

Towards Autonomous Histopathological Diagnosis: An End-to-End Multi-Agent AI Framework for Diagnostic Decision-Making and Interpretation

Mehmet Saygin Seyfioglu

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

Doctor of Philosophy

University of Washington

2025

Reading Committee:

Linda Shapiro, Chair

Ranjay Krishna, Chair

Jenq-Neng Hwang

Program Authorized to Offer Degree:
Electrical and Computer Engineering

©Copyright 2025

Mehmet Saygin Seyfioglu

University of Washington

Abstract

Towards Autonomous Histopathological Diagnosis: An End-to-End Multi-Agent AI Framework for Diagnostic Decision-Making and Interpretation

Mehmet Saygin Seyfioglu

Co-Chairs of the Supervisory Committee:

Linda Shapiro

Electrical and Computer Engineering

Ranjay Krishna

Computer Science and Engineering

The rising global incidence of cancer cases necessitates the development of AI-assisted diagnostic systems to support pathologists in analyzing Whole Slide Images (WSIs). While artificial intelligence (AI) holds promise in transforming medical imaging diagnostics, current methods often face challenges with the gigapixel scale of WSIs and struggle to provide holistic, interpretable explanations for their decisions. This dissertation introduces multiple interconnected contributions aimed at developing an end-to-end AI system for interpretable histopathological diagnosis. First, we present QUILT-1M, the largest multi-modal histopathology dataset to date, comprising over one million image-text pairs curated from open-source educational videos. QUILT-1M addresses the critical challenge of data scarcity in histopathology and enables the development of QUILTNET, a vision-language model that achieves state-of-the-art performance in zero-shot classification and cross-modal retrieval tasks. Building on this foundation, we developed QUILT-INSTRUCT to create QUILT-LLAVA, a large language and vision assistant specifically tailored for histopathology. QUILT-LLAVA can analyze WSI patches in detail, spatially localize medical concepts, and perform reasoning that extends beyond individual patches. Next, we introduce PathFinder, a multi-modal, multi-agent framework that mimics the diagnostic workflow of expert pathologists. By orchestrating

specialized AI agents for triage, navigation, description, and diagnosis, PathFinder delivers an interpretable diagnostic process for entire WSIs. Pathfinder establishes a new state-of-the-art in melanoma classification, even surpassing the average performance of human pathologists by 9%. Finally, through MedicalNarratives, we demonstrate the potential to expand this approach beyond histopathology into broader medical domains by leveraging the abundance of educational content available. This dissertation advances the field of AI-assisted pathology by creating a complete pipeline—from large-scale dataset curation to sophisticated multi-modal models to coordinated AI agents—resulting in diagnostic systems that collaborate effectively with pathologists while providing human readable predictions throughout the decision-making process.

TABLE OF CONTENTS

	Page
List of Figures	iii
List of Tables	vi
Chapter 1: Introduction	1
Chapter 2: Fueling the Pathology AI Research with Large Scale Multi-Modal Dataset - QUILT-1M	5
2.1 Introduction and Motivation	5
2.2 Related Work	7
2.3 Curating QUILT	8
2.4 Training QUILTNET with QUILT-1M	15
2.5 Conclusion	18
Chapter 3: Large Language and Vision Assistant for Histopathology - QUILT-LLAVA	20
3.1 Introduction and Motivation	20
3.2 Related work	23
3.3 Curating QUILT-INSTRUCT	25
3.4 Training QUILT-LLAVA & evaluating with QUILT-VQA	32
3.5 Experiments	35
3.6 Conclusion and Limitations	39
Chapter 4: Automating Diagnostic Process for Histopathology with Multi-Modal Multi-Agent Framework - PATHFINDER	41
4.1 Introduction and Motivation	41
4.2 Related Work	42
4.3 Datasets	44
4.4 PathFinder	45

4.5	Experiments and Results	50
4.6	Discussion	55
Chapter 5:	Beyond Histopathology - Targeting the Entire Medical Domain with MedicalNarratives	57
5.1	Introduction and Motivation	57
5.2	Related Work	59
5.3	MedicalNarratives: Curation, Formatting, and Characterization	61
5.4	GenMedCLIP: Experiments and Results	64
5.5	Conclusion	67
5.6	Limitations and Future Work	68
Chapter 6:	Conclusion	70
Bibliography	73
Appendix A:	Appendix	88

LIST OF FIGURES

Figure Number	Page
2.1 The process begins with identifying relevant histopathology YouTube videos in the Search phase. During the Image extraction stage, histopathology frames are located and de-noised using trained models. In the Text phase, Automatic Speech Recognition (ASR) is applied, followed by post-processing and error correction with the Unified Medical Language System (UMLS) and large language models (LLMs) to obtain text. Finally, the Pairing phase matches images with text, removes duplicates, and produces QUILT, a richly annotated histopathology image-text dataset [53].	11
2.2 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.	14
2.3 QUILTNET demonstrates superior performance compared to the out-of-domain CLIP baseline and state-of-the-art histopathology models in 12 zero-shot tasks across 8 distinct sub-pathologies. Accuracy percentages are shown for reference.	16
3.1 QUILT-LLAVA can describe key medical regions in a histopathology patch and also reason through the observations to support a diagnosis. Note: The image features eosinophils and lymphocytes and is taken from a WSI depicting rare benign dermatitis.	21
3.2 A sample from QUILT-INSTRUCT. Only Complex Medical Reasoning and Iterative Abductive Reasoning prompts implicitly utilize Diagnosis and Supporting Facts (which are extracted from the entire video). Note that some of the content is truncated due to space limitations. Please see Figs. A.13-A.14-A.15 in the Appendix for a more complete overview of QUILT-INSTRUCT [110]. . .	24
3.3 The QUILT-INSTRUCT dataset was created by first segmenting videos into stable chunks. For each chunk, the median frame was computed in the pixel domain and subtracted from all frames within that chunk. These cursor points were then clustered in time and space to spatially ground medical content. In the "Trace Clustering and Mapping" section of the figure, time progression is represented by color.	26

3.4	A set of instructions designed to provide varied and detailed descriptions of the visual content in images, used to create instruction-following QA pairs with comprehensive descriptions.	35
3.5	We initialized QUILT-LLAVA with the general-domain LLaVA and trained it for two stages: 1) Histopathology Domain Alignment on QUILT and 2) Instruction-tuning on QUILT-INSTRUCT. We evaluated QUILT-LLAVA on visual conversation and question answering tasks.	36
4.1	PATHFINDER consists of multiple steps carried out by multi-modal agents: 1) Initial Assessment by Triage Agent; 2) Evidence Collection by Navigation and Description Agents; and 3) Integrated Diagnosis by Diagnosis Agent. . .	42
4.2	The left panel illustrates the Navigation Agent, as outlined in Section 4.4.2. The right panel presents the iterative trajectory generation process, which employs both the Navigation Agent and Description Agent, as described in Section 4.4.4.	45
4.3	Ablation results. We ran 10 experiments, and plotted both the mean and standard deviation.	51
4.4	Preferences of expert human pathologists regarding the description quality of each model, assessed through a double-blind survey to ensure an unbiased comparison.	52
5.1	The data curation pipeline for the Video subset of the MedicalNarratives dataset. Search: Videos are collected from YouTube using a focused search approach, refined through pre-trained classifiers and heuristic methods. Image: Keyframes are extracted, cleaned, and evaluated to detect medical imagery. Text: Audio is transcribed via ASR, rectified using LLMs, and processed to extract medical or ROI-relevant text. Traces: Stable video segments are segmented, and cursor traces are mapped within these segments. Alignment: Medical/ROI text, traces, and images are aligned, with samples categorized into specific subdomains or labeled as multi-domain using LLMs.	59
5.2	MedicalNarratives: We present four dataset samples, each representing a distinct medical modality: X-ray, CT, MRI, and Histopathology images (displayed from left to right). These examples, extracted from interleaved videos, include the representative image captured, the raw input text aligned with spatial traces and bounding boxes, and the refined medical and ROI text, which has been denoised to eliminate transcription errors and non-medical content.	60

5.3	Zeroshot Classification Results shows that our model GenMedCLIP outperforms all other baselines across multiple medical benchmarks covering all 11 medical domains. Mean average precision is reported for X-ray and Mammography and accuracy is used for all other domains.	65
A.1	Using few-shot examples, the LLM is prompted to perform two key tasks. First, it refines noisy ASR text by identifying and correcting errors within the context, with corrections marked in gray and additional issues flagged in blue. Second, it extracts relevant information, distinguishing medical (MED) text in blue and ROI text in gray.	89
A.2	An example illustrating where the QUILT-LLAVA is used as an educational tool, providing clues to student pathologists to help them train.	90
A.13	An example of Conversation and Detailed Description types of Q/A in QUILT-INSTRUCT.	91
A.14	An example of Complex Medical Reasoning type of Q/A in QUILT-INSTRUCT.	92
A.15	An example of Iterative Abductive Reasoning type of Q/A in QUILT-INSTRUCT.	93

LIST OF TABLES

Table Number	Page
2.1 Linear probing. Classification accuracy results (with standard deviation). PatchCamelyon is referred to as Camelyon. Supervised benchmarks represent SOTA models for each dataset.	17
2.2 Cross-modal retrieval performance is reported for the QUILT-1M holdout set and the ARCH dataset. Results are presented in the format (%/%), with the QUILT-1M holdout scores shown on the left and ARCH scores on the right. Best is bold	18
3.1 Quantitative results on histopathology VQA datasets. For open-set questions, we report recall in column <i>Open</i> . For closed-set, we report accuracy in <i>Closed</i> . Red is best.	38
4.1 Majority voting performance for whole slide image (WSI) diagnosis on the M-Path dataset. Accuracy is reported, and the F-1 score is identical due to the balanced testing set. Finally, coverage here is the percent of patches used across all trajectories.	56
5.1 Linear Probing results (Accuracy) across datasets for classification tasks in Dermatology (<i>Isic</i>), Histopathology (<i>pcam</i> , <i>mhist</i> , <i>nck</i>), and Mammography (<i>vinDr-Mammo</i>). GenMedCLIP demonstrates superior performance compared to all baselines.	64
5.2 Retrieval results on our held-out set of 16K samples spanning all medical domains demonstrate that our model, GenMedCLIP, outperforms all baselines in both zeroshot image-to-text and text-to-image retrieval tasks.	66
A.2 VQA datasets statistics. For PathVQA, we report the evaluation set statistics. PMC-Subset is the histopathology subset retrieved from PMC-VQA.	88

ACKNOWLEDGMENTS

Completing this dissertation has been a truly a hard but rewarding journey, made possible by the guidance and support of countless individuals.

To Professor Linda Shapiro, thank you for welcoming me into your lab and tirelessly advocating for your students. Your unwavering support during challenging times, along with your generosity, patience, and caring leadership, has set a standard of academic mentorship I hope to uphold in my own career, should I follow the path of academia one day. To Professor Ranjay Krishna, your technical expertise and visionary approach have expanded the boundaries of my thinking. You challenged me to see the bigger picture and to strive for excellence every step of the way. I would also like to thank my other dissertation committee members, Professor Jenq-Neng Hwang and Professor John C. Kramlich, for their constructive feedback on my research proposal and for continually encouraging me to refine my ideas. Finally, Professor Joann Elmore, your guidance in cancer research, healthcare systems, and clinical processes greatly enriched my understanding of the complexities of patient care and the challenges faced by providers.

I would also like to express my sincere gratitude to the Fulbright Program and the Turkish Fulbright Commission. Their generous support was instrumental in allowing me to embark on this path and in creating an environment where international collaboration and academic exploration could flourish.

My thanks extend to the administrative and technical staff at UW ECE and UW CSE, whose tireless efforts made it possible for me to focus on my research goals rather than logistical hurdles. Their commitment behind the scenes ensured a smooth journey through countless details and deadlines.

I have been fortunate to collaborate with remarkable individuals who enriched my work with their energy, expertise, and camaraderie: Dr. Fatemeh Ghezloo, Dr. Nicholas Nuechterlein, Dr. Beibin Li, Dr. Sheng Wang, Dr. Tom Grabowski, Dr. Sachin Mehta, Kevin Zhang, Rustin Soraki, Zucks Liu, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Pranav Kamath, and Sadjyot Gangolli. Each of you played a part in shaping this dissertation. In particular, I would like to acknowledge Wisdom Ikezogwo—not only my most dedicated collaborator but also my closest friend during this journey. From late-night coding sessions to spirited brainstorming, our partnership fueled many of the insights that underpin this work, and your support was indispensable.

I am also grateful to my lab mates and friends who provided a supportive environment and a sense of community: Dr. Shima Nofallah, Kalyani Marathe, Mahtab Bigverdi, Dr. Wenjun Wu, Nishat Khan, Sitong Liu, Kechun Liu, Kaitlin Flores, Yuguang Li, Jiafei Duan, Jieyu Zhang, Peter Sushko. Your friendship, your willingness to discuss ideas, and the countless moments of humor and empathy have made this journey not just tolerable, but genuinely enjoyable.

Finally, I owe my deepest thanks to my parents. Your unwavering faith in my abilities and your boundless encouragement carried me through moments of doubt and difficulty. I could not have reached this milestone without your steady presence. And to my partner, Anna—your constant support and patience sustained me through the most challenging periods of this journey. Thank you for reminding me that life’s richest rewards come not just from hard work, but from sharing our burdens and hopes with those we love.

DEDICATION

In memory of my grandmother Semahat.

To my loving parents, Sukriye and Mahmut.

To my beloved Anna.

Chapter 1

INTRODUCTION

Cancer is a disease characterized by the uncontrolled growth and division of abnormal cells in the body. In healthy tissue, cells follow an orderly process of growth, division, and death. However, when this process is disrupted, damaged or mutated cells can continue to multiply unchecked, potentially forming tumors that can invade nearby tissues and spread to other parts of the body through the bloodstream or lymphatic system. With more than 100 distinct types affecting various organs and tissues, from common forms like breast and lung cancer to rarer variants, cancer presents an enormous challenge to healthcare systems worldwide. As global populations age, cancer incidence continues to rise, with millions of new cases diagnosed each year, making it one of the leading causes of mortality worldwide [113].

The microscopic examination of tissue samples through hematoxylin and eosin (H&E) stained whole slide images (WSIs) remains the gold standard for cancer diagnosis. These WSIs are gigapixel-level images created by digitally scanning glass slides containing tissue specimens stained with H&E dyes, which highlight different cellular components - hematoxylin staining cell nuclei blue-purple and eosin staining proteins pink. These high-resolution images provide pathologists with detailed views of tissue architecture and cellular characteristics critical for accurate diagnosis, allowing them to identify subtle changes in cell morphology, tissue organization, and other features that may indicate the presence and severity of cancer.

However, the traditional approach of manual WSI examination faces mounting challenges. Examining WSIs is a meticulous manual process that requires careful attention to detail across multiple magnification levels [92, 37, 79] This time-intensive nature of WSI examination, combined with the increasing volume of cancer cases and limited availability of trained pathologists, has created an unsustainable bottleneck in cancer diagnosis. The cognitive

burden of analyzing numerous WSIs daily can impact diagnostic accuracy - studies have shown that even experienced pathologists can make errors [35] when examining large volumes of WSIs, with diagnostic disagreement rates reaching up to 60% for certain types of cancer such as melanoma in-situ and T1a invasive melanoma [28]. This variability in diagnostic outcomes highlights the urgent need for advanced methods to assist pathologists in their interpretation processes and clinical practices.

While artificial intelligence (AI) has shown promise in medical image analysis, current approaches face significant limitations. Many existing AI systems either struggle to process gigapixel-level WSIs effectively or produce "black box" decisions without providing interpretable explanations - a critical shortcoming in healthcare where understanding the reasoning behind a diagnosis is essential for aid-based clinical decision-making. Furthermore, instead of analyzing the entire WSI, most AI-based approaches often focus on patch-level analysis or employ multiple instance learning techniques that lack the holistic understanding and reasoning capabilities of human pathologists [69, 40, 111, 128, 141, 13].

To address these challenges, there is a clear need for an end-to-end AI framework that can process WSIs in a manner similar to clinical workflow - from initial triage to final diagnosis - while providing human-readable explanations for its diagnostic decisions. However, developing such a comprehensive system requires addressing three fundamental challenges: the scarcity of large-scale histopathology training data, the need for sophisticated multi-modal models capable of understanding both visual and textual information, and the necessity of integrating these models into a complex framework for automating the entire diagnostic workflow.

Building on these requirements, we first introduced QUILT-1M [53], a dataset of over one million image-text pairs curated from open source educational histopathology videos. By leveraging the rich narrative content of pathology lectures and demonstrations, we developed novel algorithms to extract aligned image-caption pairs, ensuring high-quality associations between visual content and expert descriptions. QUILT-1M not only represents the largest histopathology vision-language dataset to date - five times larger than existing alternatives - but also captures the natural language explanations and reasoning processes of expert

pathologists, making it an invaluable resource for training AI models that can understand and communicate about histopathological findings in human-interpretable terms.

Building upon QUILT-1M’s foundation, we then developed QUILT-INSTRUCT [110], an instruction-tuning dataset that grounds medical concepts within histopathology images, and QUILT-LLAVA, a large language and vision assistant specifically designed for histopathology. QUILT-INSTRUCT contains over 100,000 question-answer pairs derived from the broader contextual information in educational videos, enabling more sophisticated reasoning capabilities. Using both datasets, QUILT-LLAVA was trained through a two-stage process: first aligning with the histopathology domain using QUILT-1M, then fine-tuning with QUILT-INSTRUCT. This resulted in a multi-modal model capable of not only analyzing individual image patches in detail but also reasoning beyond them, guiding users on what additional evidence might be needed for diagnosis, and providing educational support for pathology training.

Culminating these advances, we present PathFinder, a novel multi-agent system that automates the entire diagnostic workflow. PathFinder orchestrates multiple specialized AI agents working in concert: a Triage Agent that identifies cases requiring detailed examination, a Navigation Agent that systematically explores gigapixel-level WSIs to identify diagnostically relevant regions, a Description Agent (powered by QUILT-LLAVA) that analyzes these regions and articulates findings in natural language, and a Diagnosis Agent that synthesizes these observations into a final assessment. Unlike traditional AI systems that rely on patch-level analysis or exhaustive scanning, PathFinder follows a human-like diagnostic workflow, iteratively refining its focus with each step. This collaborative agent-based framework enables accurate and interpretable diagnoses, surpassing state-of-the-art AI models and even outperforming the average diagnostic accuracy of expert pathologists in challenging tasks such as melanoma classification. By providing natural language explanations for its decisions, PathFinder offers transparency, and integration into clinical workflows, setting a new bar for state of the art in AI-driven pathology.

Finally, we introduce MedicalNarratives, a multimodal dataset for the broader medical domain, leveraging the video processing framework established in QUILT-1M [53] and the

cursor detection methodology of QUILT-LLAVA [110]. MedicalNarratives expands the principles of grounded image-text curation to include over 4.2 million image-text pairs across 11 medical modalities, with 875,000 samples featuring spatial grounding through cursor traces. By sourcing pedagogical medical videos where instructors narrate and point to relevant regions, we mimicked the think-aloud protocol widely used in clinical education to ensure natural alignment between visual and textual data. MedicalNarratives enables diverse applications such as open-vocabulary segmentation, dense modeling, and grounded generative modeling, and serves as a critical resource for advancing medical AI systems, demonstrated through our vision-language model, GenMedCLIP, trained on this dataset.

This dissertation presents four interconnected projects, each building on the achievements of its predecessors to create a comprehensive framework for automated, interpretable histopathological diagnosis. Chapter 2 introduces QUILT-1M, addressing the critical challenge of data scarcity by curating a large-scale vision-language dataset. Chapter 3 delves into QUILT-LLAVA, exploring its architecture and capability to bridge visual and linguistic understanding in histopathology. Chapter 4 presents PathFinder, a multi-agent system designed to automate the entire diagnostic workflow while ensuring interpretability and accuracy. Chapter 5 extends the framework beyond histopathology with MedicalNarratives, creating a dataset that spans broader medical domains. Finally, Chapter 6 concludes the dissertation, discussing future directions. Together, these projects offer a blueprint for developing interpretable AI systems that collaborate with human experts to improve patient care while maintaining transparency in decision-making.

Chapter 2

FUELING THE PATHOLOGY AI RESEARCH WITH LARGE SCALE MULTI-MODAL DATASET - QUILT-1M

In this chapter, we outline the details of extracting QUILT-1M [53], from open-source venues such as videos and PubMed articles. We highlight how we used handcrafted algorithms alongside ML models to extract a vision-language dataset from videos, and we describe how we trained QUILTNET, which learns superior embeddings for histopathology patches that surpass the state of the art in zero-shot classification and retrieval-based tasks.

