpurkar, and Jure Leskovec. Med-flamingo: a multimodal medical few-shot learner. *arXiv preprint arXiv:2307.15189*, 2023.

[82] Dilip B Nagarkar, Ezgi Mercan, Donald L Weaver, Tad T Brunyé, Patricia A Carney, Mara H Rendi, Andrew H Beck, Paul D Frederick, Linda G Shapiro, and Joann G Elmore. Region of interest identification and diagnostic agreement in breast pathology. *Modern Pathology*, 29(9):1004–1011, 2016.

[83] Lekha S. Nair, R RamkishorPrabhu, Gowry Sugathan, Kiran V. Gireesh, and Akshay S Nair. Mitotic nuclei detection in breast histopathology images using yolov4. *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–5, 2021.

[84] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy, August 2019. Association for Computational Linguistics.

[85] Shima Nofallah, B. Li, Mojgan Mokhtari, Wenjun Wu, Stevan Knezevich, Caitlin J. May, Oliver H. Chang, Joann G. Elmore, and Linda G. Shapiro. Improving the diagnosis of skin biopsies using tissue segmentation. *Diagnostics*, 12, 2022.

[86] Shima Nofallah, Beibin Li, Mojgan Mokhtari, Wenjun Wu, Stevan Knezevich, Caitlin J May, Oliver H Chang, Joann G Elmore, and Linda G Shapiro. Improving the diagnosis of skin biopsies using tissue segmentation. *Diagnostics*, 12(7):1713, 2022.

[87] Shima Nofallah, Wenjun Wu, Kechun Liu, Fatemeh Ghezloo, Joann G. Elmore, and Linda G. Shapiro. Automated analysis of whole slide digital skin biopsy images. *Frontiers in Artificial Intelligence*, 5, 2022.

[88] Anupiya Nugaliyadde, Kok Wai Wong, Jeremy Parry, Ferdous Sohel, Hamid Laga, Upeka V Somaratne, Chris Yeomans, and Orchid Foster. Rcnn for region of interest detection in whole slide images. In *Neural Information Processing: 27th International Conference, ICONIP 2020, Bangkok, Thailand, November 18–22, 2020, Proceedings, Part V 27*, pages 625–632. Springer, 2020.

[89] Tracy Onega, Raymond L Barnhill, Michael W Piepkorn, Gary M Longton, David E Elder, Martin A Weinstock, Stevan R Knezevich, Lisa M Reisch, Patricia A Carney, Heidi D Nelson, et al. Accuracy of digital pathologic analysis vs traditional microscopy in the interpretation of melanocytic lesions. *JAMA dermatology*, 154(10):1159–1166, 2018.

[90] Yigit Ozen, Selim Aksoy, Kemal Kösemehmetoğlu, Sevgen Önder, and Ayşegül Üner. Self-supervised learning with graph neural networks for region of interest retrieval in histopathology. In *2020 25th International conference on pattern recognition (ICPR)*, pages 6329–6334. IEEE, 2021.

[91] Shreyas Malakarjun Patil, Li Tong, and May D Wang. Generating region of interests for invasive breast cancer in histopathological whole-slide-image. In *2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 723–728. IEEE, 2020.

[92] Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christoph M Friedrich. Radiology objects in context (roco): a multimodal image dataset. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*, pages 180–189. Springer, 2018.

[93] Loïc Peter, Diana Mateus, Pierre Chatelain, Denis Declara, Noemi Schworm, Stefan Stangl, Gabriele Multhoff, and Nassir Navab. Assisting the examination of large histopathological slides with adaptive forests. *Medical Image Analysis*, 35:655–668, 2017.

[94] Michael W Piepkorn, Raymond L Barnhill, David E Elder, Stevan R Knezevich, Patricia A Carney, Lisa M Reisch, and Joann G Elmore. The mpath-dx reporting schema for melanocytic proliferations and melanoma. *Journal of the American Academy of Dermatology*, 70(1):131–141, 2014.

[95] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *Computer Vision–*

ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16, pages 647–664. Springer, 2020.

[96] Dan Popescu, Mohamed El-Khatib, Hassan El-khatib, and Loretta Ichim. New trends in melanoma detection using neural networks: A systematic review. *Sensors (Basel, Switzerland)*, 22, 2022.

[97] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[98] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2022.