2.1 Introduction and Motivation

Cancer diagnosis and treatment planning heavily rely on the microscopic examination of tissue samples through whole slide images (WSIs). These WSIs are created by digitally scanning glass slides containing tissue specimens stained with hematoxylin and eosin (H&E) dyes, which highlight different cellular components - with hematoxylin staining cell nuclei blue-purple and eosin staining proteins pink. These high-resolution, gigapixel-level images provide pathologists with detailed views of tissue architecture and cellular characteristics critical for accurate diagnosis, enabling them to identify subtle changes in cell morphology, tissue organization, and other features indicative of various pathological conditions.

Within these WSIs, even individual patches contain intricate patterns and complex morphological features crucial for disease characterization [115]. The richness of this visual information poses a unique challenge: traditional categorical labels fail to capture the nuanced relationships between different tissue structures, cell types, and their spatial arrangements. With thousands of disease subtypes and continuously evolving classification systems, reducing these complex visual patterns to simple labels severely limits our ability to represent and

understand the full spectrum of histopathological features. Natural language descriptions emerge as a powerful alternative, offering the flexibility and expressiveness needed to capture detailed observations about tissue architecture, cellular characteristics, and their relationships [31].

Recent years have witnessed remarkable advances in vision-language modeling for natural images, with models like CLIP [98], VILT [66], and ALBEF [75] demonstrating unprecedented capabilities in understanding the relationship between visual content and textual descriptions. These models, trained on large-scale image-text pairs, have achieved impressive results in tasks ranging from zero-shot classification to cross-modal retrieval. However, this success has not translated to the histopathology domain, primarily due to the scarcity of comprehensive vision-language datasets. Existing resources like ARCH [31] ($\approx 8K$ samples) and OpenPath [49] ($\approx 200K$ samples) provide valuable but limited data. While PMC-15M [137] offers a larger collection of biomedical image-text pairs, its histopathology coverage remains unclear, and the dataset is not publicly accessible. To address this critical gap, we introduce QUILT, a large-scale vision-language dataset for histopathology containing 800K images and their text descriptions across multiple magnification scales (10x to 40x). We leverage an untapped resource: educational histopathology videos on YouTube, comprising 1,087 hours of expert pathologist narrations. Through a sophisticated pipeline combining specialized computer vision models, automated speech recognition, and carefully designed algorithms, we extract and align high-quality image-text pairs from these videos. By combining QUILT with additional data from diverse sources including Twitter, research papers, and other public repositories, we create QUILT-1M, a million-scale dataset that represents the largest public vision-language resource for histopathology to date.

Building on this foundation, we develop QUILTNET, a vision-language model trained using contrastive learning objectives on our dataset. Through extensive evaluation across 13 external histopathology datasets spanning various sub-pathologies, we demonstrate QUILTNET’s superior performance in zero-shot classification, linear probing, and cross-modal retrieval tasks. Our model consistently outperforms existing state of the art approaches including CLIP

[98], PLIP [49], and BiomedCLIP [137], validating the quality and utility of our dataset. This chapter presents several key contributions to the field of computational pathology: (1) the creation of QUILT, a novel large-scale vision-language dataset derived from expert pathology lectures with the development of specialized techniques for extracting aligned image-text pairs from educational videos; and (2) the implementation and validation of QUILTNET, demonstrating state-of-the-art performance across multiple histopathology tasks.

2.2 *Related Work*

The application of machine learning to histopathology has evolved significantly over the past decade. Early approaches to computational pathology were primarily based on weakly-supervised learning, where entire whole-slide images (WSIs) were assigned single labels. This simplistic labeling scheme led to suboptimal model performance at the patch level [15, 54]. To address these limitations, researchers began exploring self-supervised learning approaches, which showed promise in learning useful representations from unlabeled data [54, 13, 15].

Initial self-supervised methods in histopathology largely borrowed techniques from natural image processing, such as data augmentation strategies [17]. However, researchers soon recognized the need to incorporate domain-specific knowledge. This led to innovations such as leveraging the compositional nature of H&E stain information [54], incorporating hierarchical morphological information across different magnification levels [13], and integrating complementary data modalities like genomic features [15].

The development of vision-language datasets in the medical domain has followed an uneven trajectory. While fields like radiology have benefited from relatively straightforward data collection processes, leading to comprehensive datasets like MIMIC-CXR-JPG v2.0.0 [62], histopathology has faced unique challenges. The Cancer Genome Atlas, while providing de-identified PDF reports for WSIs, presents significant processing challenges due to the enormous size of WSI files, often exceeding $120,000^2$ pixels [82]. This has limited its utility in large-scale studies. Efforts to create vision-language datasets specifically for histopathology have been limited in scope. The ARCH dataset represented an early attempt, containing

only 7,614 image-text pairs [31, 46]. A more recent effort, OpenPath, expanded this to 200K pairs by leveraging Twitter content [49]. While PMC-15M [137] offered a larger collection of biomedical image-text pairs, its histopathology coverage remains unclear, and the dataset is not publicly available.

Recent years have seen increasing interest in leveraging video data for model training. Researchers have successfully utilized millions of YouTube videos for vision-language model development [135, 134], and gaming videos for training causal video models [5]. The concept of Localized Narratives [122, 95] has demonstrated the value of dense, interconnected supervision for single images. While video sources typically introduce more noise compared to static image sources, advances in automatic speech recognition have made it increasingly feasible to curate large-scale, clean datasets from video content [135, 5, 137].

2.3 Curating QUILT

As mentioned, a large-scale image-text dataset is needed in histopathology and YouTube offers a plethora of educational videos that remain untapped and can be used to extract an image-text dataset.

Creating a vision-language dataset from videos is a non trivial task, as not all educational histopathology videos are usable. Many either lack voiced audio, are not in English, fail to contain medically relevant content, or have insufficient medical relevance. For instance, some videos may only display static histopathology images without further analysis, or they might mention histopathology briefly for unrelated purposes. The use of standard automatic speech recognition (ASR) models is often inadequate due to the specific jargon and terminology used in histopathology, presenting a challenge that requires sophisticated solutions. The de-noising of text and image modalities adds further complexity as the videos are typically conversational and, therefore, inherently noisy. Instructors pan and zoom at varying speeds, recording a mix of relevant and irrelevant histopathological visual content in their videos. Simple frame extraction methods are insufficient for accurately capturing the intended data. To assemble QUILT, we have developed and utilized models along with custom algorithms

that take advantage of the specific ways instructors convey information visually and textually, ensuring the precise collection and synchronization of both image and text modalities.

Our dataset curation pipeline involves (1) gathering relevant videos covering the histopathology domain, (2) filtering videos based on a certain "narrative style", (3) extracting and denoising image and text modalities from videos using various models, tools, and algorithms, (4) postprocessing denoised text by LLMs to extract medical text and finally, (5) aligning image and text modalities in time to obtain the final version of QUILT. See Fig. 2.1 for a detailed overview of the pipeline.

2.3.1 Collecting representative channels and videos.

The initial step in our data curation pipeline involves identifying YouTube channels and videos that are pertinent to histopathology. We employ a strategy that utilizes keywords related to 18 different sub-fields of pathology to streamline the search process. This approach prioritizes channel searches over individual video queries to enhance efficiency and mitigate the constraints imposed by API request limits [135]. To refine our search further, we exclude channels with a subscriber count exceeding 300K to bypass broad-based scientific channels, focusing instead on niche educational histopathology content, which typically attracts a smaller audience. Following selection, we download all identified videos.

2.3.2 Filtering for narrative-style medical videos.

We filter out the videos less than one minute in duration, lacking voiced narration, or in languages other than English. For videos meeting these heuristics, we check for two more conditions: **1)** the presence of essential medical content, specifically histopathology image-text pairs, and **2)** whether the video demonstrates a narrative style. This narrative style is defined by the extent to which narrators engage with the WSIs, including spending time on panning and zooming within the WSI, accompanied by detailed vocal descriptions of the visual content.

For **1)**, we employ an automated method to pinpoint relevant videos by extracting keyframes, utilizing FFmpeg ¹ for this task. These keyframes, which indicate the start or finish of a scene marked by significant visual changes, help in identifying content-rich videos. The software requires a threshold that determines the minimum amount of visual change required to trigger a keyframe. Through experimentation, we set different thresholds for various video durations, with smaller thresholds for longer videos. Next, we train a simple histopathology image classifier to identify videos with histopathology images.

For **2)**, the process of determining whether a video exhibits a narrative style involves analyzing randomly chosen keyframes predicted to contain histopathology content. For each selected keyframe, we extract the subsequent three frames deemed to be histopathology-related and calculate the cosine similarity between the initial frame and each of the following three. If all three frames have similarity scores equal to or above a defined threshold of 0.9, the sequence is classified as a narrative streak, which means that the narrator talked about a certain region within the broader WSI. A video is identified as narrative style if at least 10% of the selected frames exhibit a narrative streak. Consequently, we download all narrative-style videos at high-resolution. Narrative-style videos typically cover WSIs at various magnifications, hence, we train a tissue-image-magnification classifier to predict the following three scales: $\{(1 - 10)\times, (> 10 - 20)\times, (> 20)\times\}$. This provides relevant metadata for downstream objectives.

2.3.3 Text extraction using ASR and text denoising.

The prohibitive costs of specialized medical ASR APIs² led us to opt for the more accessible ASR model, Whisper [99], despite its known limitations in accurately recognizing medical terminology. This necessitates a subsequent phase of post-processing to correct misinterpretations and reduce error rates significantly.

We developed a comprehensive four-step process for enhancing our text quality: **i)** Initially,

¹<https://ffmpeg.org/>

²nuance.com/en-au/healthcare/provider-solutions/speech-recognition/dragon-medical-one.html

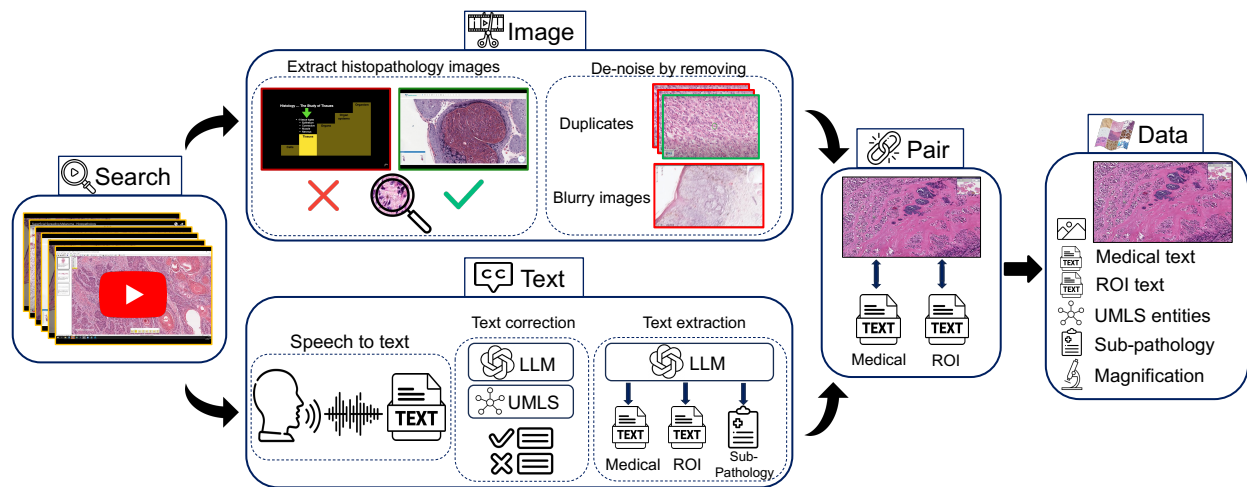


Figure 2.1: The process begins with identifying relevant histopathology YouTube videos in the **Search** phase. During the **Image** extraction stage, histopathology frames are located and de-noised using trained models. In the **Text** phase, Automatic Speech Recognition (ASR) is applied, followed by post-processing and error correction with the Unified Medical Language System (UMLS) and large language models (LLMs) to obtain text. Finally, the **Pairing** phase matches images with text, removes duplicates, and produces QUILT, a richly annotated histopathology image-text dataset [53].

we employ the RAKE keyword extraction algorithm to identify and refine keywords or phrases, removing any stopwords in the process [105]. **ii)** The refined entries are then cross-checked against the Unified Medical Language System (UMLS) [6] using the SciSpacy entity linking package [89]. For entries that are not found in UMLS, we apply a spell-checker³ preloaded with a comprehensive list of histopathology-related terms derived from various histopathology ontology labels and definitions. This step generates a list of potentially misspelled medical keywords. **iii)** With this probable list of misspelled keywords, we *condition* and prompt the LLM with examples to correct the misspelled entry within its context (sentence), and

³<https://github.com/barrust/pyspellchecker>

secondly, we task the LLM with identifying additional *unconditioned* errors/misspelled entries. For both, we leverage a set of manually curated examples to prompt the LLM in-context. (Shown in Fig. A.1 in the Appendix). **iv)** The final step involves confirming the accuracy of the corrections by cross-referencing the adjusted terms with UMLS and our specialized histopathology terminology list. This dual verification ensures that only correctly spelled and contextually appropriate terms replace the original, error-prone transcriptions, effectively de-noising the text.

Upon refining the text with ASR corrections, we extract two key types of textual content: *medical text*, which broadly describes the images, and *ROI text*, derived from instances where the speaker points out specific visual regions of interest with phrases like "look here...". (Please note that, in this work, we did not utilize ROI text, however we made our datasets available for other researchers to utilize and thought ROI text could have a use case.) To filter medically relevant text from the corrected transcriptions, we leverage LLMs. The decision to use LLMs for this task stems from two fundamental reasons. First, the cost of hiring experts for such a large-scale task is very high. LLMs, despite not being perfect, offer a viable alternative because a dataset at our scale can handle a certain amount of noise. Second, we are using LLMs not to create new information but to check the existing content for relevance. This is a simpler task for LLMs. By using specific prompts, we guide LLMs to specifically exclude any text that is not medically relevant, thus ensuring that the final dataset strictly consists of histopathological content. Also, please note that we instruct the LLMs to ensure they do not add any new words beyond what is present in the ASR-corrected text, and set the LLMs temperature to zero. Finally, we further leverage LLMs to classify our images into one of the predefined 18 sub-pathology categories to generate additional metadata.

2.3.4 Image Frame Extraction and Denoising.

In each video, we utilize a technique similar to the one outlined in Sec.2.3.1 for isolating key histopathology frames. This approach uses the timestamps of these key frames as markers to segment the video into discrete time segments, referred to as *chunks*, from which we then

extract representative images. Subsequently, we perform a deduplication process on these key histopathology frames to ensure each chunk is represented by a unique image.

Overall QUILT is collected from 4475 videos with a total runtime exceeding 1087 hours. From these videos, we extracted 437K unique images with 802K associated text pairs (one to many relationship). The descriptions vary in length and we have an average of 22.76 words per image. QUILT’s medical content is quite rich in terminology, featuring nearly 1.5M medical terms (as defined by UMLS), with about 28K unique terms. The visual data presents a diverse range of microscopic detail, captured at three primary magnification levels: low (0-10x), medium (10-20x), and high (20-40x), containing 280K, 75K, and 107K samples respectively. The high-quality images maintain substantial resolution, typically measuring 882 pixels in height by 1,468 pixels in width.

2.3.5 Combining QUILT with other histopathology data sources to obtain QUILT-1M

We enhanced QUILT by incorporating medical imagery and descriptions from three additional open-source data sources, creating an expanded collection called QUILT-1M.

PubMed Open Access Articles. First, we conducted an extensive analysis of medical literature published between 2010 and 2022 in PubMed’s freely accessible archive. This yielded nearly 60K matching pairs of microscopic tissue images and their descriptions. Our automated systems identified three distinct categories: standalone microscopic images, composite medical figures containing tissue samples, and individual tissue sections extracted from larger medical illustrations.

Histopathology Image Sourcing from LAION. The LAION-5B dataset [109] is a massive collection of over 5 billion image-text pairs gathered from diverse sources across the web, including a small amount of histopathology-related content, from which we identified and extracted \approx 23K pairs of images and associated text.

Twitter Dataset from OpenPath. Leveraging a dataset of tweets compiled by [49], we accessed a total of 55K unique tweets, which corresponded to 134K distinct image-text pairs. This dataset introduced a one-to-many mapping, where individual images were associated

with multiple captions. For consistency, we adopted their text pre-processing protocol outlined in [49].

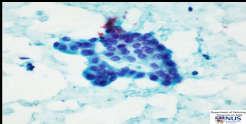
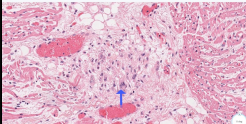
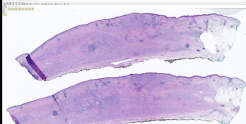
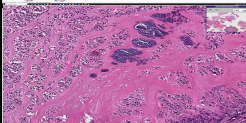
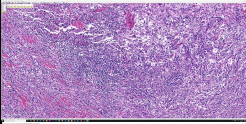
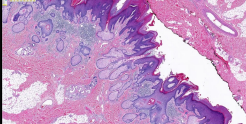
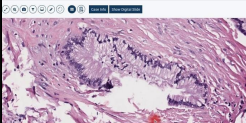
Image	Medical TEXT	ROI Text	Sub-pathology Classification
	['There are clusters of cells with micro-follicular formations.', 'Nuclear pseudo-inclusions, oval nuclei, nuclear grooves, and small nucleoli are present in some cells.']	['clusters of cells', 'micro-follicular formations', 'nuclear pseudo-inclusions', '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 2.2: 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.

2.4 Training QUILTNET with QUILT-1M

We leverage the Contrastive Language-Image Pre-training (CLIP) objective [98] to pretrain QUILTNET on QUILT-1M. CLIP operates on a batch of N (image, text) pairs and optimizes a contrastive objective that creates a joint embedding space. The training process involves optimizing the InfoNCE loss, where the image and text encoders are maximizing the cosine similarity of embeddings for matching pairs while minimizing it for mismatched pairs. Mathematically, InfoNCE loss is defined as follows:

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

Here, I_i and T_i represent the embeddings for the i -th aligned image and text, respectively. The image encoder utilizes ViT-B/32 and ViT-B/16 architectures [22], while the text encoder employs GPT-2 [100] with a context length of 77 (as an ablation we also tried PubMedBert [137]). We train QUILTNET by fine-tuning an OpenAI-pretrained CLIP model [98] on QUILT-1M. After fine-tuning, we evaluate its performance on two downstream tasks: image classification (using both zero-shot and linear probing) and cross-modal retrieval (zero-shot).

Histopathology Datasets for Downstream Tasks. We assess the performance of QUILTNET using 13 distinct datasets: **PatchCamelyon** [121] includes lymph node histology scans, categorized with binary labels indicating the presence or absence of metastatic tissue. **NCT-CRC-HE-100K** [64] comprises colorectal cancer images, divided into normal and cancerous. **SICAPv2** [114] provides images labeled across categories ranging from non-cancerous to Grades 3 to 5. **Databiox** [7] contains cases of invasive ductal carcinoma classified into Grades I-III. **BACH** [2] features breast tissue samples labeled as normal, benign, in-situ carcinoma, and invasive carcinoma. **Osteo** [3] includes tissue patches showcasing the variability within osteosarcoma samples. **RenalCell** [9] consists of images from clear-cell renal cell carcinoma, annotated across five different tissue texture types. **SkinCancer** [67] provides tissue patches derived from skin biopsies, representing 12 anatomical regions and 4 types of neoplasms, forming the **SkinTumor** subset. **MHIST** [126] offers tissue patches

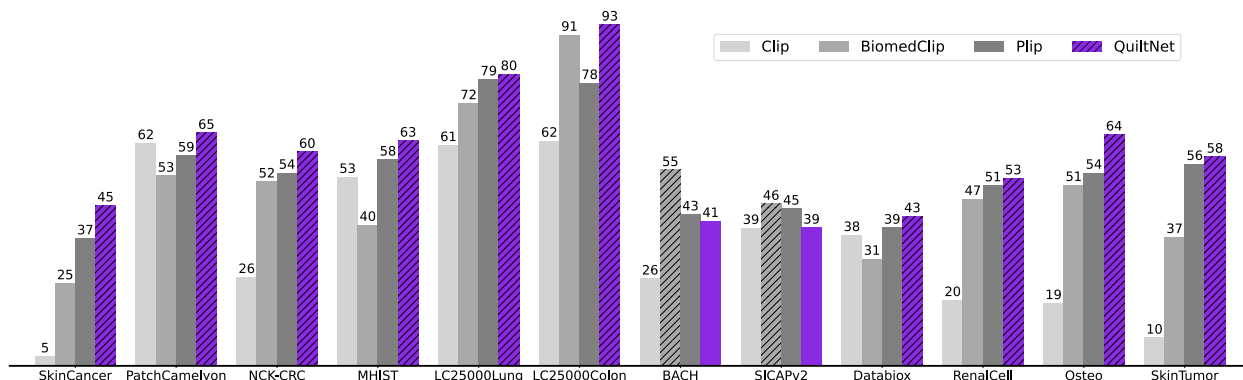


Figure 2.3: QUILTNET demonstrates superior performance compared to the out-of-domain CLIP baseline and state-of-the-art histopathology models in 12 zero-shot tasks across 8 distinct sub-pathologies. Accuracy percentages are shown for reference.

extracted from Formalin-Fixed Paraffin-Embedded Whole Slide Images (WSIs) of colorectal polyps. **LC25000** [8] is divided into two subsets: **LC25000 (Lung)** and **LC25000 (Colon)**, containing images of lung and colon adenocarcinomas, respectively.

Zero-Shot Learning Results. The diverse nature of cancer sub-types in histopathology necessitates models that can generalize effectively without retraining on specific datasets. To assess this capability, we evaluate QUILTNET’s zero-shot performance in comparison with three state-of-the-art models: CLIP, BiomedCLIP, and PLIP. As shown in Figure 2.3, our model achieves the best performance across all but two datasets, where BiomedCLIP shows slightly better results.

Linear Probing Results. We evaluate our model’s performance in few-shot and full-shot scenarios by conducting linear probing with 1%, 10%, and 100% of the training data. Sampling is performed using three different seeds, and the average accuracy along with standard deviation is reported in Table 2.1. This evaluation is conducted on four datasets from our external collection, each of which includes separate training and testing sets.

Table 2.1: **Linear probing.** Classification accuracy results (with standard deviation). PatchCamelyon is referred to as Camelyon. Supervised benchmarks represent SOTA models for each dataset.

Dataset	%shot	ViT-B/32			ViT-B/16			
		CLIP	PLIP	QUILTNET	CLIP	QUILTNET	BiomedCLIP	QUILTNET
Supervised(%)		GPT/77	GPT/77	GPT/77	GPT/77	GPT/77	PMB/256	PMB/256
NCT-CRC [64]	1	91.0 (0.1)	93.75 (0.1)	94.64 (0.2)	90.96 (0.1)	93.36 (0.2)	92.14 (0.1)	93.55 (0.2)
	10	92.02 (1.3)	93.83 (0.1)	95.30 (0.1)	92.58 (0.1)	93.85 (0.1)	92.90 (0.1)	93.72 (0.1)
	(94.0)	100	91.83 (0.1)	94.16 (0.1)	95.22 (0.1)	92.26 (0.1)	93.76 (0.1)	92.97 (0.1)
Camelyon [121]	1	80.38 (0.2)	87.26 (0.2)	87.62 (0.4)	80.28 (0.2)	84.78 (0.2)	83.63 (0.4)	83.48 (0.2)
	10	82.67 (0.2)	87.48 (0.1)	87.55 (0.1)	82.20 (0.1)	86.77 (0.1)	84.18 (0.2)	84.42 (0.1)
	(97.5)	100	82.80 (0.1)	87.34 (0.1)	87.48 (0.1)	82.55 (0.1)	86.81 (0.1)	84.23 (0.1)
SkinCancer [67]	1	84.27 (0.2)	91.07 (0.3)	90.93 (0.3)	85.62 (0.2)	88.29 (0.1)	87.53 (0.2)	88.06 (0.2)
	10	89.0 (0.1)	93.39 (0.1)	92.99 (0.1)	90.28 (0.1)	91.20 (0.0)	89.23 (0.0)	90.03 (0.1)
	(93.3)	100	89.02 (0.0)	93.29 (0.0)	93.03 (0.0)	90.29 (0.0)	91.20 (0.0)	89.16 (0.0)
SICAPv2 [114]	1	52.45 (2.4)	65.76 (2.6)	69.92 (1.0)	56.01 (0.6)	66.86 (1.2)	69.43 (1.0)	68.49 (1.1)
	10	62.24 (0.6)	69.23 (0.4)	74.14 (0.4)	63.70 (0.7)	72.37 (0.7)	71.61 (0.3)	72.48 (0.4)
	(67.0)	100	65.75 (0.2)	73.0 (0.1)	75.48 (0.1)	68.74 (0.1)	74.14 (0.2)	74.57 (0.1)

Notably, our model, built on the ViT-B/32 architecture with GPT/77, achieves superior results compared to BiomedCLIP, PLIP, and CLIP on the majority of datasets. On the NCT-CRC and SICAPv2 datasets, our model exceeds even the fully supervised performance while utilizing only 1% of the labels. Interestingly, in some cases, the 10% subset outperforms the 100% subset. We argue, this is because the smaller subset has better class balance, as each class is represented equally whereas on 100% case some classes are overrepresented

compared to others.

Cross-Modal Retrieval Results. We assess the effectiveness of cross-modal retrieval by evaluating both zero-shot text-to-image and image-to-text retrieval tasks. This is achieved by identifying the nearest neighbors for each modality and checking if the corresponding pair appears within the top N nearest neighbors, where $N \in \{1, 50, 200\}$. These experiments are performed on two datasets: our holdout subset from QUILT-1M and the ARCH dataset. The retrieval results are presented in Table 2.2.

Table 2.2: Cross-modal retrieval performance is reported for the QUILT-1M holdout set and the ARCH dataset. Results are presented in the format (%/%), with the QUILT-1M holdout scores shown on the left and ARCH scores on the right. Best is bold

		Text-to-Image (%)			Image-to-Text (%)		
model	config	R@1	R@50	R@200	R@1	R@50	R@200
CLIP	ViT-B/32 GPT/77	0.5/0.1	4.7/2.4	10.1/7.2	0.4/0.1	4.0/2.5	8.8/7.2
PLIP	ViT-B/32 GPT/77	1.1/0.6	10.8/13.1	21.8/29.9	0.9/0.7	11.0/13.7	21.6/29.5
QUILTNET	ViT-B/32 GPT/77	1.2/1.4	16.3/19.9	32.0/39.1	1.2/1.3	14.9/19.2	28.9/38.5
CLIP	ViT-B/16 GPT/77	0.8/0.1	5.6/2.7	11.2/8.7	0.7/0.1	5.0/3.1	10.8/9.0
QUILTNET	ViT-B/16 GPT/77	2.4/1.3	22.4/20.3	41.0/40.9	2.0/1.0	21.7/16.2	39.3/34.2
BiomedCLIP	ViT-B/16 PMB/256	4.3/ 8.9	15.0/53.2	25.6/71.4	3.9/ 10.0	14.0/52.1	23.5/68.5
QUILTNET	ViT-B/16 PMB/256	6.2/8.8	30.3/55.1	50.6/77.6	6.3/9.8	31.1/53.1	50.8/73.4

2.5 Conclusion

This chapter presented QUILT-1M, the largest publicly available histopathology dataset to date. Our experiments demonstrate the effectiveness of pre-training on QUILT, with QUILTNET surpassing larger state-of-the-art models such as BiomedCLIP across multiple

sub-pathology categories and tasks, including zero-shot, few-shot, full-shot, and cross-modal retrieval. We set new benchmarks in zero-shot learning, linear probing, and cross-modal retrieval tasks within the field of histopathology. To support the research community, we have made our data, code, and models available at <https://quilt1m.github.io/>.

Limitations. While our results are encouraging, the curation of QUILT relied on a combination of handcrafted algorithms and LLMs, which can introduce biases and errors. For example, our histopathology classifier occasionally produced false positives ($\approx 5\%$), as verified through human evaluation. Additionally, ASR systems sometimes misinterpreted medical terms. Regarding societal biases, a substantial proportion of narrators in our dataset are affiliated with western institutions, a bias compounded by our focus on English-language videos. As a result, QUILTNET may exhibit preferential performance on data representing these demographics, potentially leading to reduced effectiveness when applied to culturally or linguistically diverse groups. Finally, adhering to YouTube’s privacy policy, we provide only video IDs, enabling individuals to opt out of our dataset if desired.

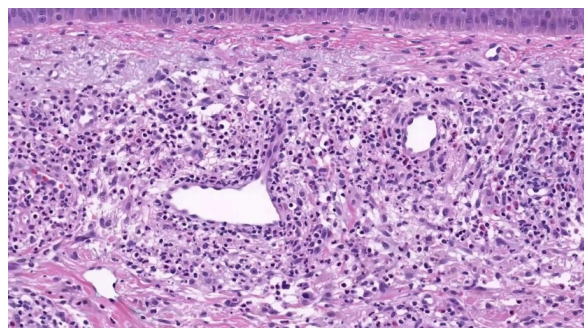
Chapter 3

LARGE LANGUAGE AND VISION ASSISTANT FOR HISTOPATHOLOGY - QUILT-LLAVA

3.1 Introduction and Motivation

In histopathology, interpreting giga-pixel-level whole slide images (WSIs) demands a comprehensive approach that goes beyond analyzing individual patches within the WSI. Pathologists typically examine these WSIs holistically, collecting evidence from multiple regions to form a diagnosis. As they navigate through WSI regions, pathologists iteratively reason, deliberating over morphological features within the regions to either support or rule out specific diagnoses, resulting in a process that is both dynamic and context-dependent. For instance, in liver tissue WSIs, the presence of steatosis (fat accumulation) alone may not indicate significant pathology. However, when accompanied by hepatocyte ballooning and lobular inflammation, it strongly suggests non-alcoholic steatohepatitis, a severe condition that can progress to cirrhosis or liver cancer [38]. This diagnostic approach relies on understanding the relevance of histological visual features, their spatial relationships across the WSI [10]. While current multi-modal models in histopathology are effective at analyzing isolated image patches, they often lack the capability to perform reasoning that extends beyond those patches to guide subsequent actions, thereby limiting their applicability.

In natural images, the success of the Large Language and Vision Assistant (LLaVA) [72] highlights the potential of multi-modal models as conversational partners. LLaVA is trained using instruction-tuning data generated by Large Language Models (LLMs), such as GPT-4 [94]. Additionally, natural images often feature multiple objects that can be annotated with bounding boxes, enabling the creation of spatially grounded instruction-tuning datasets [72]. Recently, multi-modal models have expanded to incorporate video content, allowing for



User: Can you provide a concise description of the histopathology image shown?



Quilt-LLaVA: Upon examination of the histopathology image, it is evident that there is a significant presence of lymphocytes. These cells are densely packed and exhibit a high degree of pleomorphism, indicating a possible pathological condition. The lymphocytes are irregular in shape and size, with some displaying enlarged nuclei and prominent nucleoli. In addition to the lymphocytes, there is also a noticeable presence of eosinophils. These cells are characterized by their bright pink cytoplasm and bilobed nuclei. The eosinophils are scattered throughout the tissue, suggesting an ongoing immune response. The overall cellular architecture appears disrupted, with the normal tissue architecture replaced by a dense infiltrate of lymphocytes and eosinophils. This could be indicative of an inflammatory or immune-mediated process. Further diagnostic tests may be required to confirm the exact nature of this pathological condition.



User: Given the presence of a mixture of lymphocytes and eosinophils, what can be deduced from this single image regarding a possible diagnosis?



Quilt-LLaVA: The presence of a mixture of lymphocytes and eosinophils in this patch suggests an inflammatory response. This could be seen in a variety of conditions, including infections, autoimmune diseases, and certain types of skin lesions. However, without additional context, it's challenging to make a definitive diagnosis based solely on this feature. It's important to note that the presence of these cells alone does not necessarily indicate a malignant condition. In fact, these cells can be seen in a variety of benign conditions as well. Therefore, a comprehensive evaluation of the whole slide image, including the overall architecture of the lesion, the presence of a junctional and dermal component, and the absence of atypia or mitotic activity, would be necessary to make a more accurate diagnosis.

Figure 3.1: QUILT-LLaVA can describe key medical regions in a histopathology patch and also reason through the observations to support a diagnosis. Note: The image features eosinophils and lymphocytes and is taken from a WSI depicting rare benign dermatitis.

more advanced reasoning capabilities beyond single images [70]. Despite these advancements, similar progress has not yet been achieved in histopathology. Current multi-modal models for histopathology frequently depend on PubMed articles to generate image-caption pairs [72, 87]. However, these articles lack three key components: **1)** a sufficiently large pool of papers to curate a large-scale, diverse image-caption dataset specific to histopathology, **2)** visually grounded captions for connecting visual concepts in histopathology, and **3)** comprehensive contextual information from the entire WSI, which is crucial for holistic reasoning capability. These limitations restrict the effectiveness of current PubMed-trained models in supporting detailed histopathological analysis.

To address this gap, in the previous Chapter, **1)** we introduced QUILT [53], the largest image-caption dataset in histopathology, comprising 723k image-caption pairs sourced from educational YouTube videos—making it five times larger than the next largest dataset [49]. In this Chapter, leveraging QUILT videos, we generated QUILT-INSTRUCT [110], an instruction-tuning dataset containing 107,131 histopathology-specific question/answer pairs. While generating QUILT-INSTRUCT, we addressed the previously identified challenges: **2)** to improve spatial awareness, we extracted narrators’ mouse cursor movements from the videos, enabling spatial grounding of histopathology concepts within each patch through spatio-temporal clustering. **3)** to provide reasoning, we developed novel instruction-tuning QA prompting techniques tailored for histopathology, including complex reasoning and iterative abductive reasoning, which integrate the global WSI diagnosis and its supporting evidence with image captions, ensuring factual grounding and minimizing hallucinations (see Fig3.2).

Using QUILT and QUILT-INSTRUCT, we trained QUILT-LLAVA [110], a multi-modal model tailored for histopathology. QUILT-LLAVA excels at describing histopathology patches, reasoning toward diagnoses, and spatially localizing critical histopathology concepts, as shown in Fig3.1. The training of QUILT-LLAVA followed a two-stage process (see Fig 3.5). First, the model was aligned with the histopathology domain by leveraging 723K image-text pairs from QUILT [53], and then it underwent further instruction-tuning using QUILT-INSTRUCT. QUILT-LLAVA provides detailed image analysis, localizes medical concepts spatially, and

goes beyond isolated image patches by guiding users on the additional evidence required to confirm or rule out specific diagnoses. Furthermore, it can serve as an educational tool by offering subtle diagnostic hints rather than directly providing a diagnosis, making it an effective aid for pathology student training (See Section A.0.1 in the Appendix).

Lastly, educational videos are highly valuable due to their interactive nature, where narrators frequently engage viewers by posing and answering questions during their presentations. Building on this, we introduce QUILT-VQA, an evaluation dataset organically derived from naturally occurring questions and answers within QUILT videos, created with the assistance of GPT-4 and manual verification. We evaluated QUILT-LLAVA using QUILT-VQA and two public histology VQA test sets, covering both open and closed questions. QUILT-LLAVA achieved state-of-the-art performance, surpassing the previous best by 4% on open questions and 9% on closed questions. To further assess QUILT-LLAVA’s reasoning abilities, we utilized GPT-4 to compare its responses with those generated by two other state-of-the-art multi-modal models: LLAVA [77] and LLAVA-MED [72]. QUILT-LLAVA outperformed LLAVA and LLAVA-MED by over 16% and 7%, respectively.

3.2 Related work

Visual instruction-tuning in natural images. The open-source availability of LLMs [120, 19, 59, 118] has facilitated advancements in the vision-language domain by enabling the training of multi-modal models that leverage implicit visual understanding through the generation of instruction-tuning datasets from image captions. Previous work [78, 144, 32] has showcased impressive results, with LLaVA-1.5 achieving performance comparable to GPT-4 in certain multi-modal tasks [77].

Visual instruction-tuning in medical images. Visual Med-Alpaca [44] generated 54K question-answer pairs for instruction-tuning using GPT-3.5, while PMC-VQA [139] compiled a larger multiple-choice dataset from general medical domains using PubMed. However, its coverage of histopathology remains limited. Similarly, LLaVA-Med [72] leverages image captions from PubMed articles, supplementing them with sentences from the article in cases

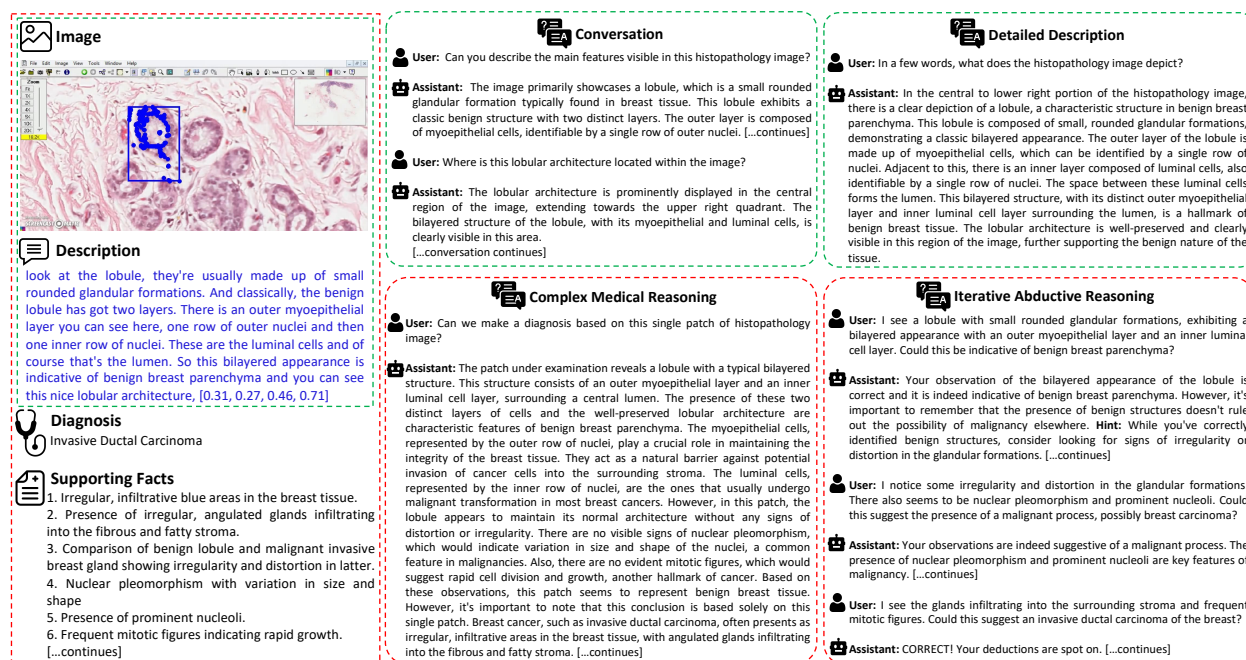


Figure 3.2: A sample from QUILT-INSTRUCT. Only Complex Medical Reasoning and Iterative Abductive Reasoning prompts implicitly utilize Diagnosis and Supporting Facts (which are extracted from the entire video). Note that some of the content is truncated due to space limitations. Please see Figs. A.13-A.14-A.15 in the Appendix for a more complete overview of QUILT-INSTRUCT [110].

of short captions, which may not always directly relate to the referenced figure. Within LLaVA-Med, 17K images are of histology, resulting in 49K question-answer pairs. These approaches predominantly rely on isolated image-caption pairs to construct instruction-tuning datasets, which restricts GPT-4’s ability to reason beyond its immediate context and increases the risk of hallucination when attempting to do so.

Video-based image-text datasets. Several studies have focused on curating datasets from videos in the natural domain [135, 134, 123, 70, 125]. For example, MIMIC-IT [70] utilizes video data to generate various prompts, including reasoning-based ones, to create

instruction-tuning datasets. Similarly, Video Localized Narratives [123] involve annotators verbally describing images while simultaneously moving their mouse cursors over the corresponding regions, producing dense visual groundings. However, replicating such approaches in histopathology is costly. PathNarratives [136], for instance, enlisted eight pathologists to manually annotate spatially grounded pathology data, but the dataset is not publicly available. More recently, we introduced QUILT [53], which collects large-scale data from educational histopathology videos on YouTube to create image-caption pairs. While QUILT is extensive, it lacks spatially grounded captions. Building on QUILT’s video content, we generate spatially grounded histopathology data. To the best of our knowledge, this is the first effort to use videos for creating spatially grounded instruction-tuning datasets in the medical domain.