[99] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[100] Vignesh Raghunath, Melissa O Braxton, Stephanie A Gagnon, Tad T Brunyé, Kimberly H Allison, Lisa M Reisch, Donald L Weaver, Joann G Elmore, and Linda G Shapiro. Mouse cursor movement and eye tracking data as an indicator of pathologists' attention when viewing digital whole slide images. *Journal of pathology informatics*, 3(1):43, 2012.

[101] Richard J Roberts. Pubmed central: The genbank of the published literature, 2001.

[102] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[103] Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. Automatic keyword extraction from individual documents. *Text mining: applications and theory*, pages 1–20, 2010.

[104] Abhijit Guha Roy, Nassir Navab, and Christian Wachinger. Recalibrating fully convolutional networks with spatial and channel "squeeze and excitation" blocks. *IEEE transactions on medical imaging*, 38(2):540–549, 2018.

[105] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning internal representations by error propagation, parallel distributed processing, explorations in the

microstructure of cognition, ed. de rumelhart and j. mcclelland. vol. 1. 1986. *Biometrika*, 71(599-607):6, 1986.

[106] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[107] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

[108] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13183–13192, 2024.

[109] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021.

[110] Julio Silva-Rodríguez, Adrián Colomer, María A Sales, Rafael Molina, and Valery Naranjo. Going deeper through the gleason scoring scale: An automatic end-to-end system for histology prostate grading and cribriform pattern detection. *Computer methods and programs in biomedicine*, 195:105637, 2020.

[111] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[112] Hardeep Singh and Mark L Graber. Improving diagnosis in health care–the next imperative for patient safety. *The New England journal of medicine*, 373(26):2493–2495, 2015.

[113] Josef Sivic and Andrew Zisserman. Efficient visual search of videos cast as text retrieval. *IEEE transactions on pattern analysis and machine intelligence*, 31(4):591–606, 2008.

[114] Karin Stacke, Jonas Unger, Claes Lundström, and Gabriel Eilertsen. Learning representations with contrastive self-supervised learning for histopathology applications. *arXiv preprint arXiv:2112.05760*, 2021.

[115] Yuxuan Sun, Yunlong Zhang, Yixuan Si, Chenglu Zhu, Zhongyi Shui, Kai Zhang, Jingxiong Li, Xingheng Lyu, Tao Lin, and Lin Yang. Pathgen-1.6 m: 1.6 million

pathology image-text pairs generation through multi-agent collaboration. *arXiv preprint arXiv:2407.00203*, 2024.

[116] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7, 2023.

[117] Thomas E Tavolara, Ziyu Su, Metin N Gurcan, and M Khalid Khan Niazi. One label is all you need: Interpretable ai-enhanced histopathology for oncology. In *Seminars in Cancer Biology*. Elsevier, 2023.

[118] E Llewellyn Thomas and EL Lansdown. Visual search patterns of radiologists in training. *Radiology*, 81(2):288–292, 1963.

[119] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019.

[120] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[121] Darren Treanor, Chee Hooi Lim, Derek Magee, Andy Bulpitt, and Phil Quirke. Tracking with virtual slides: a tool to study diagnostic error in histopathology. *Histopathology*, 55(1):37–45, 2009.

[122] Jeroen Van der Laak, Geert Litjens, and Francesco Ciompi. Deep learning in histopathology: the path to the clinic. *Nature medicine*, 27(5):775–784, 2021.

[123] Mike Van Zon, Nikolas Stathonikos, Willeke AM Blokx, Selim Komina, Sybren LN Maas, Josien PW Pluim, Paul J Van Diest, and Mitko Veta. Segmentation and classification of melanoma and nevus in whole slide images. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 263–266. IEEE, 2020.

[124] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*, pages 210–218. Springer, 2018.

[125] Paul Voigtlaender, Soravit Changpinyo, Jordi Pont-Tuset, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with video localized narratives. In *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2461–2471, 2023.

[126] Noorul Wahab and Asifullah Khan. Multifaceted fused-cnn based scoring of breast cancer whole-slide histopathology images. *Applied Soft Computing*, 97:106808, 2020.

[127] Jerry Wei, Arief Suriawinata, Bing Ren, Xiaoying Liu, Mikhail Lisovsky, Louis Vaickus, Charles Brown, Michael Baker, Naofumi Tomita, Lorenzo Torresani, et al. A petri dish for histopathology image analysis. In *Artificial Intelligence in Medicine: 19th International Conference on Artificial Intelligence in Medicine, AIME 2021, Virtual Event, June 15–18, 2021, Proceedings*, pages 11–24. Springer, 2021.