3.3 Curating QUILT-INSTRUCT

To create a robust instruction-tuning dataset for histopathology with visual groundings in WSIs, we leveraged the detailed narrative content found in educational YouTube videos.

3.3.1 Data Preparation

We observed that in educational videos, narrators frequently pause while navigating large-scale WSIs before highlighting critical areas using their cursor [63, 123, 57, 96]. Our workflow involves three key steps to transform these unstructured videos into visually-grounded instruction data:

1. **Cursor Localization:** Identify and track the narrators’ cursors throughout the video.
2. **Spatio-Temporal Clustering:** Cluster cursor movements over time and space to visually anchor histopathology concepts within specific image regions, as illustrated in Fig. 3.3.

3. Instruction-Tuning Dataset Generation: Use the extracted grounded captions to prompt an LLM for generating QUILT-INSTRUCT. This involves diverse prompting strategies, ranging from independent prompts that create varied Q/A pairs for individual image patches to reasoning-based prompts that synthesize information across patches for diagnostic Q/A pairs.

Data source. We constructed QUILT-INSTRUCT from QUILT [53], which comprises 4,149 educational YouTube videos totaling over 1,000 hours of content. The videos follow a *narrative style*, where pathologists articulate histopathological concepts while interacting with WSIs. Although QUILT provides extensive image-text data, it lacks spatial annotations that link text descriptions to specific regions in the images—a critical component often missing in most medical datasets. Fortunately, the narrators’ mouse cursors naturally highlight morphological features, providing the visual grounding necessary to bridge this gap. We now detail the process of extracting these cursors from raw video footage.

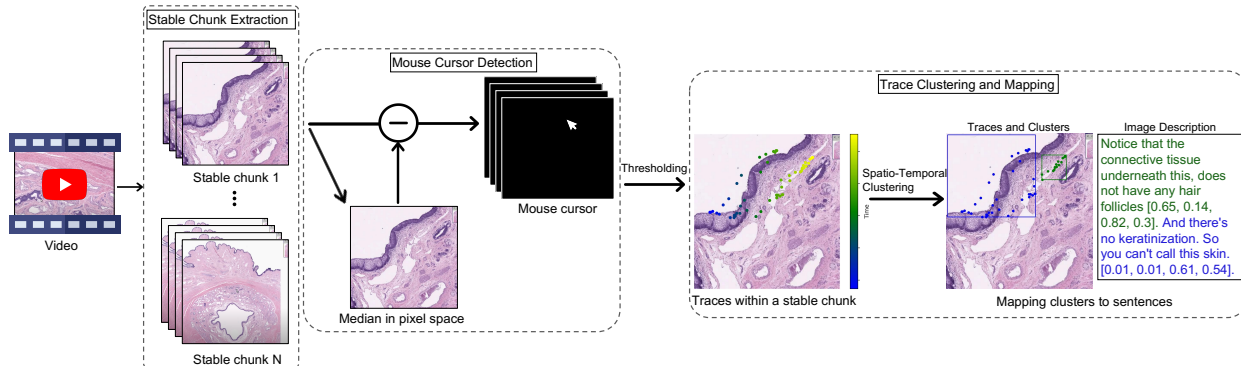


Figure 3.3: The QUILT-INSTRUCT dataset was created by first segmenting videos into stable chunks. For each chunk, the median frame was computed in the pixel domain and subtracted from all frames within that chunk. These cursor points were then clustered in time and space to spatially ground medical content. In the "Trace Clustering and Mapping" section of the figure, time progression is represented by color.

Localizing Cursors. Extracting cursor locations from histopathology images presents a notable challenge due to the uniformity in color and texture, coupled with dynamic elements such as minor pixel variations and differences in cursor movement speed and style across videos. However, a key observation simplifies this task: narrators often pause their navigation of WSIs before gesturing with the cursor to direct the audience’s attention. This behavior suggests that if a stable background occurs in consecutive frames, it is highly likely that the narrator may be using their cursor to mark medically relevant regions. Leveraging this insight, we developed an efficient method to detect mouse cursors. First, we identify segments in each video where the background remains largely static, referred to as stable chunks (illustrated in the leftmost section of Fig. 3.3). To detect these stable chunks, we employ a simple frame-differencing approach. The algorithm calculates absolute differences between consecutive frames and applies a Gaussian filter with adaptive thresholding to identify frames with minimal background movement. While histopathology images typically feature uniform backgrounds due to dye-based staining, this pixel-wise difference approach often produces false positives by misclassifying chunks with minor changes as stable. To address this, we integrate a perceptual metric, the Structural Similarity Index Measure (SSIM), to verify changes between frames. By evaluating SSIM on randomly sampled patches, we refine the detection of stable chunks with improved reliability. For each stable chunk, we calculate a median frame in the pixel domain to represent the background. This median frame is subtracted from the individual frames within the chunk, followed by thresholding to mitigate noise and artifacts, effectively isolating the cursor (as shown in the middle section of Fig. 3.3). The cursor location is then identified by extracting the coordinates of the maximum pixel value. This approach assumes minimal interference from other subtle pixel changes, such as movements caused by the narrator’s facial expressions within the scene. To counter this, we apply a face detection model [108] to mask facial regions, ensuring the algorithm focuses solely on cursor movement. The result is a computationally efficient and highly generalizable method for extracting cursor traces from educational videos with similar *narrative styling*.

After identifying the candidate stable chunks, we retained those with a minimum duration

of three seconds. To ensure only histological content is processed, we applied a histology classifier as described in [53], using the background frame of each stable chunk to filter out non-histological segments. This approach yielded 132K image-caption pairs, denoted as $\langle I_j^v, c_j^v \rangle$, where I_j^v represents the median frame of the j^{th} stable chunk within the v^{th} video, and c_j^v is the corresponding caption for I_j^v , derived by transcribing the narrator’s speech to text, following the methodology outlined in [53]. On average, each pair spanned 24 seconds of video and contained approximately 55 words in its caption. From these 132K image-caption pairs, 44,163 exhibited active cursor movement, where the cursor’s position is preserved as (x_j^t, y_j^t) , with $x \in [0, W]$ and $y \in [0, H]$, where W and H denote the image width and height, respectively, and t spans from 0 to the total duration of the j^{th} stable chunk. Next, we outline our approach for clustering the extracted cursor positions to visually ground the captions.

Spatio-temporal clustering of the mouse cursor. To spatially ground narrators’ words, we clustered the tracer points (x_j^t, y_j^t) for each stable chunk. First, we transform the inputs into (x_j^t, y_j^t, t, w_j^t) , where t represents time and w_j^t denotes the cumulative word count at each cursor position (to encode the pace of speech better). To balance the spatial and temporal aspects in cluster formation, we applied an exponential decay to the spatial coordinates, $e^{-\lambda t}$, with $\lambda = 0.05$, prioritizing the clustering of points closer in time. The number of clusters was dynamically selected based on the word count in c_j^v . To refine the assignment of words to clusters, we first determined the temporal center of each cluster by calculating the mean timestamp of the tracer points within it. This "temporal midpoint" served as a reference for assigning words to clusters based on their temporal proximity. Each word was mapped to the cluster whose temporal midpoint was closest to the word’s occurrence time. The resulting clusters are represented as bounding boxes in the format $[x_1, y_1, x_2, y_2]$, where the coordinates are normalized between 0 and 1. These coordinates define the top-left (x_1, y_1) and bottom-right (x_2, y_2) corners of the bounding boxes. An example of this process is illustrated in Fig. 3.3. Next, we demonstrate how these grounded captions are used to generate our instruction-tuning dataset, QUILT-INSTRUCT.

3.3.2 Generating QUILT-INSTRUCT

We generated two distinct sets of question-answer types. The first set utilizes *Independent Prompts*, inspired by the Conversations and Detailed Descriptions prompts in [78]. These prompts take single patch-level grounded text as input to generate Q/A pairs that are constrained to the independent context of each input sample, i.e., they do not aim to have a WSI-level reasoning. The second set introduces *Reasoning-Based Prompts*, which leverage the contextual continuity present in our dataset. Specifically, for videos that review a single WSI, we exploit the sequential development of concepts and diagnostic clues toward a final conclusion. These prompts use both patch-level grounded text and global WSI-level information as input. This approach enables an LLM (GPT-4) to extend its reasoning beyond the immediate context of individual patches, while remaining anchored in factual insights extracted from the entire video. By doing so, these prompts effectively reduce hallucinative behavior and enhance the logical coherence of generated responses.

Independent Prompts

Following the methodology outlined in [78], we generated Q/A pairs from each image caption c_j^v , aiming to create an array of questions that describe the context presented in the image. These prompts were intentionally designed to elicit answers based only on the information within the caption, avoiding any extrapolation beyond the immediate input.

Conversation. The purpose of the Conversation prompt is to simulate a dialogue between an AI assistant and an individual inquiring about a histopathology image. Using the image caption c_j^v , which includes bounding boxes of medical concepts, we generated diverse Q/A pairs. The AI assistant, equipped with the provided descriptions and bounding box coordinates, responds as though directly analyzing the image. These responses address aspects such as cell types, activities, and the spatial relationships of medical concepts within the image. Each dialogue typically consists of 3–4 Q/A pairs, maintaining a conversational tone while focusing on the visual and contextual content of the image.

Detailed Descriptions. In line with [78], GPT-4 is instructed to use c_j^v to produce a detailed description of the given image. Subsequently, questions are randomly selected from a pre-compiled list given in Fig. 3.4.

Reasoning-based Prompts

Most image-caption pairs used in instruction-tuning datasets, such as those in PubMed or COCO [76], are contextually isolated, meaning $\langle I_i, c_i \rangle$ and $\langle I_j, c_j \rangle$ where $i \neq j$ lack shared context. Consequently, Q/A pairs generated by GPT-4 are constrained to the context of a single image. However, in the case of gigapixel WSIs, deriving a comprehensive diagnosis from a single image patch is rarely sufficient. Ideally, the model should reason beyond the given image, guiding the user on subsequent areas to examine or next steps to take. While GPT-4’s inherent medical knowledge may assist in such reasoning, prompting it to extrapolate without interconnected context often results in hallucinations, thereby compromising the quality of the generated data.

In contrast, our approach leverages video-derived image-caption pairs that cover WSIs. To this end, we manually reviewed the content of QUILT identifying 2,066 videos that focus exclusively on a single WSI for reasoning-based prompts. Using the complete text from these videos, GPT-4 was prompted to infer the final diagnosis d_j^v and the supporting facts f_j^v that lead up to that diagnosis. These supporting facts and diagnosis provide the foundation for reasoning-based prompts, encouraging GPT-4 to engage in extrapolative reasoning. This approach fosters extrapolations that remain grounded in context, significantly reducing the tendency for hallucination. We designed two distinct prompt types, **Complex Medical Reasoning** and **Iterative Abductive Reasoning**, to generate Q/A pairs for reasoning-based instruction-tuning datasets.

Complex Medical Reasoning. Given a caption c_j^v , along with a diagnosis d_j^v and supporting facts f_j^v , we prompt GPT-4 with a diagnostic reasoning task designed to extrapolate beyond the immediate context of c_j^v . Specifically, GPT-4 is instructed to utilize its inherent medical knowledge to interpret the contents of c_j^v , while subtly integrating the diagnosis and supporting

facts derived from the entire video. If the observations from c_j^v are sufficient to make a diagnosis using GPT-4’s general medical knowledge, it provides the diagnosis along with the reasoning behind it. If the information is insufficient, GPT-4 extrapolates using f_j^v to propose additional features or regions within the WSI that should be examined to validate or rule out specific diagnoses. We term this approach "Subconscious Knowledge Injection," as it enables GPT-4 to leverage its broader medical expertise while being gently guided by the contextual constraints of d_j^v and f_j^v . This methodology ensures that reasoning remains focused and contextually grounded, significantly reducing the likelihood of hallucinations.

Iterative Abductive Reasoning. We simulated a conversation between two GPT-4 agents to mimic a scenario where a professional pathologist uses the model to ask medically intricate and extended questions about an image. Unlike shorter prompts, this approach emphasizes longer, contextually rich exchanges. The first agent, referred to as Human-GPT, receives an image patch caption c_j^v and is tasked with abductively reasoning about potential diagnoses and the supporting facts. This reasoning is presented in the format: *User: [{Abduction: xxx}, {Facts Used: xxx}]*. The second agent, termed AI Assistant GPT, is provided with the diagnosis d_j^v and supporting facts f_j^v , simulating an individual who has access to the entire WSI for the patient. The AI Assistant evaluates the accuracy of the abduction proposed by Human-GPT, providing comments or hints on potentially overlooked details using its medical expertise and the information from d_j^v and f_j^v . This feedback is presented in the format: *GPT: [{Comments: xxx}, {Hint: xxx}]*. If Human-GPT exhausts all available information from c_j^v without arriving at a diagnosis, the AI Assistant suggests examining additional patches for further evidence. For example, it might propose: *"Consider looking for evidence of X in other patches to validate your diagnosis."*

By leveraging Iterative Abductive Reasoning-based instruction-tuning data, QUILT-LLAVA can also function as an educational tool, assisting human users in brainstorming diagnostic steps by providing hints without explicitly revealing the answer. (Refer to Fig. A.2 in the Appendix.) The conversation between AI Assistant GPT and Human-GPT proceeds iteratively, with a randomly determined upper limit of two to four exchanges per agent. Each

new exchange incorporates the post-exchange history, enabling GPT to retain memory of the ongoing conversation. The dialogue may terminate earlier if Human-GPT arrives at a conclusive diagnosis or if the AI Assistant determines that Human-GPT has fully utilized all relevant information from c_j^v , subsequently guiding it on the next steps using f_j^v .

Dataset Statistics. From QUILT, we extracted 162,566 image-caption pairs. To refine this data, captions containing fewer than 20 words or exceeding 150 words were filtered out, resulting in a curated dataset of 114,343 pairs with an average caption length of 55 words. From this refined dataset, we generated QUILT-INSTRUCT, which includes 107,131 question/answer pairs. On average, the questions contain 16.5 words, while the answers are more detailed, averaging 101 words. For reasoning-based prompts, we manually reviewed 4,149 videos and selected 2,066 that focused exclusively on a single WSI from a single patient.

3.4 Training QUILT-LLAVA & evaluating with QUILT-VQA

In this section, we detail using QUILT-INSTRUCT to train QUILT-LLAVA. Next, we curate QUILT-VQA, to evaluate QUILT-LLAVA. Finally, we generate an Instruction Following Test Set from QUILT-VQA to assess QUILT-LLAVA’s ability to follow instructions.

3.4.1 Training QUILT-LLAVA

We adopt the LLAVA autoregressive model architecture for its simplicity and efficiency. This choice also ensures consistency in evaluation against baselines such as LLAVA [78] and LLaVA-MED [72], both of which employ a curriculum-learning strategy [72] for instruction tuning. LLAVA integrates a vision module, an LLM, and an MLP connector, enabling the LLM to process visual information. The training process begins with the MLP, which serves as a projector. During this initial phase, the MLP is trained until convergence while both the LLM and vision module remain frozen. Once the MLP is trained, the second stage involves fine-tuning both the MLP and LLM using instruction-following data to align the model’s outputs with those of human pathologists. Our overall architecture is shown in Fig. 3.5. While LLAVA typically utilizes a pre-trained CLIP image encoder, we employ a

domain-specific pre-trained CLIP model QUILTNET [53]. Additionally, we perform ablations with various image encoders, training strategies, and visual prompts to evaluate their impact on performance.

Histopathology Domain Alignment. To align our vision and language models within the histopathology domain, we extracted 723K image-text pairs from QUILT and converted their captions into Q/A format. This was achieved by randomly selecting an instruction (question) from a predefined list and prepending it to the caption (answer) to create Q/A pairs. These instructions are designed to variably describe the visual content of the images, ensuring diverse and contextually relevant Q/A data. For the questions, we adapted a list used in [71]. At this stage, we froze both the vision and language models, training only the MLP layer. The MLP projects the image encoder’s embeddings into a space compatible with the language model, enabling it to predict the image captions based on the provided questions. This step effectively aligns histology image embeddings with their corresponding text embeddings.

Histopathology instruction-tuning. Next, we fine-tuned our model using QUILT-INSTRUCT. At this stage, we kept the visual encoder weights frozen and continued to train the MLP layer and the LLM.

3.4.2 Evaluation Data Generation: QUILT-VQA

In histopathology, evaluation datasets such as PathVQA [46] and PMC-VQA [139] are commonly used to benchmark model performance. However, these datasets suffer from significant limitations, including excessive repetitiveness caused by paraphrasing the same question and, more critically, contradictory answers to identical questions. In contrast, educational videos offer a unique opportunity to address these issues. These videos often include an interactive element, with narrators posing questions during their presentations and subsequently providing the answers. For example, a narrator might ask, "*Do you know what kind of organ we’re dealing with?*" and then elaborate with, "*Yes, this is a colon.*" This natural Q/A format provides a rich and organic dataset, which can be extracted and

repurposed for evaluation, enabling more accurate and contextually relevant benchmarking.

To harness this potential, we mapped the occurrences of "?" in the video’s transcript to our stable chunks. If a "?" appeared within a 45-second time frame of a stable chunk, we expanded the stable chunk’s caption to include the complete sentence containing the question mark. This approach ensures that the questions are directly related to the visual content being displayed. Following data preprocessing and strategic mapping of question marks, we prompted GPT-4 to extract question-answer pairs directly from the text. For this, GPT-4 was provided with the stable chunk’s text, including any sentences ending in a question mark, which indicates a posed question. After GPT-4’s initial extraction, we conducted a manual verification to ensure that each question-answer pair in the dataset is both medically relevant and accurately corresponds to the content in the stable chunk’s text. Additionally, we categorized the questions into two groups: *Image-dependent* (1,055 Q/A pairs), which reference the narrator’s specific image, and *General-knowledge* (228 Q/A pairs), which pertain to broader medical understanding. This categorization enables the creation of evaluation sets that comprehensively test the model’s capabilities in both image analysis and medical reasoning.

3.4.3 Evaluation Data Generation: *Instruction Following Test Set*

In addition to QUILT-VQA, which focuses on assessing QUILT-LLAVA’s medical knowledge, we also aimed to evaluate the model’s capability to follow instructions in multi-modal conversations. For this purpose, we constructed a test set consisting of 326 questions, including 256 conversational questions and 70 detailed description questions, derived from image-text pairs extracted from unseen videos within QUILT-VQA. To generate this evaluation set, we utilized the same Conversation and Detailed Description prompts employed during the creation of QUILT-INSTRUCT. This ensured consistency in evaluation while testing the model’s ability to handle both conversational and descriptive instruction-following tasks.

- "Can you provide a concise description of the histopathology image shown?"
- "How would you summarize the features of the histopathology image displayed?"
- "What are the key observations in the presented histopathology image?"
- "Briefly outline the characteristics of the histopathology image in view."
- "In a few words, what does the histopathology image depict?"
- "What essential information emerges from the histopathology image provided?"
- "Could you distill the contents of the displayed histopathology image into a brief description?"
- "Offer a succinct report on the histopathological aspects evident in the image."
- "Identify and briefly discuss the components of the histopathology image."
- "How would you encapsulate the findings presented in the histopathology image?"
- "What insights can be concisely derived from the histopathology image on display?"

Figure 3.4: A set of instructions designed to provide varied and detailed descriptions of the visual content in images, used to create instruction-following QA pairs with comprehensive descriptions.

3.5 Experiments

This section evaluates the performance of QUILT-LLAVA against existing state-of-the-art multi-modal models on histopathology VQA benchmarks using various metrics. First, we assessed the alignment of output generations with ground truth answers using GPT-4. Second, we evaluated QUILT-LLAVA on both open-set and closed-set VQA tasks to measure its ability to generalize and perform under constrained conditions. Finally, we performed ablations, exploring the impact of visual promptings and different trained model configurations on overall performance.

Oracle (GPT-4) Alignment Evaluation. To evaluate the effectiveness of QUILT-LLAVA in multi-modal conversations, we used the Instruction Following test set (see Section 3.4.3) and leveraged GPT-4 (language-only model) to assess the *helpfulness, relevance, accuracy, and level of detail* responses generated by two assistants: the candidate model (QUILT-LLAVA) and GPT-4. GPT-4 assigns an overall score on a scale of 1 to 10 for each answer, where a higher score indicates better performance. Additionally, GPT-4 provides comprehensive

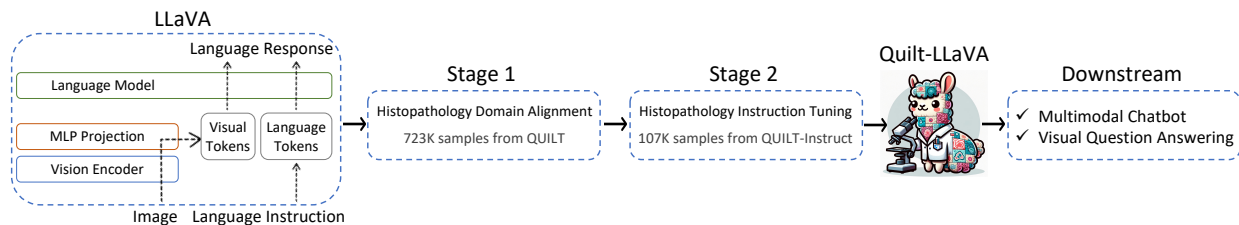


Figure 3.5: We initialized QUILT-LLAVA with the general-domain LLaVA and trained it for two stages: 1) Histopathology Domain Alignment on QUILT and 2) Instruction-tuning on QUILT-INSTRUCT. We evaluated QUILT-LLAVA on visual conversation and question answering tasks.

explanations for its evaluations, offering insights into the strengths and weaknesses of each answer. Table 3.1a presents the results across 14 sub-pathologies and 2 QA types. All QUILT-LLAVA variants outperformed the baselines, with QUILT-LLAVA trained on a single epoch of stage-1 pretraining and a balanced subset of 40K instruction-tuning pairs for stage-2 fine-tuning outperforming LLaVA and LLaVA-MED by over 16% and 7%, respectively. By increasing the instruction-tuning dataset size and pretraining for three epochs, we achieved an additional improvement, with QUILT-LLAVA surpassing LLaVA-MED by 10.8%.

Visual Question Answering. We evaluated QUILT-LLAVA on three histopathology VQA datasets, detailed in Table A.2 in the Appendix. The datasets include a mix of open-ended and closed-ended Q/A pairs. For closed-set questions, accuracy was used as the primary metric to determine the proportion of correct responses generated by the model. In contrast, for open-set questions, we focused on recall to assess how frequently the model’s responses contain ground-truth tokens [72]. The results, presented in Table 3.1b, compare all variants of QUILT-LLAVA against the medical domain LLaVA-MED and the general domain LLaVA where all QUILT-LLAVA variants outperformed LLaVA. Notably, models initialized with the text encoder from Vicuna [19] paired with QUILTNET models exhibited superior performance on open-set questions, achieving an average improvement of 4% over the state-of-the-art. On

the other hand, models initialized with the text encoder from LLAVA excelled in closed-set questions, showing an average improvement of 9% over the SOTA. The performance of QUILT-LLAVA on closed-set questions highlights its instruction-following capabilities, particularly in binary *yes/no* and true multi-choice $\{A/B/C/D\}$ formats. In contrast, LLAVA-MED underperformed in multi-choice QA due to its limited ability to follow such instructions. QUILT-LLAVA outperformed both the PathVQA baselines and QUILT-VQA for the closed-set evaluation, although the performance on the PMC-VQA-Subset was limited, which we attribute to the issues inherent in PMC-VQA. For open-set evaluation, QUILT-LLAVA demonstrated superior performance on PathVQA and QUILT-VQA, achieving significant margins over the baselines.

(Question Count)	Question Types		Sub-Domains														Overall
	Conv	Desc	Bone	Breast	Cyto	Derm	Endo	Gastro	Bone	Geni	Gyne	H&N	Neuro	Pulm	Renal	Soft	
	(256)	(70)	(25)	(23)	(23)	(21)	(23)	(23)	(23)	(22)	(24)	(22)	(24)	(25)	(23)	(25)	(326)
LLAVA [77]	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 [72]	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) Instruction-following abilities measured by GPT-4. QUILT-LLAVA outperforms the baselines.

Instruct	QUILT-LLAVA Model Variants		PathVQA		PMC-VQA-Subset	QUILT-VQA			
	Stage 1	Stage 2	Open	Closed	Closed	Open	Closed		
<i>QUILTNET ViT-B-32 Vision Encoder[53], 7B Language Model</i>									
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
<i>LLAVA [77] checkpoint, 7B Language Model</i>									
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
<i>Baselines</i>									
LLaVA-Med [72] 7B			0	0	11.97	56.15	1.34	54.81	61.22
LLaVA [77] 7B			0	0	11.65	54.02	33.91	55.81	57.73

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

QUILT-LLAVA Instruction Data				PathVQA		PMC-VQA-Subset	QUILT-VQA	
Conv	Detail	Complex	Abductive	Open	Closed	Closed	Open	Closed
10k	10k	0	0	12.56	50.99	12.04	57.52	64.14
0	0	10k	10k	17.53	44.50	30.46	73.95	39.65
5k	5k	5k	5k	14.04	57.12	22.13	61.75	65.60
10k	10k	10k	10k	14.10	57.33	22.17	61.42	65.60

(c) Finetuning QUILT-LLAVA on the instruction-tuning data subsets with varying number of samples.

Table 3.1: Quantitative results on histopathology VQA datasets. For open-set questions, we report recall in column *Open*. For closed-set, we report accuracy in *Closed*. Red is best.

Instruction-Tuning Ablation Studies. To investigate the impact of QUILT-INSTRUCT on performance, we trained several models: *I*—trained only on balanced independent prompts data (Section 3.3.2), *R*—trained only on balanced reasoning prompts data (Section 3.3.2), and *IR*—trained on a balanced combination of all prompt types. Our findings indicate that the *I* model performed better on closed-set questions, while the *R* model excelled in open-set questions. We attribute this to the contextual distribution of the questions and answers: *I* is well-suited for closed-set questions that directly ask option-constrained queries, whereas *R* demonstrates strength in open-set questions that often require reasoning to arrive at a diagnosis. Interestingly, the *IR* model improved performance on closed-set questions but exhibited a drop in open-set VQA performance. This suggests that while a balanced dataset combining all prompt types enhances certain capabilities, it may introduce trade-offs in tasks that require specialized reasoning.

3.6 Conclusion and Limitations

Limitations. Our dataset, derived from raw video footage, is not perfect. Noise within the data can lead to inaccuracies, such as incorrectly generated clusters or erroneous mouse cursor detections. Additionally, despite attempts to constrain its context, GPT-4 occasionally hallucinates, which may propagate to QUILT-LLAVA’s outputs. In some cases, GPT-4 also extracts information directly from captions rather than the image itself, contrary to explicit instructions.

Conclusion. We presented: **1)** QUILT-INSTRUCT, a dataset constructed by automatically extracting human narrators’ mouse movements to spatially ground concepts and leveraging WSIs to develop novel reasoning-based prompts; **2)** QUILT-LLAVA, a multi-modal model trained on QUILT-INSTRUCT, which surpasses existing baselines in both open-ended and closed-ended histopathology question answering; and **3)** QUILT-VQA, a curated evaluation dataset consisting of human-generated question-answer pairs derived from pedagogical videos,

designed to rigorously test QUILT-LLAVA’s reasoning capabilities. To support the research community, all resources, including code and datasets, are made available as open-source at <https://quilt-llava.github.io/>.

Chapter 4

AUTOMATING DIAGNOSTIC PROCESS FOR HISTOPATHOLOGY WITH MULTI-MODAL MULTI-AGENT FRAMEWORK - PATHFINDER

4.1 *Introduction and Motivation*

Histopathology diagnosis through whole slide image (WSI) analysis is fundamental to modern pathology. WSIs are high-resolution digital scans that provide comprehensive views of tissue architecture and cellular details. Pathologists navigate these massive gigapixel images to identify critical morphological features and spatial relationships essential for accurate diagnosis. Starting with low magnification to locate suspicious regions, they then zoom into specific patches for closer inspection [37, 79]. Evidence is gathered from multiple patches and synthesized into a holistic diagnosis. While this method remains the gold standard, it is highly labor-intensive, demands significant expertise, and is increasingly unsustainable given the growing global burden of cancer cases.

Adopting more efficient diagnostic approaches in medical imaging is crucial but must not compromise accuracy. Advances in deep learning have shown potential to achieve human expert level performance, offering a scalable solution [119]. However, most existing methods analyze WSIs by dividing them into smaller patches, often missing the holistic context necessary for comprehensive diagnosis [69, 133, 143, 110, 117, 1, 131, 53, 40]. While transformer-based models aim to integrate local and global patterns, they struggle to scale effectively for the high-resolution requirements of WSIs [111, 128, 42, 141, 13].

In this chapter, we propose PATHFINDER, a multi-modal and multi-agent system designed to mimic the decision-making process of expert pathologists for skin melanoma grading [36]. PATHFINDER integrates four AI agents: Triage, Navigation, Description, and Diagnosis. The

system begins with the Triage Agent, classifying the WSI as benign or risky. If risky, the Navigation and Description Agents iteratively examine patches, generating natural language descriptions and refining their focus with each cycle. Finally, these detailed insights are integrated by the Diagnosis Agent to produce an accurate and holistic diagnostic classification. Overall, the system is shown in Figure 4.1. Our experiments demonstrate that this model not only surpasses state-of-the-art AI methods on melanoma classification but also exceeds the average diagnostic performance of pathologists. To the best of our knowledge, PATHFINDER is the first AI-based system capable of surpassing the average performance of pathologists on this challenging melanoma classification task. It is also fully explainable, from the patches visited to the description of the patches and the final diagnoses, incorporating all the patch-wise information.

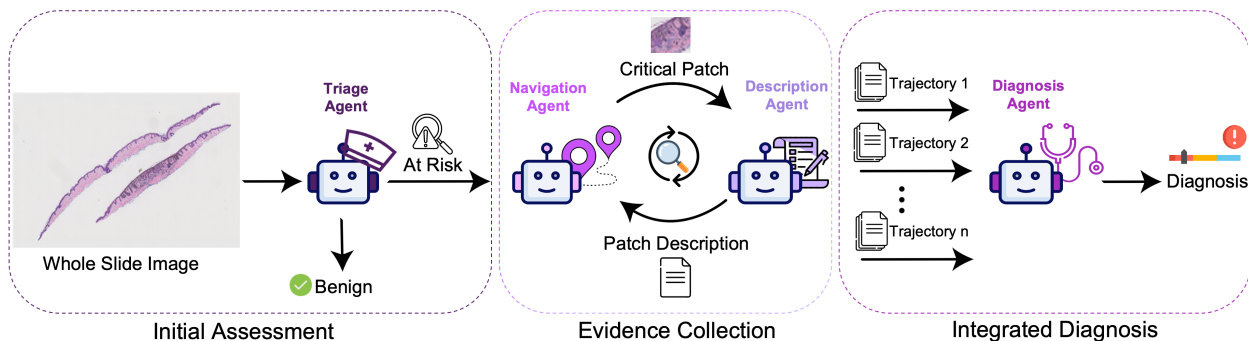


Figure 4.1: **PATHFINDER** consists of multiple steps carried out by multi-modal agents: 1) Initial Assessment by Triage Agent; 2) Evidence Collection by Navigation and Description Agents; and 3) Integrated Diagnosis by Diagnosis Agent.

4.2 Related Work

Multi-modal Histopathology Models. Recent advancements in histopathology have focused on leveraging whole-slide images (WSIs) and patch-level images to train unimodal classifiers using multiple instance learning approaches with pretrained feature extractors

[111, 129, 83]. More recently, unimodal foundational models trained on various self-supervised objectives have achieved notable improvements in downstream performance [131, 124, 54, 14]. The emergence of large-scale multi-modal datasets in histopathology has further advanced the development of large language models and vision-language models for the field. Using image-text datasets like Quilt-1M [53] or PathGen-1.6M [117], researchers trained models to learn joint vision-language representations. These models significantly enhance downstream clinical histology tasks at the patch level. At the WSI level, models like PathAlign [1] align diagnostic texts from pathology reports with their corresponding WSIs, enabling applications such as automated report generation and visual question answering. This aligns diagnostic tasks more closely with clinical workflows. Other works, such as QUILT-LLAVA [110], SlideChat [18], and PathChat [80], train Multi-modal Large Language Models (MLLMs) for histopathology to improve diagnostic reasoning. However, these models do not effectively address automated navigation of gigapixel-scale WSIs to support diagnosis, so they are not holistic.

The Role of Navigation in Histopathology Diagnosis. In computational pathology, several studies have explored the navigation behaviors of pathologists when examining digital slide images [103, 85, 86, 37]. These works often analyze patterns such as mouse movements, zooming, and field-of-view (FOV) panning to gather morphological clues for diagnosis. A common theme is comparing the navigation strategies of junior versus senior pathologists.

Multi-agent Systems. Recent research highlights the potential of large foundation models to enable interactive agent-based systems in robotics, involving interactions between robots, their environments, and humans [26, 43, 127]. These systems excel at handling complex tasks through collaboration and coordination among multiple task-specific agents. Applications span various domains, including role-playing [73], reasoning [24], gaming [47], and software engineering [45]. In medicine, some studies simulate clinicians treating patients and developing expertise through iterative interactions [74, 30].

4.3 Datasets

To lay the groundwork for describing our agents, we first start by introducing the different datasets used for training and evaluating our system. In this work, we specifically focused on skin pathology as that is the only cancer type that we have data modeling pathologists' viewing behavior.

M-Path Skin Biopsy WSIs. The skin biopsy WSIs in this dataset originate from the M-Path study [28, 11, 93], consisting of 238 melanocytic lesion specimens. A consensus reference panel of three dermatopathologists, each with internationally recognized expertise, independently interpreted all 238 cases and established a consensus diagnosis for each case through a series of review meetings. There are 4 diagnostic classes in this dataset: class 1 with 35 cases (mild and moderate dysplastic nevi); class 2 with 86 cases (severe dysplasia/melanoma in situ); class 3 with 70 cases (invasive melanoma stage pT1a); and class 4 with 47 cases (advanced invasive melanoma stage pT1b or more). For model development, the dataset was divided into training, validation, and test sets with a 168/35/35 case split, maintaining consistent class distribution across these sets.

M-Path Pathologists' Viewport Data. The M-Path study conducted viewport data collection, recruiting 87 pathologists from 10 U.S. states. Eligibility criteria included completion of residency and/or fellowship training and recent experience interpreting skin specimens in clinical practice. Pathologists' viewport data was gathered through an online digital slide viewer developed using Microsoft's open-source Silverlight-based HD View SL, a gigapixel image viewer. This viewer enabled pathologists to navigate each image by panning and zooming up to 60x magnification. During interpretation, the web-based viewer automatically logged viewport tracking data, capturing a rectangular image area displayed on the pathologist's screen at any given moment. For each interpretation (unique pathologist-case pair), the system recorded a list of viewport coordinates, magnification levels, and timestamps. Data from 32 pathologists who completed the M-Path study were included in the current study. Detailed methodology of the M-Path study is available in [93].

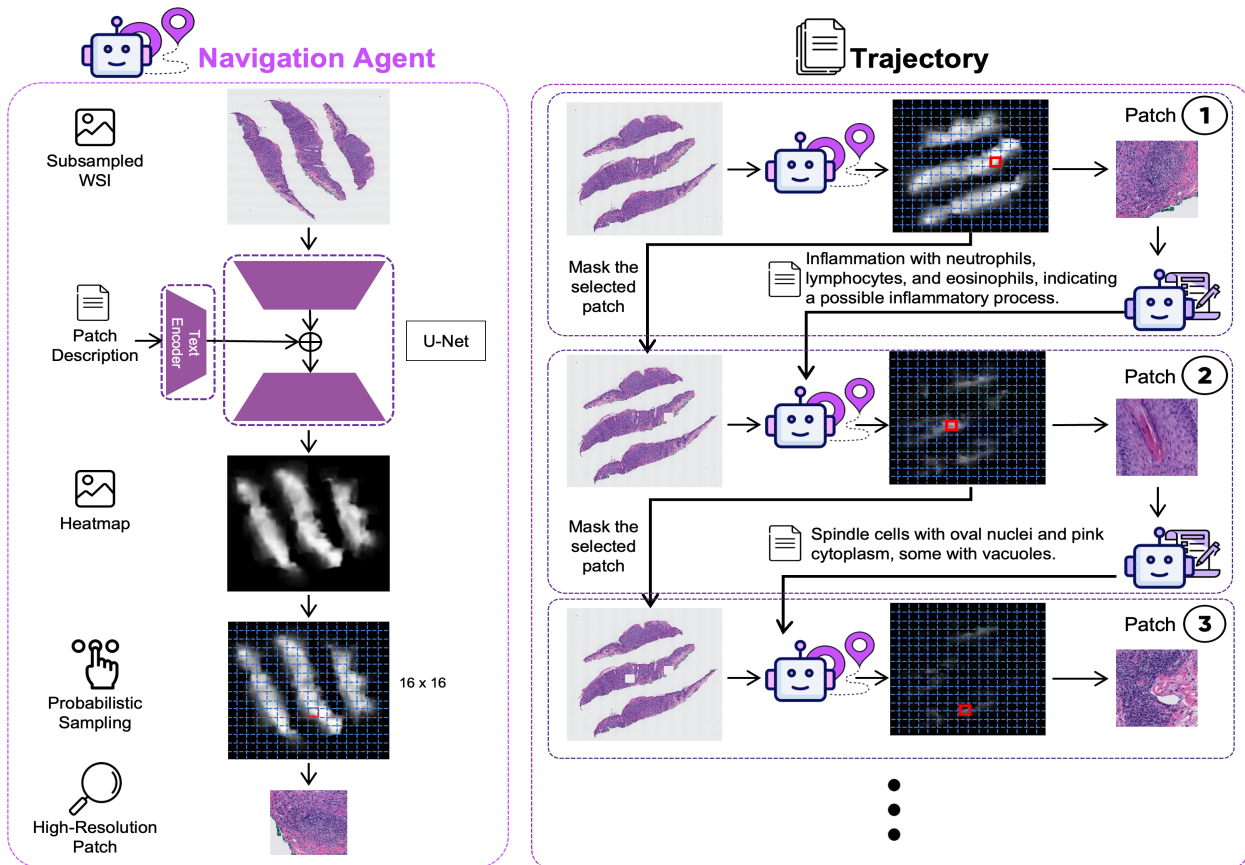


Figure 4.2: The left panel illustrates the Navigation Agent, as outlined in Section 4.4.2. The right panel presents the iterative trajectory generation process, which employs both the Navigation Agent and Description Agent, as described in Section 4.4.4.

4.4 PathFinder

The multi-agent multi-modal framework proposed in this chapter includes four agents: 1) Triage Agent ; 2) Navigation Agent ; 3) Description Agent ; and 4) Diagnosis Agent. The details of training data and model architectures are described below. Figure 4.1 demonstrates how the four agents interact with each other towards the final goal, which is diagnosing a WSI.

4.4.1 Triage Agent

The Triage Agent performs a triaging role, streamlining the diagnostic process by focusing attention on potentially malignant cases. Its goal is to filter out non-cancerous samples, ensuring only relevant cases are passed on for further analysis. The Triage Agent is an image-only transformer-based model. Below, we describe the data preparation, model architecture, and training details.

The Triage Agent includes several sequential stages: The feature vector is initially projected from $(N, 768)$ to (N, dim) using a linear layer to align with the model’s embedding dimension dim . For compatibility with 2D processing, the vector is reshaped into a square grid through padding to dimensions $H \times H$, where H is the smallest integer satisfying $H \times H \geq N$, with padding achieved by repeating the first $M = H^2 - N$ features. The padded vector is then processed through a transformer block, followed by positional encoding) [111], capturing spatial patterns across scales and emphasizing key features. The processed features are flattened, reprojected to the embedding dimension, and appended with a learnable class token for capturing the global context. After additional transformer blocks and positional encoding, the class token is pooled and passed through an MLP head for the final output. Our approach achieved higher F1-score and accuracy compared to other methods that are directly comparable for this task.

4.4.2 Navigation Agent

The Navigation Agent is designed to mimic a pathologist’s methodical approach to identifying regions of interest (ROIs) in WSIs. Unlike traditional systems that scan the entire WSI in a single, mechanistic sweep, our Navigation Agent adopts a more human-like, iterative process collaborating with the Description Agent. It begins by pinpointing an initial ROI, much as a pathologist would focus on one area at a time. This selected ROI is then sent to the Description Agent, which provides a natural-language description of the area. Figure 4.2 illustrates the workflow of the Navigation Agent in the left panel.