[128] Gezheng Wen, Avigael Aizenman, Trafton Drew, Jeremy M Wolfe, Tamara Miner Haygood, and Mia K Markey. Computational assessment of visual search strategies in volumetric medical images. *Journal of Medical Imaging*, 3(1):015501–015501, 2016.

[129] Wenjun Wu, Sachin Mehta, Shima Nofallah, Stevan Knezevich, Caitlin J May, Oliver H Chang, Joann G Elmore, and Linda G Shapiro. Scale-aware transformers for diagnosing melanocytic lesions. *IEEE Access*, 9:163526–163541, 2021.

[130] Bin Xin, Yaning Yang, Xiaolan Xie, Jiandong Shang, Zhengyu Liu, and Shaoliang Peng. Detecting and classifying nuclei using multi-scale fully convolutional network. *Journal of computational biology : a journal of computational molecular cell biology*, 2022.

[131] Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, pages 1–8, 2024.

[132] Pavel Yakubovskiy. Segmentation models pytorch. https://github.com/qubvel/segmentation_models.pytorch, 2020.

[133] Zhaochang Yang, Ting Wei, Ying Liang, Xin Yuan, Ruitian Gao, Yujia Xia, Jie Zhou, Yue Zhang, and Zhangsheng Yu. A foundation model for generalizable cancer diagnosis and survival prediction from histopathological images. *bioRxiv*, pages 2024–05, 2024.

[134] Moi Hoon Yap, Manu Goyal, Fatima Osman, Robert Martí, Erika Denton, Arne Juette, and Reyer Zwiggelaar. Breast ultrasound region of interest detection and lesion localisation. *Artificial Intelligence in Medicine*, 107:101880, 2020.

[135] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

[136] Hai Yu, Wei Yang, Shi Wu, Shaohui Xi, Xichun Xia, Qianchuan Zhao, Wai-Kit Ming, Lifang Wu, Yunfeng Hu, Liehua Deng, and Jun Lyu. Deep-learning-based survival prediction of patients with cutaneous malignant melanoma. *Frontiers in Medicine*, 10, 2023.

[137] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, et al. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*, 2023.

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

[139] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pages 2–25. PMLR, 2022.

[140] Zizhao Zhang, Pingjun Chen, Mason McGough, Fuyong Xing, Chunbao Wang, Marilyn Bui, Yuanpu Xie, Manish Sapkota, Lei Cui, Jasreman Dhillon, et al. Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. *Nature Machine Intelligence*, 1(5):236–245, 2019.

[141] Chen Zhao, Ren Shuai, Li Ma, Wenjia Liu, Die Hu, and Menglin Wu. Dermoscopy image classification based on stylegan and densenet201. *IEEE Access*, 9:8659–8679, 2021.

[142] Yi Zheng, Rushin H Gindra, Emily J Green, Eric J Burks, Margrit Betke, Jennifer E Beane, and Vijaya B Kolachalama. A graph-transformer for whole slide image classification. *IEEE transactions on medical imaging*, 41(11):3003–3015, 2022.

[143] Yushan Zheng, Zhiguo Jiang, Fengying Xie, Jun Shi, Haopeng Zhang, Jianguo Huai, Ming Cao, and Xiaomiao Yang. Diagnostic regions attention network (dra-net) for histopathology wsi recommendation and retrieval. *IEEE transactions on medical imaging*, 40(3):1090–1103, 2020.

[144] Yushan Zheng, Zhiguo Jiang, Haopeng Zhang, Fengying Xie, Jun Shi, and Chenghai Xue. Histopathology wsi encoding based on gcns for scalable and efficient retrieval of diagnostically relevant regions. *arXiv preprint arXiv:2104.07878*, 2021.

[145] Qifeng Zhou, Wenliang Zhong, Yuzhi Guo, Michael Xiao, Hehuan Ma, and Junzhou Huang. Pathm3: A multimodal multi-task multiple instance learning framework for whole slide image classification and captioning. *arXiv preprint arXiv:2403.08967*, 2024.

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

[147] Xiaochun Zou, Xinbo Zhao, Yongjia Yang, and Na Li. Learning-based visual saliency model for detecting diabetic macular edema in retinal image. *Computational intelligence and neuroscience*, 2016:1–1, 2016.

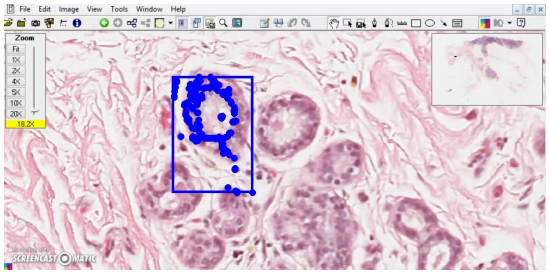
Appendix A

VISION-LANGUAGE MODELS IN HISTOPATHOLOGY

A.1 Downstream histopathology datasets.

We evaluate the utility of QuiltNet on 13 downstream datasets: **PatchCamelyon** [124] contains histopathology scans of lymph node sections labeled for metastatic tissue presence as a binary label. **NCT-CRC-HE-100K** [59] consists of colorectal cancer images and is categorized into cancer and normal tissue. For **SICAPv2** [110] the images are labeled as non-cancerous, Grade 3-5. **Databiox** [8] consists of invasive ductal carcinoma cases of Grades I-III. **BACH** [3] consists of breast tissues labeled as normal, benign, in-situ, and invasive carcinoma. **Osteo** [4] is a set of tissue patches representing the heterogeneity of osteosarcoma. **RenalCell** [11] contains tissue images of clear-cell renal cell carcinoma annotated into five tissue texture types. **SkinCancer** [61] consists of tissue patches from skin biopsies of 12 anatomical compartments and 4 neoplasms that make up the **SkinTumor** Subset. **MHIST** [127] contains tissue patches from Formalin-Fixed Paraffin-Embedded WSIs of colorectal polyps. **LC25000** [9], which we divide into **LC25000 (Lung)** and **LC25000 (Colon)**, contains tissue of lung and colon adenocarcinomas. For more details see supplemental material.

Image



Description

look at the lobule, they're usually made up of small rounded glandular formations. And classically, the benign lobule has got two layers. There is an outer myoepithelial layer you can see here, one row of outer nuclei and then one inner row of nuclei. These are the luminal cells and of course that's the lumen. So this bilayered appearance is indicative of benign breast parenchyma and you can see this nice lobular architecture, [0.31, 0.27, 0.46, 0.71]

Conversation

User: Can you describe the main features visible in this histopathology image?

Assistant: The image primarily showcases a lobule, which is a small rounded glandular formation typically found in breast tissue. This lobule exhibits a classic benign structure with two distinct layers. The outer layer is composed of myoepithelial cells, identifiable by a single row of outer nuclei. The inner layer, on the other hand, consists of luminal cells, marked by another row of nuclei. The space between these cells forms the lumen. This bilayered appearance is a characteristic feature of benign breast parenchyma.

User: Where is this lobular architecture located within the image?

Assistant: The lobular architecture is prominently displayed in the central region of the image, extending towards the upper right quadrant. The bilayered structure of the lobule, with its myoepithelial and luminal cells, is clearly visible in this area.

User: What does this bilayered appearance indicate about the condition of the breast tissue?

Assistant: The bilayered appearance, with distinct myoepithelial and luminal cell layers, is indicative of benign breast parenchyma. This means that the breast tissue appears to be healthy and non-cancerous. However, it's important to note that a definitive diagnosis would require a comprehensive clinical evaluation and possibly additional tests.

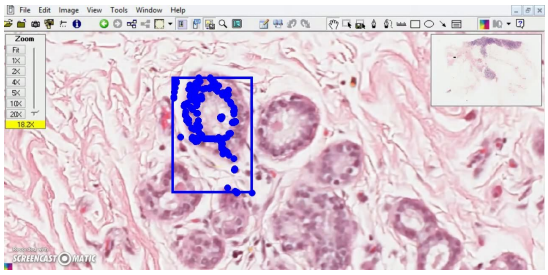
Detailed Description

User: In a few words, what does the histopathology image depict?