In our initial attempt, we designed the Navigation Agent using a multi-modal architecture inspired by LLaVA [77], integrating an image encoder and an LLM. The image encoder extracted features from a low-resolution version of the WSI, and the LLM processed these features along with previous text descriptions to predict the next ROI. Specifically, the WSI was divided into a grid of patches, and the LLM would output the grid coordinates of the most relevant patch based on both visual and textual inputs. However, this approach faced significant challenges due to the limited size of our training dataset. The model tended to overfit, frequently selecting central patches regardless of the input. This limitation prompted the exploration of more data-efficient methods that could better generalize from limited samples.

To overcome these challenges, we restructured the Navigation Agent to directly generate an importance map over the WSI, conditioned on textual descriptions from previous observations. This approach removes the dependency on the LLM for spatial selection and leverages a feedback mechanism between the image and text modalities. Let $I^{(t)}$ be the input WSI at iteration t , with previously selected patches masked out to avoid re-sampling and $D^{(1:t)} = \{D^{(1)}, D^{(2)}, \dots, D^{(t)}\}$ be the set of textual descriptions up to iteration t . At each iteration t , the Navigation Agent processes the masked WSI $I^{(t)}$ to predict an importance map $M^{(t)}$, indicating the likelihood of each region being the next ROI. The importance map is conditioned on the aggregated textual information from previous descriptions. We define the importance map generation as $M^{(t)} = f_{\text{Nav}}(I^{(t)}, E^{(t-1)})$ where, f_{Nav} is the Navigation Agent’s function (implemented as a lightweight U-Net [104]) that has four layers in both encoder and decoder and is conditioned with text embeddings of descriptions, as well as the masked version of the WSI that masks the earlier predicted ROIs, and $E^{(t-1)}$ is the aggregated text embedding up to iteration $t - 1$. $E^{(t-1)}$ is computed by encoding each description $D^{(k)}$ using a pre-trained Text-to-Text-Transfer-Transformer (T5) text encoder [101] and averaging the embeddings:

$$E^{(t-1)} = \frac{1}{t-1} \sum_{k=1}^{t-1} \text{T5}_{\text{text}}(D^{(k)}) \quad (4.1)$$

At the first iteration ($t = 1$), since there are no prior descriptions, the importance map is generated solely from the unmasked WSI $M^{(1)} = f_{\text{Nav}}(I^{(1)})$. From the importance map $M^{(t)}$, we then statistically sample the next patch to analyze. The probability $p_{(i,j)}^{(t)}$ of selecting a location (i, j) is proportional to its importance score:

$$p_{(i,j)}^{(t)} = \frac{M_{(i,j)}^{(t)}}{\sum_{(i',j')} M_{(i',j')}^{(t)}} \quad (4.2)$$

We then sample the patch coordinates (i^*, j^*) based on this probability distribution: $(i^*, j^*) \sim p_{(i,j)}^{(t)}$. The selected high-resolution patch corresponding to (i^*, j^*) is sent to the Description Agent, which generates a new textual description $D^{(t)}$. The new description $D^{(t)}$ is encoded and incorporated into the aggregated text embedding $E^{(t)}$:

$$E^{(t)} = \frac{1}{t} \sum_{k=1}^t \text{T5}_{\text{text}}(D^{(k)}) \quad (4.3)$$

This updated embedding $E^{(t)}$ is then used to condition the Navigation Agent in the next iteration, enabling the model to refine its importance map $M^{(t+1)}$ based on both the visual information from $I^{(t+1)}$ and the accumulated textual insights. Therefore, we refer to it as the Text-conditioned Visual Navigator.

To train the Navigation Agent, we constructed a dataset from M-Path [93] consisting of WSIs and sequences of textual descriptions for the most important patches. Each training sample includes: The WSI and the corresponding masked versions, the set of descriptions $D^{(1:t)}$ for each iteration generated by QUILT-LLAVA [110] and the ground truth importance maps derived from pathologist annotations. We minimized the binary cross-entropy loss between the predicted importance maps $M^{(t)}$ and the ground truth maps $\hat{M}^{(t)}$.

4.4.3 Description Agent

As our Description Agent, we utilize QUILT-LLAVA [110], a multi-modal large language model capable of describing individual histopathology patches. While the original QUILT-LLAVA

generates highly detailed findings, in this work, we instruction-tuned the model to produce more concise summaries, optimizing for computational efficiency as we needed to handle multiple descriptions. Using captions from the QUILT-1M dataset [53], we prompted GPT-4 to generate a list of findings that were as concise as possible. This process yielded 102,000 instruction-tuning samples. QUILT-LLAVA 7B model was instruction-tuned for one epoch to obtain the Description Agent.

4.4.4 *Diagnosis Agent*

The Diagnosis Agent is a language-only model that analyzes the natural text descriptions generated by the Description Agent for all patches identified by the Navigation Agent. It classifies the histopathological findings into three categories: classes 2, 3, and 4 (with class 1 already excluded by the Triage Agent). The Diagnosis Agent is based on a pre-trained GPT-2 model [100], which we fine-tuned for this task. We selected GPT-2 due to its lightweight architecture, making it a resource-efficient choice for our application.

To train the Diagnosis Agent, we generated diagnostic trajectories—sequences of patch descriptions that simulated how a pathologist examined a whole slide image (WSI). Using our Navigation Agent, we first obtained a heatmap for a sub-sampled WSI (512×512 pixels) using a text-conditioned U-Net model, which highlighted regions of diagnostic significance. The WSI was divided into a 16×16 grid, creating 256 patches of 32×32 pixels each. Each patch received an importance score based on the mean intensity of the heatmap over the patch, indicating its diagnostic relevance. These scores were normalized across all patches.

To generate a single trajectory, we iterated the following steps ten times, yielding ten patches per case: At each iteration, a patch was selected using weighted probabilistic sampling based on the normalized importance scores, introducing variability and ensuring different patches were chosen across iterations. The selected patch was then cropped from the high-resolution $10\times$ WSI, and a description was generated by the Description Agent. Selected patches were masked on the WSI to prevent reselection, and all descriptions generated thus far were combined into a single text input for the next iteration. The generation of a single

trajectory is presented in the right panel of Figure 4.2.

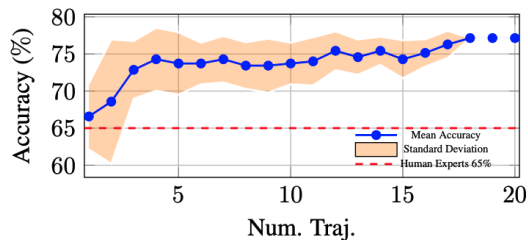
For each WSI in the training and validation sets, we generated five ($n = 5$) different trajectories, each containing ten patch descriptions, to capture various examination patterns. For the test set, we extracted additional trajectories ($n = 20$) to assess the effect of trajectory number on diagnosis results. To introduce further variability, we used a LLaMA 3.1 Instruct model [25] at each iteration to rephrase the text descriptions. This approach effectively simulated the variability among pathologists, who might examine a single case using different patterns while seeking diagnostically relevant regions. The Diagnosis Agent consists of an LLM with a classification head on top. The classification head maps the LLM’s output (vocabulary size) to the number of classes, producing the final classification probabilities using a single linear layer.

We expanded the training set to enhance diversity and robustness by resampling to create 20,000 cases, resulting in 100,000 trajectories for training. Each trajectory consists of a randomly selected number of descriptions (between five and ten), and we shuffled the sequence of descriptions within each trajectory to prevent over-fitting to any specific order. Each trajectory was formatted as a prompt to the LLM:

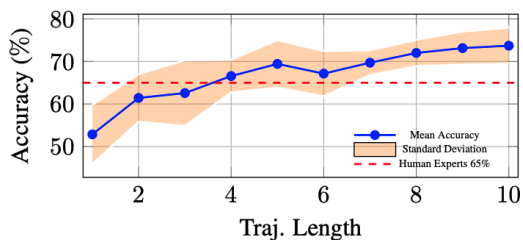
“The image descriptions below are extracted from different patches from the same whole slide image (WSI); please tell me which class the image belongs to: descriptions”, where *descriptions* is the list of selected descriptions.

4.5 Experiments and Results

This section outlines the experimental setup and evaluates the performance of the proposed PATHFINDER framework. First, we conducted a qualitative assessment of the descriptions generated by the Description Agent, comparing them to two vision-language models (VLMs). Next, we evaluated PathFinder on the M-Path dataset for melanoma diagnosis (see Section 4.3), benchmarking it against state-of-the-art transformer-based and MIL-based baselines, as well as public and private LLMs using prompting without additional training. Finally, we analyzed PathFinder’s performance under various configurations, altering the Triage,



(a) Majority voting accuracy for 1-20 trajectories. Each trajectory contains 10 patches visited by the navigator.



(b) Majority voting accuracy for 5 trajectories with 1-10 patches per trajectory.

Figure 4.3: Ablation results. We ran 10 experiments, and plotted both the mean and standard deviation.

Navigation, and Description Agents. Detailed evaluations are provided in the following subsections.

4.5.1 Pathologist Evaluation of Description Quality

To assess the quality of descriptions generated by our Description Agent, we conducted a survey in which two expert pathologists rated descriptions produced by our Description agent in comparison to those generated by GPT4-o [50] and LLaVA-Med [72]. We selected 25 cases from the M-Path dataset, sampling across the four diagnostic classes. For each case, we cropped the consensus region of interest, manually labeled by a panel of expert dermatopathologists as the area most representative of the diagnosis. Using this region, we prompted our Description Agent, LLaVA-Med and GPT4-o to generate concise descriptions of each histopathology patch. These descriptions were then presented to two expert pathologists in a randomized, double-blind format. Each pathologist was asked to respond to two questions for each case to indicate their preferred description and the reason for their preference. The results shown in Figure 4.4 indicate that PathFinder’s Description Agent achieved comparable performance to GPT-4o while being significantly more cost-effective, operating with just 7B parameters - a fraction of GPT-4o’s size.

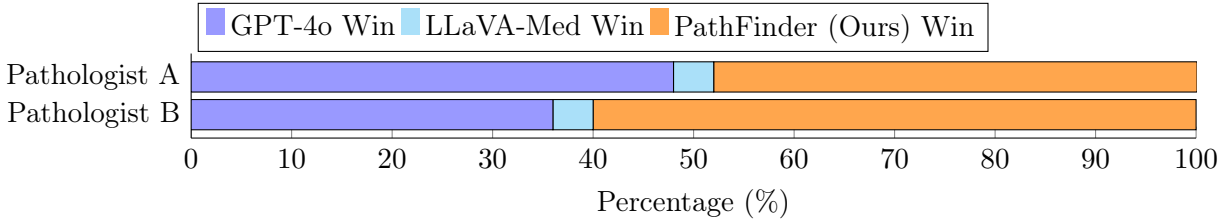


Figure 4.4: Preferences of expert human pathologists regarding the description quality of each model, assessed through a double-blind survey to ensure an unbiased comparison.

4.5.2 PathFinder Evaluation

For evaluating PathFinder, we utilize the M-Path dataset which contains histopathology WSIs of melanocytic skin tissue. As outlined in 4.4.4, multiple trajectories are generated per case to simulate the variability in diagnostic patterns observed among pathologists, who may assess a single case with diverse visual strategies to identify diagnostically significant regions. To mitigate randomness in our results, we evaluated PathFinder 10 times on the test set, each time using a different random subset of 5 trajectories selected from the total of 20. For each Whole Slide Image (WSI), majority voting is performed on the predictions from the 5 selected trajectories to produce the final result. The overall performance is then reported in Table 4.1 as the mean of the results across the 10 runs. We balanced the testing dataset to ensure that each diagnostic class is represented by an equal number of samples. Consequently, the micro-averaged F1 score, precision, and recall are equivalent to the accuracy reported in the table. We opted to use micro-averaged metrics in our clinical evaluation, because they appropriately balance the importance of different stages of skin cancer, which is crucial for assessing the overall reliability and effectiveness of the diagnostic tool.

We compared Pathfinder to four state-of-the-art baseline models: 1) three transformer-based models all utilizing the ScAtNet architecture [128, 34, 79] and 2) an MIL-based model, ABMIL [56]. ScAtNet utilizes a MobileNetV2 backbone [107] to extract multi-scale features from images at 7.5x, 10x, and 12.5x magnification. For the first baseline model, these feature

vectors are subsequently fed into ScATNet which aggregates information of the three scales to perform the diagnostic task using Transformer blocks. The second approach [34] augments the WSI with ROI heatmaps generated by the U-Net model, appending these maps as a fourth input channel and using ScAtNet for classification. The third baseline model, SAG [79], converts diagnostically relevant entities into attention signals, integrating these with ScAtNet and employing an attention-guiding loss function that combines heuristic guidance (HG) and tissue guidance (TG) based on disease-specific prior knowledge such as tissue, structure, and cellular information. Lastly, ABMIL [56] aggregates information across instances, ABMIL uses an attention mechanism that assigns weights to each instance, allowing the model to capture its contribution to the final bag label in a permutation-invariant manner. Then, we conducted comprehensive experiments to evaluate PathFinder by examining different architectures for each agent component, achieving 74% accuracy that surpasses both human experts (65%) and previous state of the art (63% best). Our evaluation focused on three main aspects:

Navigator Architectures. First, to quantify the importance of Description Agent feedback, we tested a Vision-Only Navigator that employs weighted probabilistic sampling for patch selection without iterative feedback in a single pass over the WSI. Additionally, we implemented Imitated Sampling, which leverages pathologists' viewing pattern distributions (viewport width, height, and zoom level) from our M-Path dataset (Section 4.3) to statistically sample patches as important WSI regions. If the pathologists spend more time focusing on a region, we gave a higher chance of sampling to that region. Both Imitated Sampling and Vision-Only Navigator performed similarly (64% and 63% respectively), indicating that both pure statistical and learned "sampling", regardless of source, has limited effectiveness. Furthermore, we evaluated text-conditioned visual navigators using either CLIP-based or T5-based text encoders. The T5-based navigator significantly outperformed its CLIP-based counterpart (74% vs 62%), suggesting CLIP's 77-token limit constrains its ability to effectively process multiple descriptions (we simply truncate descriptions exceeding 77 tokens, then average if a description is long). Finally, our navigation-based approach (74%) outperformed exhaustive

search (68%), which utilizes all non-background patches of the WSI, suggesting that selective patch sampling helps avoid confusion from irrelevant regions.

Description Agents. We compared a fine-tuned version of Quilt-LLaVA (optimized for concise descriptions) against off-the-shelf LLaVA-Med. The fine-tuned version showed superior performance (74% vs 56%), demonstrating better guidance for the navigator. Notably, when paired with LLaVA-Med descriptions, the T5-based Navigator showed no advantage over the CLIP-based version (56% vs 60%). This suggests that a more powerful text encoder like T5 can actually be detrimental when processing lower-quality descriptions, potentially steering the Navigator toward irrelevant regions. This finding emphasizes the importance of high-quality descriptions for effective navigation.

Diagnosis Agents. We evaluated various public and private LLMs as baselines for our Diagnosis Agent. Specifically, we used PathFinder with T5-Based Text-Conditioned Visual Navigator and Quilt-LLaVA Description Agent to generate multiple descriptions for each WSI and prompted LLMs to make the classification given the descriptions.

It is worth noting that without the Triage Agent, the performance of the best Pathfinder-variant dropped below baselines, likely due to Quilt-LLaVA’s train dataset’s bias toward malignant cases. The evaluation of the baseline models are similarly done using the majority voting over 10 runs. PathFinder achieves 11% improvement compared to the best image-only baseline approach, ScAtNet + ROI Heatmap [34]. Considering that GPT-2 is a relatively small LLM compared to the current state-of-the-art, we believe that utilizing larger LLMs could further improve diagnostic outcomes. Lastly, to investigate the impact of the number of trajectories on model performance, we evaluated the model using between 1 and 20 trajectories for majority voting, as well as the effect of varying trajectory lengths. Figure 4.3 illustrates this analysis, indicating optimal performance with 5 trajectories and 10 patches per trajectory. We ran every experiment for 10 rounds and reported the mean and standard deviation.

4.6 Discussion

This chapter presented PathFinder, a multi-modal, multi-agent AI framework designed to emulate the multi-scale, iterative diagnostic approach of expert pathologists for histopathology WSIs. By integrating Triage, Navigation, Description, and Diagnosis Agents, PathFinder collaboratively gathers evidence to deliver accurate, interpretable diagnoses with natural language explanations. Notably, it surpasses state-of-the-art methods and the average performance of human experts in melanoma diagnosis, setting a new benchmark in AI-driven pathology. PathFinder has the potential to accelerate diagnostic workflows, reducing the reliance on manual examination and enabling timely patient care in clinical settings. Its natural language descriptions provide interpretability, facilitating the validation of AI-generated diagnoses by pathologists. Moreover, its integration of VLMs and LLMs highlights the promise of multi-modal AI in delivering scalable, specialized diagnostic tools that could improve access to pathology expertise.

Limitations. Despite its strengths, PathFinder has limitations. The framework relies on pre-existing datasets and significant computational resources, posing challenges in resource-constrained environments. Additionally, the complexity of the Navigation Agent’s decision-making process and occasional hallucinations by the Description Agent could affect transparency and accuracy of the decision-making process. Future work should address these issues by enhancing dataset diversity, computational efficiency, and patch selection strategies, further advancing PathFinder’s potential as a transformative tool in AI-assisted pathology.

Methods	Accuracy	F-1 score
<i>Baselines</i>		
Human Experts [28]	0.65	0.65
ScAtNet [128]	0.62	0.62
ScAtNet + ROI Heatmap [34]	0.63	0.63
ScAtNet + SAG [79]	0.60	0.60
ABMIL [56]*	0.46	0.47
<i>LLM Prompting Baselines</i>		
BioMistral-7B	0.43	0.43
Mistral-Nemo-Instruct-2407	0.41	0.41
GPT-4o	0.49	0.49
Meta-Llama-3-8B-Instruct	0.31	0.31
LLaVA-Med-v1.5-Mistral-7b	0.43	0.43
Quilt-LLaVA-v1.5-7b	0.29	0.29
<i>Ours</i>		
PathFinder + T5-Based Text-Conditioned Visual Navigator + No Triage	0.58	0.58
PathFinder + T5-Based Text-Conditioned Visual Navigator + LLaVA-Med	0.56	0.56
PathFinder + CLIP-Based Text-Conditioned Visual Navigator + LLaVA-Med	0.60	0.60
PathFinder + Imitated Sampling	0.63	0.63
PathFinder + Vision-Only Navigator	0.64	0.64
PathFinder + CLIP-Based Text-Conditioned Visual Navigator	0.62	0.62
PathFinder + Exhaustive search	0.68	0.68
PathFinder + T5-Based Text-Conditioned Visual Navigator	0.74	0.74

* ABMIL result is based on a single run and does not use majority voting

Table 4.1: Majority voting performance for whole slide image (WSI) diagnosis on the M-Path dataset. Accuracy is reported, and the F-1 score is identical due to the balanced testing set. Finally, coverage here is the percent of patches used across all trajectories.

Chapter 5

BEYOND HISTOPATHOLOGY - TARGETING THE ENTIRE MEDICAL DOMAIN WITH MEDICALNARRATIVES

In this chapter, we build upon the video processing framework of MedicalNarratives [53] and the cursor detection framework of GenMedCLIP [110] to create a large-scale grounded image-text dataset for the broader medical domain.

5.1 Introduction and Motivation

Medical image analysis is inherently perceptual and conceptual, requiring experts to extract visual information and apply clinical knowledge to form a diagnosis [88]. This process necessitates spatial grounding to link regions of the image with clinical understanding. Despite advancements, semantic tasks (e.g., classification) and dense tasks (e.g., segmentation) are often studied separately due to the lack of integrated models and datasets capable of addressing both fields.

Localized Narratives [97] demonstrated the potential of datasets that combine image-text data with spatial traces by capturing annotators' spoken descriptions alongside cursor movements. Such datasets are easier to collect and provide robust grounding for training models on tasks like classification, segmentation, and multimodal generation. However, the medical domain lacks similar large-scale datasets. A recent work, MedTrinity-25M [130], attempts to address this by releasing a multimodal dataset with spatial annotations, but relies on sub-optimally pretrained models to generate text descriptions and Regions of Interests (ROIs) for medical images lacking ground truth annotations, potentially propagating model biases and errors.