Assistant: In the central to lower right portion of the histopathology image, there is a clear depiction of a lobule, a characteristic structure in benign breast parenchyma. This lobule is composed of small, rounded glandular formations, demonstrating a classic bilayered appearance. The outer layer of the lobule is made up of myoepithelial cells, which can be identified by a single row of nuclei. Adjacent to this, there is an inner layer composed of luminal cells, also identifiable by a single row of nuclei. The space between these luminal cells forms the lumen. This bilayered structure, with its distinct outer myoepithelial layer and inner luminal cell layer surrounding the lumen, is a hallmark of benign breast tissue. The lobular architecture is well-preserved and clearly visible in this region of the image, further supporting the benign nature of the tissue.

Figure A.1: A full example of Conversation and Detailed Description types of Q/A in Quilt-Instruct.

Image



Description

look at the lobule, they're usually made up of small rounded glandular formations. And classically, the benign lobule has got two layers. There is an outer myoepithelial layer you can see here, one row of outer nuclei and then one inner row of nuclei. These are the luminal cells and of course that's the lumen. So this bilayered appearance is indicative of benign breast parenchyma and you can see this nice lobular architecture, [0.31, 0.27, 0.46, 0.71]

Diagnosis

Invasive Ductal Carcinoma

Supporting Facts

1. Irregular, infiltrative blue areas in the breast tissue.
2. Presence of irregular, angulated glands infiltrating into the fibrous and fatty stroma.
3. Comparison of benign lobule and malignant invasive breast gland showing irregularity and distortion in latter.
4. Nuclear pleomorphism with variation in size and shape
5. Presence of prominent nucleoli.
6. Frequent mitotic figures indicating rapid growth.
7. Increased nuclear cytoplasmic ratios.
8. Poorly formed, irregular glands on low power view.
9. Infiltrative growth pattern where the glands invade into the surrounding stroma.

Complex Medical Reasoning

User: Can we make a diagnosis based on this single patch of histopathology image?

Assistant: The patch under examination reveals a lobule with a typical bilayered structure. This structure consists of an outer myoepithelial layer and an inner luminal cell layer, surrounding a central lumen. The presence of these two distinct layers of cells and the well-preserved lobular architecture are characteristic features of benign breast parenchyma.

The myoepithelial cells, represented by the outer row of nuclei, play a crucial role in maintaining the integrity of the breast tissue. They act as a natural barrier against potential invasion of cancer cells into the surrounding stroma. The luminal cells, represented by the inner row of nuclei, are the ones that usually undergo malignant transformation in most breast cancers.

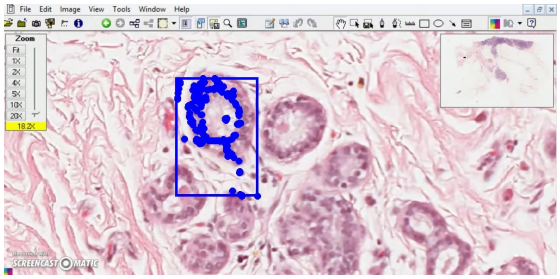
However, in this patch, the lobule appears to maintain its normal architecture without any signs of distortion or irregularity. There are no visible signs of nuclear pleomorphism, which would indicate variation in size and shape of the nuclei, a common feature in malignancies. Also, there are no evident mitotic figures, which would suggest rapid cell division and growth, another hallmark of cancer.

Based on these observations, this patch seems to represent benign breast tissue. However, it's important to note that this conclusion is based solely on this single patch. Breast cancer, such as invasive ductal carcinoma, often presents as irregular, infiltrative areas in the breast tissue, with angulated glands infiltrating into the fibrous and fatty stroma. These features are not evident in this patch, but they could be present in other areas of the tissue.

Therefore, while this patch appears benign, a definitive diagnosis would require a comprehensive examination of the whole slide image. This would allow for the identification of any areas of irregularity or distortion, increased nuclear cytoplasmic ratios, or poorly formed glands that might suggest an invasive growth pattern. Additionally, correlation with clinical information and possibly additional immunohistochemical studies would be necessary to confirm the diagnosis.

Figure A.14: A full example of Complex Medical Reasoning type of Q/A in QUILT-INSTRUCT.