To address this gap, in this chapter, we introduce MedicalNarratives, a multimodal

dataset derived from pedagogical medical videos where instructors narrate and point to relevant regions with a cursor, mimicking the think-aloud protocol [27]. Our dataset includes 4.2M image-text pairs across 11 medical modalities and 1 pseudo-medical domain, with 875K grounded samples usable for dense tasks like segmentation. We evaluate GenMedCLIP, trained on MedicalNarratives, across classification and retrieval tasks and show that it outperforms prior SOTA models by 3% and 14%, respectively. By releasing this dataset, alongside video clips and metadata such as UMLS entities, we aim to enable future research on grounded generative models and open-vocabulary segmentation in the medical domain.

Building on prior work Quilt [53] and Quilt-LLaVA [110], we present MedicalNarratives [52], a dataset derived from pedagogical medical videos where instructors narrate while pointing to relevant regions with their cursor, closely mirroring the think-aloud protocol used in clinical practice. MedicalNarratives includes 4.2M image-text pairs spanning 11 medical modalities and 1 pseudo-medical domain, with interleaved samples across modalities (e.g., X-ray and CT from the same patient). Notably, 875K samples are grounded with traces that can be reformatted into bounding boxes or masks, enabling training for tasks such as segmentation and dense modeling with tools like MedSAM [81, 16], while remaining adaptable to clinical workflows.

To evaluate the quality of our dataset, we trained a vision-language model, GenMedCLIP, and tested it on a new benchmark covering all 11 modalities for classification and retrieval tasks. GenMedCLIP outperformed prior SOTA models like BiomedCLIP, achieving gains of 3% in classification and 14% in retrieval. We envision future research leveraging MedicalNarratives for training grounded generative models like Quilt-LLaVA [110], LLaVA-Med++ [130], and PixelLLM [132], as well as for advancing open-vocabulary segmentation and detection. Additionally, we release the video clips and accompanying metadata, including UMLS entities, to support broader clinical and research applications.

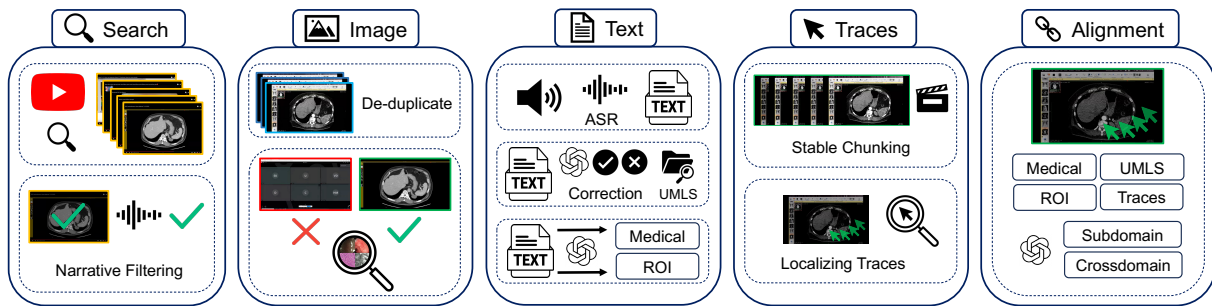


Figure 5.1: The data curation pipeline for the Video subset of the MedicalNarratives dataset. **Search:** Videos are collected from YouTube using a focused search approach, refined through pre-trained classifiers and heuristic methods. **Image:** Keyframes are extracted, cleaned, and evaluated to detect medical imagery. **Text:** Audio is transcribed via ASR, rectified using LLMs, and processed to extract medical or ROI-relevant text. **Traces:** Stable video segments are segmented, and cursor traces are mapped within these segments. **Alignment:** Medical/ROI text, traces, and images are aligned, with samples categorized into specific subdomains or labeled as multi-domain using LLMs.

5.2 Related Work

Vision-Language Representation Advances in vision-language (VL) models, particularly contrastive learning approaches [98, 140], have enabled strong performance in both supervised and zero-shot tasks by aligning visual and textual embeddings. In medical imaging, early models focused on narrow domains, such as radiology [140, 48], and later extended to other fields like ophthalmology [112], histopathology [51], and dermatology [65]. While these models excel in specific domains, they struggle to generalize due to limited datasets, driving the push for broader, multi-domain medical VL models [138, 142].

Medical (Localized) Narratives Most medical VL datasets rely on textual data from reports [61], PubMed [29], or videos [51]. However, these datasets often lack spatial annotations, limiting their use for tasks like segmentation and detection. Localized Narratives [97] overcome

this by grounding text to image regions via point or trace annotations. These datasets have supported semantic reasoning [97], dense tasks [39], and generative modeling [132, 110]. In medical contexts, works like Quilt-LLaVA [110] and PathNarratives [136] demonstrate the potential of grounded datasets to enhance explainability and improve diagnostic workflows.

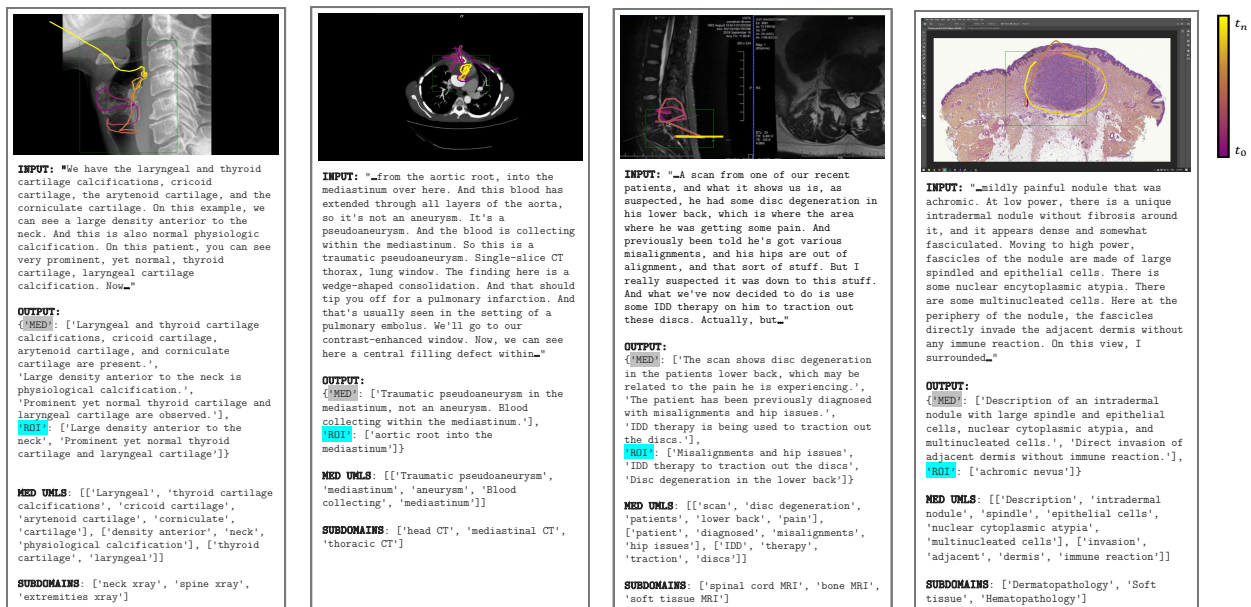


Figure 5.2: **MedicalNarratives**: We present four dataset samples, each representing a distinct medical modality: **X-ray**, **CT**, **MRI**, and **Histopathology** images (displayed from left to right). These examples, extracted from interleaved videos, include the representative image captured, the raw input text aligned with spatial traces and bounding boxes, and the refined medical and ROI text, which has been denoised to eliminate transcription errors and non-medical content.

5.3 *MedicalNarratives: Curation, Formatting, and Characterization*

5.3.1 *Overview*

MedicalNarratives builds on prior methodologies introduced in Quilt-1M and Quilt-LLaVA to curate a multimodal vision-language dataset with spatial traces from unstructured pedagogical videos and scientific articles. Challenges such as noisy data, irrelevant content, and specialized medical language transcription were addressed through carefully designed algorithms and workflows. This section outlines the sourcing, filtering, and processing steps for curating MedicalNarratives.

5.3.2 *Sourcing and Filtering Medical Data*

Videos: We primarily followed the Quilt-1M pipeline to source videos from YouTube, focusing on 12 medical imaging domains, including Magnetic Resonance Imaging (MRI), X-Rays, Computed Tomography (CT), Ultrasound (US), Mammography (Mammo), Surgery (Surg), Endoscopy (Endo), Dentistry, Dermatology, Ophthalmology (Ophtha), Histopathology (Histo), and General Medical Illustrations. Slight modifications were made to heuristics for identifying relevant videos, including filtering by duration, voice presence, and domain-specific visual content. Narrative-style videos were selected by analyzing keyframes, transcripts, and domain-specific thresholds for consistency and medical relevance.

Articles: Articles were sourced from the PubMed Central Open Access Subset (PMC-OA), extracting figures and captions from 5.47M articles. Inline figure references were also retrieved to pair figures with their contextual descriptions. This approach ensured a well-structured and diverse dataset across all 12 medical imaging domains.

5.3.3 *Text Extraction and Denoising*

Videos: We used Whisper for ASR transcription, following the methodology outlined in Quilt-1M [53]. Errors in transcription, particularly for medical terminology, were corrected

using language models and domain-specific medical verifiers to ensure accurate and clean textual data.

Articles: Captions for figures were extracted from article XML documents, and compound captions were split into sub-captions using rule-based methods and language models. Inline references were matched to their corresponding sub-figures to provide clear associations between text and visual elements.

5.3.4 Image Extraction and Denoising

Videos: Keyframes were identified using the MedicalNarrativescombo pipeline and processed to extract representative images from stable video segments. Median images from static frames were used for domains like X-rays and CT scans, leveraging the natural pauses instructors make while explaining content.

Articles: Figures from articles were processed to extract sub-figures, addressing challenges like irregular layouts and non-standard labels. We trained an object detection model based on the YOLO architecture [60] using sub-figure annotation datasets MedICaT and ImageCLEF 2016 [116, 33] to split and label sub-figures, ensuring proper alignment with their captions.

5.3.5 Extracting Localizing Traces

Cursor traces were extracted using the GenMedCLIP [110] methodology, which identifies stable video chunks with minimal background movement. Frames within these chunks were analyzed to detect cursor positions based on pixel intensity changes. Structural similarity metrics and face detection models were used to minimize false positives caused by minor background changes or distractions. This algorithm offers a surprisingly generalizable way to collect cursor traces from any medical video in static domains like X-rays and histopathology. Finally, we extracted the smallest bounding boxes around the traces.

5.3.6 *Characterizing MedicalNarratives*

MedicalNarratives combines medical narratives from videos with image-text pairs from PubMed, totaling 4.7M image-text samples, 1M of which include localized narratives.

Narratives

From 738K videos, we curated 74K narrative-style videos (10.1% yield), amounting to 4526 hours of content. These contributed 809K unique images and 1.62M image-text pairs, with 1M samples grounded by 101.6M trace points and 546K bounding boxes. The dataset spans four trace domains (CT, MRI, X-ray, Histopathology) with bounding box averages of H: 291px and W: 357px. Captions average 22.37 words, with 1.88 medical sentences per image, covering over 4M UMLS entities across 300K unique types.

Non-Narratives

From 5.4M PubMed articles, we filtered 23M figures down to 1.03M medical figures from 273K articles. After sub-figure separation, we obtained an average of 2.62 subfigure-caption pairs per article figure, with captions averaging 45.45 words.

5.3.7 *Quality*

Although MedicalNarratives lacks human-annotated spatial data for direct evaluation, we assessed the pipeline’s performance through key metrics. LLM-based text correction achieved a conditioned precision of 47.99%, validated against UMLS, and an unconditioned precision of 17.58%, indicating its ability to handle both domain-specific and general corrections. The ASR system also performed well, with an average error rate of 0.81%, demonstrating its reliability in transcribing medical terminology accurately. To ensure ethical compliance, we further refined the dataset by prompting the LLM to exclude identifiable information, such as clinic names or addresses, while retaining only medically relevant content. These measures ensured that MedicalNarratives is accurate, anonymized, and suitable for medical research.

5.4 GenMedCLIP: Experiments and Results

To evaluate MedicalNarratives, we tested its utility on two tasks: image classification (both zero-shot and linear probing) and cross-modal information retrieval (zero-shot) across all medical imaging modalities. These experiments follow the same protocols as in Quilt, leveraging the Contrastive Language-Image Pre-training (CLIP) objective [98].

We pre-trained a vision-language model, GenMedCLIP, using OpenCLIP [55] with adaptations to encoder and tokenizer choices informed by prior work [138, 51]. For the image tower, we fine-tuned Vision Transformer (ViT-Base) models [23] pre-trained on supervised (ViT-B/16, ViT-B/32) and unsupervised contrastive objectives (ViT-B/16). For the text tower, we used GPT2 [100] and BioMedBERT [41]. Training was conducted on 4 NVIDIA A40 GPUs for 20 epochs over 4.2M training samples. We trained three variants of GenMedCLIP for comparison: GENMEDCLIP-32 (ViT-B/32 and GPT2/77), GENMEDCLIP-PMB (ViT-B/16 and BioMedBERT/256), and GenMedCLIPPMB (ViT-B/16 and GPT2/77).

Model	Isic	Til	Pcam	Mhist	Nck	Mammo	Avg
CLIP-ViT-B-16 [98]	71.23	91.23	82.42	63.97	92.26	83.30	80.74
PUBMEDCLIP [29]	68.58	91.32	84.07	72.16	92.29	83.90	82.06
BIOMEDCLIP [138]	68.25	91.82	83.43	66.73	93.05	83.70	81.17
GENMEDCLIP-32	72.75	93.26	86.77	72.06	92.77	83.70	83.55
GENMEDCLIP-PMB	69.38	91.51	84.54	67.66	88.02	84.20	80.88
GENMEDCLIP	74.87	93.34	87.69	72.16	90.84	84.90	83.97

Table 5.1: **Linear Probing** results (Accuracy) across datasets for classification tasks in Dermatology (*Isic*), Histopathology (*pcam*, *mhist*, *nck*), and Mammography (*vinDr-Mammo*). GenMedCLIP demonstrates superior performance compared to all baselines.

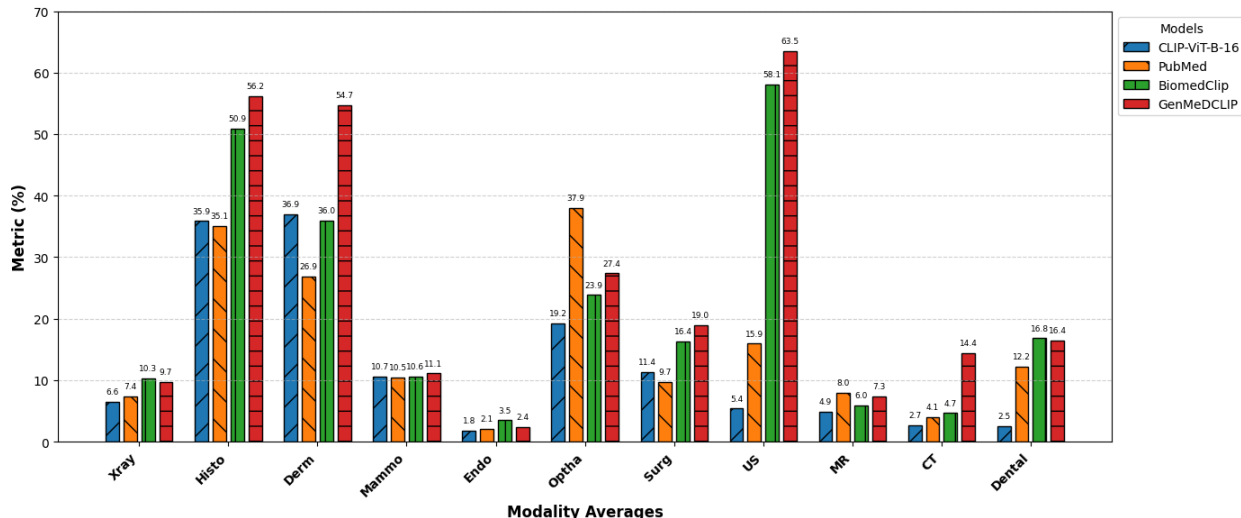


Figure 5.3: **Zeroshot Classification Results** shows that our model GenMedCLIP outperforms all other baselines across multiple medical benchmarks covering all 11 medical domains. Mean average precision is reported for X-ray and Mammography and accuracy is used for all other domains.

5.4.1 Benchmarking on Downstream Medical Tasks

We evaluated GenMedCLIP on a comprehensive medical imaging benchmark encompassing all domains represented in MedicalNarratives. This included 29 downstream classification tasks and a held-out set of 1000 unique images for cross-modal retrieval.

For MRI, we used subsets of **RadImageNet** [84], covering seven anatomical regions: ankle/foot, brain, knee, abdomen/pelvis, hip, shoulder, and spine, with class counts ranging from 9 to 26. For CT, we evaluated on RadImageNet’s lung and abdomen/pelvis subsets, covering 6 and 28 classes, respectively. Ultrasound tasks were based on thyroid and abdomen/pelvis anatomical regions, spanning 15 classes. For X-ray, we used **VinDr-CXR** [90], reporting mAP across 28 findings, while for mammography, we tested on **VinDr-Mammo** [91] using the bilateral craniocaudal view, reporting mAP across findings. Surgical organ classification was evaluated using **Dresden** [12] (8 abdominal organs), and for endoscopy,

we used **GastroVision** [58] (27 procedures). Dermatology tasks included **DDI** [21] for binary classification (benign vs. malignant) and **ISIC 2018** [20]. Dental classification was performed on the **OPG** dataset [102] with 6 classes. Ophthalmology benchmarks included **G1020** [4] for glaucoma detection and **OCTDL** [68] with 6 disease classes. Histopathology evaluations were conducted on diverse datasets: **PatchCamelyon** [121] for binary lymph node metastasis detection, **NCT-CRC-HE-100K** [64] (8 morphological classes), **BACH** [2] (4 breast tissue classes), **Osteo** [3] (osteosarcoma with 3 classes), **SkinCancer** [67] (skin biopsy patches with 12 anatomical and 4 neoplasm categories), **MHIST** [126] (colorectal polyp classification), **LC25000** [8] (lung and colon adenocarcinoma), and **TCGA-TIL** [106] for binary tumor-infiltrating lymphocyte prediction.

Models	T2I retrieval			I2T retrieval			Avg
	@5	@50	@200	@5	@50	@200	
CLIP-ViT-B-16 [98]	3.48	20.38	35.69	3.56	20.39	35.42	19.82
PubMedclip [29]	1.44	12.68	25.44	1.10	12.30	24.07	12.84
BiomedCLIP [138]	16.50	51.48	67.46	15.71	48.85	64.61	44.10
GenMedCLIPthreetwo	22.36	76.33	88.60	20.75	75.15	88.23	61.90
GenMedCLIPPMB	28.29	82.91	92.43	29.21	82.91	92.43	68.03
GenMedCLIP	34.89	83.83	92.27	34.26	83.48	92.32	70.17

Table 5.2: **Retrieval** results on our held-out set of 16K samples spanning all medical domains demonstrate that our model, GenMedCLIP, outperforms all baselines in both zeroshot image-to-text and text-to-image retrieval tasks.

5.4.2 Zero-shot classification

To evaluate the generalization ability of GenMedCLIP, we tested its zero-shot performance against state-of-the-art models: CLIP, BiomedCLIP, and PubMedclip. Across the benchmark,

GenMedCLIP variants achieved average metrics of 31.33% (GenMedCLIP_{threetwo}), 31.46% (GenMedCLIP_{PMB}), and 32.55% (GenMedCLIP), all outperforming BiomedCLIP (27.80%) by up to 4.75%. As shown in Figure 5.3, GenMedCLIP outperformed baselines in five domains: Histopathology, Dermatology, Mammography, Surgery, Ultrasound, and CT, while remaining competitive in Chest X-ray, Endoscopy, Ophthalmology, and MRI.

5.4.3 Supervised linear probing

We evaluated full-shot performance via linear probing using the entire training data on five benchmark datasets from Dermatology, Histopathology, and Mammography domains. On average, GenMedCLIP outperformed BiomedCLIP and CLIP, achieving a 2.8% and 1.9% improvement, respectively. The ViT-B/32 architecture with GPT/77 consistently demonstrated strong performance, surpassing other baselines across most datasets.

5.4.4 Cross-Modal Retrieval

To assess zero-shot text-to-image and image-to-text retrieval, we tested GenMedCLIP on a held-out set of 16K image-text pairs spanning 12 medical modalities, roughly equally distributed. Retrieval was conducted by identifying the top-N nearest neighbors ($N \in \{1, 50, 200\}$). Results in Table 5.2 show that GenMedCLIP outperformed baselines, exceeding BiomedCLIP by 26.07% on average. Notably, consistent with BiomedCLIP findings [138], general CLIP outperformed in-domain PubMedclip by 6.98%.

5.5 Conclusion

In this chapter, we introduced a comprehensive protocol for annotating medical narratives, aimed at standardizing the creation of multimodal medical datasets. Our curated dataset, MedicalNarratives, tackles the challenges of large-scale medical data collection by balancing utility with cost and ease of collection. We propose this protocol as a potential standard for future multimodal medical datasets due to its versatility in capturing grounded text that

effectively describes medical images. By implementing and scaling this approach, we enable systematic and scalable dataset generation across broader medical domain.

In semantic tasks, our GenMedCLIP demonstrated superior performance compared to prior vision-language (VL) models, achieving state-of-the-art results in zero-shot and full-shot classification as well as retrieval tasks on a carefully curated medical imaging benchmark. Despite using only 4.7M samples, compared to BiomedCLIP’s 15M, our in-domain data filtration methods proved highly effective. We envision future research building upon our models, dataset, and protocol to drive further advancements.

5.6 *Limitations and Future Work*

While our approach and findings have contributed significantly to the development of multi-modal medical datasets and model performance in the medical domain, certain limitations need to be acknowledged:

1. **Bias in Abnormality Representation:** Our dataset over represents abnormal cases, reflecting diagnostic imaging practices. This imbalance may affect model generalizability to more balanced populations and could introduce bias in clinical decision support.
2. **Focus on 2D Imaging:** The dataset is limited to 2D images, excluding 3D modalities like MRI and CT. While 2D approaches can be effective, incorporating 3D imaging would enhance applicability and performance in certain tasks.
3. **Exclusion of Other Data Modalities:** This work focuses solely on imaging, omitting critical data such as laboratory results, time series signals, and patient histories. Including these is essential for developing comprehensive medical models.
4. **Lack of Ground Truth for Localization:** The dataset lacks human-annotated localization data, limiting the ability to directly evaluate the accuracy of trace annotations.

Future work involves building an MLM like Quilt-LLaVA for the entire medical domain from MedicalNarratives and finally automating the entire medical diagnostic pipeline for other fields with frameworks similar to PathFinder.

Chapter 6

CONCLUSION

This dissertation has investigated the integration of artificial intelligence (AI) and digital pathology to enhance histopathological diagnosis through interpretable and automated frameworks. Across four major chapters, we introduced large-scale vision-language datasets, explored multi-modal architectures, and developed agent-based diagnostic systems that collectively address the challenges of data scarcity, diagnostic reasoning, and workflow automation in histopathology.

In Chapter 2, we introduced QUILT-1M, a large-scale vision-language dataset designed to address the critical challenge of data scarcity in histopathology. By curating over one million image-text pairs from educational pathology videos, QUILT-1M serves as a foundational resource for training AI models capable of understanding and communicating histopathological findings in human-interpretable terms.

In Chapter 3, we developed QUILT-LLAVA, a multi-modal large language model fine-tuned on QUILT-1M and the QUILT-INSTRUCT dataset. This system bridges the gap between visual and linguistic understanding in histopathology, enabling sophisticated reasoning capabilities and offering novel tools for education, training, and diagnostic support. QUILT-LLAVA demonstrated significant advancements in multi-modal tasks, such as visual question answering and patch-level analysis.

In Chapter 4, we presented PathFinder, a multi-agent AI system that emulates the diagnostic workflow of expert pathologists. By integrating Triage, Navigation, Description, and Diagnosis Agents, PathFinder not only automates the examination of gigapixel whole slide images (WSIs) but also provides interpretable natural language explanations for its decision. This system achieved diagnostic accuracy exceeding that of experienced pathologists,

highlighting its potential to enhance both efficiency and reliability in histopathological workflows.

Finally, Chapter 5 expanded the scope of our methods to broader medical imaging applications. MedicalNarratives, a multimodal dataset featuring spatially grounded image-text pairs, was created to support diverse tasks beyond histopathology, including open-vocabulary segmentation and retrieval. Building upon the principles established in Quilt [53], MedicalNarratives empowers the development of models like GenMedCLIP, which set new benchmarks in classification and retrieval tasks across multiple medical modalities.

Collectively, these contributions illustrate how AI can transform digital pathology workflows, enabling more efficient, accurate, and interpretable diagnostic processes. However, histopathology is an exceptionally complex field, where even seasoned experts with years of experience often struggle to achieve consistently high success rates. Despite significant advancements in AI, these models still suffer from hallucinations and make incorrect predictions, making them unreliable as standalone diagnostic tools. Nevertheless, the AI models could be utilized as copilots, or validator tools to assist in automating the diagnostic process, as demonstrated in the case of PathFinder. In this approach, an expert pathologist can verify the results by reviewing highlighted patches, AI-generated descriptions, and the final diagnostic concurrence. Thus, human experts could save valuable time by validating cases rather than solving them from scratch.

The following could be interesting avenues for future research:

- **Integrative AI-Pathologist Collaboration.** Developing advanced human-AI interaction frameworks that enable seamless collaboration between pathologists and AI systems. This includes creating tools for real-time feedback, adaptive learning based on pathologist input, and mechanisms for incorporating expert corrections to iteratively improve model performance over time.
- **Uncertainty Quantification and Communication.** Incorporating robust uncertainty quantification mechanisms into diagnostic models, enabling them to express

confidence levels and highlight ambiguous cases. Such systems would assist pathologists by explicitly identifying cases requiring additional scrutiny, thereby improving reliability and trust in AI-assisted workflows.

- **Multimodal Domain Expansion.** Extending PathFinder framework beyond histopathology to embrace diverse imaging modalities (e.g., MRI, X-ray, ultrasound) can enable more generalized interpretable diagnostic AI frameworks. Comprehensive multi-modality datasets will further facilitate robust cross-domain analysis.

In conclusion, the research outlined in this dissertation highlights the potential of AI to reshape histopathology by accelerating and improving diagnostic processes. By pushing the boundaries of data curation, multi-modal model design, and agent-based automation, we have taken decisive steps toward a future where AI augments—rather than replaces—the nuanced expertise of clinical professionals. Continuing to refine and expand these methods promises to further elevate diagnostic accuracy, streamline workflows, and ultimately improve patient 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] 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.
- [3] 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.
- [4] Muhammad Naseer Bajwa, Gur Amrit Pal Singh, Wolfgang Neumeier, Muhammad Imran Malik, Andreas Dengel, and Sheraz Ahmed. G1020: A benchmark retinal fundus image dataset for computer-aided glaucoma detection. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2020.
- [5] Bowen Baker, Ilge Akkaya, Peter Zhokov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. Video pretraining (vpt): Learning to act by watching unlabeled online videos. *Advances in Neural Information Processing Systems*, 35:24639–24654, 2022.
- [6] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Res.*, 32(Database-Issue):267–270, 2004.
- [7] 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.
- [8] 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.

- [9] 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.
- [10] 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.
- [11] 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.
- [12] Matthias Carstens, Franziska M Rinner, Sebastian Bodenstedt, Alexander C Jenke, Jürgen Weitz, Marius Distler, Stefanie Speidel, and Fiona R Kolbinger. The dresden surgical anatomy dataset for abdominal organ segmentation in surgical data science. *Scientific Data*, 10(1):1–8, 2023.
- [13] 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.
- [14] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024.
- [15] Richard J Chen, Ming Y Lu, Wei-Hung Weng, Tiffany Y Chen, Drew FK Williamson, Trevor Manz, Maha Shady, and Faisal Mahmood. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4025, 2021.
- [16] Tianrun Chen, Lanyun Zhu, Chaotao Ding, Runlong Cao, Yan Wang, Zejian Li, Lingyun Sun, Papa Mao, and Ying Zang. Sam fails to segment anything? – sam-adapter: Adapting sam in underperformed scenes: Camouflage, shadow, medical image segmentation, and more, 2023.
- [17] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

- [18] Ying Chen, Guoan Wang, Yuanfeng Ji, Yanjun Li, Jin Ye, Tianbin Li, Bin Zhang, Nana Pei, Rongshan Yu, Yu Qiao, et al. Slidechat: A large vision-language assistant for whole-slide pathology image understanding. *arXiv preprint arXiv:2410.11761*, 2024.
- [19] 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.
- [20] Noel Codella, Veronica Rotemberg, Philipp Tschandl, M Emre Celebi, Stephen Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael Marchetti, et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1902.03368*, 2019.
- [21] Roxana Daneshjoui, Kailas Vodrahalli, Roberto A Novoa, Melissa Jenkins, Weixin Liang, Veronica Rotemberg, Justin Ko, Susan M Swetter, Elizabeth E Bailey, Olivier Gevaert, et al. Disparities in dermatology ai performance on a diverse, curated clinical image set. *Science advances*, 8(31):eabq6147, 2022.
- [22] 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.
- [23] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.
- [24] 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.
- [25] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [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] Steven J Durning, John Graner, Anthony R Artino Jr, Louis N Pangaro, Thomas Beckman, Eric Holmboe, Terrance Oakes, Michael Roy, Gerard Riedy, Vincent Capaldi, et al. Using functional neuroimaging combined with a think-aloud protocol to explore clinical reasoning expertise in internal medicine. *Military Medicine*, 177(suppl_9):72–78, 2012.
- [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] Zhihao Fan, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, Jun Xi, Fei Huang, and Jingren Zhou. Ai hospital: Interactive evaluation and collaboration of llms as intern doctors for clinical diagnosis. *arXiv preprint arXiv:2402.09742*, 2024.
- [31] 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.
- [32] 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.
- [33] Alba García Seco de Herrera, Roger Schaer, Stefano Bromuri, and Henning Müller. Overview of the ImageCLEF 2016 medical task. In *Working Notes of CLEF 2016 (Cross Language Evaluation Forum)*, September 2016.
- [34] 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.
- [35] 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*, pages 1–16, 2024.

- [36] Fatemeh Ghezloo, Mehmet Saygin Seyfioglu, Rustin Soraki, Wisdom O Ikezogwo, Beibin Li, Tejoram Vivekanandan, Joann G Elmore, Ranjay Krishna, and Linda Shapiro. Pathfinder: A multi-modal multi-agent system for medical diagnostic decision-making applied to histopathology. *arXiv preprint arXiv:2502.08916*, 2025.
- [37] 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.
- [38] Shah Giashuddin and Mouyed Alawad. Histopathological diagnosis of nonalcoholic steatohepatitis (nash). In *Non-Alcoholic Steatohepatitis: Methods and Protocols*, pages 1–18. Springer, 2022.
- [39] Cristina González, Nicolás Ayobi, Isabela Hernández, José Hernández, Jordi Pont-Tuset, and Pablo Arbeláez. Panoptic narrative grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1364–1373, October 2021.
- [40] 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.
- [41] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing, 2020.
- [42] 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.
- [43] 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.
- [44] 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.

- [45] 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.
- [46] 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.
- [47] 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.
- [48] 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 (ICCV)*, pages 3942–3951. IEEE, 2021.
- [49] 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.
- [50] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- [51] Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1m: One million image-text pairs for histopathology. *Advances in neural information processing systems*, 36, 2024.
- [52] Wisdom O Ikezogwo, Kevin Zhang, Mehmet Saygin Seyfioglu, Fatemeh Ghezloo, Linda Shapiro, and Ranjay Krishna. Medicalnarratives: Connecting medical vision and language with localized narratives. *arXiv preprint arXiv:2501.04184*, 2025.
- [53] 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.
- [54] Wisdom Oluchi Ikezogwo, Mehmet Saygin Seyfioglu, and Linda Shapiro. Multi-modal masked autoencoders learn compositional histopathological representations. *arXiv preprint arXiv:2209.01534*, 2022.

- [55] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hananeh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. If you use this software, please cite it as below.
- [56] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.
- [57] Julie A Jacko. Human computer interaction handbook: Fundamentals, evolving technologies, and emerging applications. 2012.
- [58] Debesh Jha, Vanshali Sharma, Neethi Dasu, Nikhil Kumar Tomar, Steven Hicks, M. K. Bhuyan, Pradip K. Das, Michael A. Riegler, Pål Halvorsen, Ulas Bagci, and Thomas de Lange. Gastrovision: A multi-class endoscopy image dataset for computer aided gastrointestinal disease detection, 2023.
- [59] 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.
- [60] Glenn Jocher, Jing Qiu, and Ayush Chaurasia. Ultralytics YOLO, 2023.
- [61] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, 2019.
- [62] 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.
- [63] Daniel Kahneman. *Attention and effort*, volume 1063. Citeseer, 1973.
- [64] Jakob Nikolas Kather, Niels Halama, and Alexander Marx. 100,000 histological images of human colorectal cancer and healthy tissue. *Zenodo10*, 5281, 2018.
- [65] Chanwoo Kim, Soham U Gadgil, Alex J DeGrave, Jesutofunmi A Omiye, Zhuo Ran Cai, Roxana Daneshjou, and Su-In Lee. Transparent medical image ai via an image–text foundation model grounded in medical literature. *Nature Medicine*, pages 1–12, 2024.

- [66] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR, 2021.
- [67] 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.
- [68] Mikhail Kulyabin, Aleksei Zhdanov, Anastasia Nikiforova, Andrey Stepichev, Anna Kuznetsova, Mikhail Ronkin, Vasilii Borisov, Alexander Bogachev, Sergey Korotkich, Paul A. Constable, and Andreas Maier. Octdl: Optical coherence tomography dataset for image-based deep learning methods. *Scientific Data*, 11(1), April 2024.
- [69] 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.
- [70] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*, 2023.
- [71] 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.
- [72] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024.
- [73] 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.
- [74] Junkai Li, Siyu Wang, Meng Zhang, Weitao Li, Yunghwei Lai, Xinhui Kang, Weizhi Ma, and Yang Liu. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*, 2024.

- [75] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.
- [76] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.
- [77] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.
- [78] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.
- [79] 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.
- [80] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Melissa Zhao, Aaron K Chow, Kenji Ikemura, Ahronng Kim, Dimitra Pouli, Ankush Patel, et al. A multimodal generative ai copilot for human pathology. *Nature*, pages 1–3, 2024.
- [81] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.
- [82] Niccolò Marini, Stefano Marchesin, Sebastian Otálora, Marek Wodzinski, Alessandro Caputo, Mart Van Rijnthoven, 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.
- [83] Sachin Mehta, Ximing Lu, Wenjun Wu, Donald Weaver, Hannaneh Hajishirzi, Joann G Elmore, and Linda G Shapiro. End-to-end diagnosis of breast biopsy images with transformers. *Medical image analysis*, 79:102466, 2022.
- [84] Xueyan Mei, Zelong Liu, Philip M Robson, Brett Marinelli, Mingqian Huang, Amish Doshi, Adam Jacobi, Chendi Cao, Katherine E Link, Thomas Yang, et al. Radimagenet: an open radiologic deep learning research dataset for effective transfer learning. *Radiology: Artificial Intelligence*, 4(5):e210315, 2022.

- [85] 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.
- [86] Jesper Molin, Morten Fjeld, Claudia Mello-Thoms, and Claes Lundström. Slide navigation patterns among pathologists with long experience of digital review. *Histopathology*, 67(2):185–192, 2015.
- [87] 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.
- [88] Junya Morita, Kazuhisa Miwa, Takayuki Kitasaka, Kensaku Mori, Yasuhito Suenaga, Shingo Iwano, Mitsuru Ikeda, and Takeo Ishigaki. Interactions of perceptual and conceptual processing: Expertise in medical image diagnosis. *International Journal of Human-Computer Studies*, 66(5):370–390, 2008.
- [89] 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.
- [90] Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, Dung D Le, Chi M Pham, Hang TT Tong, Diep H Dinh, et al. Vindr-cxr: An open dataset of chest x-rays with radiologist’s annotations. *Scientific Data*, 9(1):429, 2022.
- [91] Hieu T Nguyen, Ha Q Nguyen, Hieu H Pham, Khanh Lam, Linh T Le, Minh Dao, and Van Vu. Vindr-mammo: A large-scale benchmark dataset for computer-aided diagnosis in full-field digital mammography. *Scientific Data*, 10(1):277, 2023.
- [92] 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*, 5:1005086, 2022.
- [93] 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.
- [94] R OpenAI. Gpt-4 technical report. *arXiv*, pages 2303–08774, 2023.

- [95] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *ECCV*, 2020.
- [96] 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.
- [97] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives, 2020.
- [98] 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.
- [99] 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.
- [100] 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.
- [101] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.
- [102] Rubaba Binte Rahman, Sharia Arfin Tanim, Nazia Alfaz, Tahmid Enam Shrestha, M Saef Ullah Miah, and Firoz Mridha. Dental OPG XRAY Dataset, 2024.
- [103] Lucia Roa-Peña, Francisco Gómez, and Eduardo Romero. An experimental study of pathologist’s navigation patterns in virtual microscopy. *Diagnostic pathology*, 5:1–11, 2010.
- [104] 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.
- [105] Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. Automatic keyword extraction from individual documents. *Text mining: applications and theory*, pages 1–20, 2010.

- [106] Joel Saltz, Rajarsi Gupta, Liang Hou, Tahsin Kurc, Parul Singh, Vu Nguyen, Dimitris Samaras, Kenneth R. Shroyer, Ting Zhao, Robert Batiste, Jonathan Van Arnam, The Cancer Genome Atlas Research Network, Ilya Shmulevich, Aniruddha U. K. Rao, Alexander J. Lazar, Arvind Sharma, and Vesteinn Thorsson. Tumor-infiltrating lymphocytes maps from tcga h&e whole slide pathology images, 2018. Data set.
- [107] 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.
- [108] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.
- [109] 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.
- [110] 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.
- [111] 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.
- [112] Danli Shi, Weiyi Zhang, Jiancheng Yang, Siyu Huang, Xiaolan Chen, Mayinuer Yusufu, Kai Jin, Shan Lin, Shunming Liu, Qing Zhang, and Mingguang He. Eyeclip: A visual-language foundation model for multi-modal ophthalmic image analysis, 2024.
- [113] Rebecca L Siegel, Angela N Giaquinto, and Ahmedin Jemal. Cancer statistics, 2024. *CA: a cancer journal for clinicians*, 74(1):12–49, 2024.
- [114] 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.
- [115] 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.

- [116] Sanjay Subramanian, Lucy Lu Wang, Sachin Mehta, Ben Bogin, Madeleine van Zuylen, Sravanthi Parasa, Sameer Singh, Matt Gardner, and Hannaneh Hajishirzi. Mediat: A dataset of medical images, captions, and textual references, 2020.
- [117] 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.
- [118] 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.
- [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] 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.
- [122] Paul Voigtlaender, Soravit Changpinyo, Jordi Pont-Tuset, Radu Soricut, and Vittorio Ferrari. Connecting Vision and Language with Video Localized Narratives. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.
- [123] 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.
- [124] Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature medicine*, pages 1–12, 2024.

- [125] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023.
- [126] 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.
- [127] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversations. 2024.
- [128] 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.
- [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] Yunfei Xie, Ce Zhou, Lang Gao, Juncheng Wu, Xianhang Li, Hong-Yu Zhou, Sheng Liu, Lei Xing, James Zou, Cihang Xie, and Yuyin Zhou. Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine, 2024.
- [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] Jiarui Xu, Xingyi Zhou, Shen Yan, Xiuye Gu, Anurag Arnab, Chen Sun, Xiaolong Wang, and Cordelia Schmid. Pixel-aligned language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13030–13039, 2024.
- [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] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16375–16387, 2022.

- [135] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34:23634–23651, 2021.
- [136] Heyu Zhang, Yan He, Xiaomin Wu, Peixiang Huang, Wenkang Qin, Fan Wang, Juxiang Ye, Xirui Huang, Yanfang Liao, Hang Chen, et al. Pathnarratives: Data annotation for pathological human-ai collaborative diagnosis. *Frontiers in Medicine*, 9:1070072, 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] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.
- [139] 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.
- [140] 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.
- [141] 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.
- [142] Hong-Yu Zhou, Subathra Adithan, Julián Nicolás Acosta, Eric J Topol, and Pranav Rajpurkar. A generalist learner for multifaceted medical image interpretation. *arXiv preprint arXiv:2405.07988*, 2024.
- [143] 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.
- [144] 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.