Image



Description

look at the lobule, they're usually made up of small rounded glandular formations. And classically, the benign lobule has got two layers. There is an outer myoepithelial layer you can see here, one row of outer nuclei and then one inner row of nuclei. These are the luminal cells and of course that's the lumen. So this bilayered appearance is indicative of benign breast parenchyma and you can see this nice lobular architecture, [0.31, 0.27, 0.46, 0.71]

Diagnosis

Invasive Ductal Carcinoma

Supporting Facts

1. Irregular, infiltrative blue areas in the breast tissue.
2. Presence of irregular, angulated glands infiltrating into the fibrous and fatty stroma.
3. Comparison of benign lobule and malignant invasive breast gland showing irregularity and distortion in latter.
4. Nuclear pleomorphism with variation in size and shape
5. Presence of prominent nucleoli.
6. Frequent mitotic figures indicating rapid growth.
7. Increased nuclear cytoplasmic ratios.
8. Poorly formed, irregular glands on low power view.
9. Infiltrative growth pattern where the glands invade into the surrounding stroma.

Iterative Abductive Reasoning

User: I see a lobule with small rounded glandular formations, exhibiting a bilayered appearance with an outer myoepithelial layer and an inner luminal cell layer. Could this be indicative of benign breast parenchyma?

Assistant: Your observation of the bilayered appearance of the lobule is correct and it is indeed indicative of benign breast parenchyma. However, it's important to remember that the presence of benign structures doesn't rule out the possibility of malignancy elsewhere. **Hint:** While you've correctly identified benign structures, consider looking for signs of irregularity or distortion in the glandular formations. Also, pay attention to the nuclear characteristics. Are there any signs of nuclear pleomorphism or prominent nucleoli? These could be indicative of a more sinister process.

User: I notice some irregularity and distortion in the glandular formations. There also seems to be nuclear pleomorphism and prominent nucleoli. Could this suggest the presence of a malignant process, possibly breast carcinoma?

Assistant: Your observations are indeed suggestive of a malignant process. The presence of nuclear pleomorphism and prominent nucleoli are key features of malignancy. However, the specific type of breast carcinoma might still be unclear. **Hint:** Consider the growth pattern of the glands. Are they infiltrating into the surrounding stroma? Also, look for signs of rapid growth, such as frequent mitotic figures. These features could help you narrow down the type of breast carcinoma.

User: I see the glands infiltrating into the surrounding stroma and frequent mitotic figures. Could this suggest an invasive ductal carcinoma of the breast?

Assistant: CORRECT!!! Your observations and deductions are spot on. The presence of glands infiltrating into the surrounding stroma and frequent mitotic figures are indeed indicative of an invasive process, specifically invasive ductal carcinoma of the breast. **Hint:** None needed. You've made the correct diagnosis based on the evidence available.

Figure A.15: A full example of Iterative Abductive Reasoning type of Q/A in QUILT-INSTRUCT.

	PathVQA	**PMC-Subset**	**Quilt-VQA**
# Images	858	2209	985
# QA Pairs	6761	2318	1283
# Open	3370	-	940
# Closed	3391	2318	343

Table A.1: VQA datasets statistics. For PathVQA, we report the evaluation set statistics. PMC-Subset is the histopathology subset retrieved from PMC-VQA.

Category	**Details and Example**
Wrong answers	PathVQA includes questions with wrong answers. E.g., Q: Did early ischemic injury increase eosinophilia? A: No. (Correct answer is "Yes" according to Figure 1-9 in [15].)
Inconsistent Q/A pairs	PathVQA has similar questions with inconsistent answers. E.g., Q: Does early ischemic injury show surface blebs, increase eosinophilia of cytoplasm, and swelling of occasional cells? A: Yes. (Contradicts with example in "Wrong answers".)
Ambiguous questions	PathVQA includes ambiguously phrased questions. The question below starts with "What is showing" and the answer provides a diagnosis. The question should have been more clear as "what is the diagnosis of" or "what has the following symptoms?" E.g., Q: What is showing increased eosinophilia of cytoplasm? A: early (reversible) ischemic injury.
Repetitive Q/A pairs	Both PathVQA and PMC-VQA generate multiple Q/A pairs for a single image-text pair, leading to repetition. E.g., two different questions in PathVQA about early ischemic injury. (Refer to examples in "Wrong answers" and "Inconsistent Q/A pairs".)
General-domain Q/A pairs	Most VQA datasets focus on general-domain biomedical questions. E.g., PMC-VQA is generated from PubMed articles and we extracted a histopathology subset from it for evaluation purposes. In addition, PathVQA includes many non-histopathology images.

Table A.2: Summary and examples of issues in public VQA datasets.