Investigating Zero-Shot Diagnostic Pathology in Vision-Language Models with Efficient Prompt Design

Vasudev Sharma

A Thesis

in

The Department

of

Computer Science and Software Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Science (Computer Science) at

Concordia University

Montréal, Québec, Canada

April 2025

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Vasudev Sharma**

Entitled: **Investigating Zero-Shot Diagnostic Pathology in Vision-Language Models with Efficient Prompt Design**

and submitted in partial fulfillment of the requirements for the degree of

Master of Science (Computer Science)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

_____ Chair
Dr. Yang Wang

_____ Examiner
Dr. Yang Wang

_____ Examiner
Dr. Thomas Fevens

_____ Supervisor
Dr. Mahdi S. Hosseini

Approved by _____
Dr. Charalambos Poullis, Chair
Department of Computer Science and Software Engineering

_____ 2025

_____
Dr. Mourab Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Investigating Zero-Shot Diagnostic Pathology in Vision-Language Models with Efficient Prompt Design

Vasudev Sharma

Vision-Language Models (VLMs) have emerged as powerful tools in computational pathology, offering the ability to perform zero-shot diagnostic inference on gigapixel whole slide images (WSIs). However, a core challenge remains: these models exhibit high sensitivity to the linguistic structure and specificity of prompts, which can significantly impact diagnostic accuracy, reproducibility, and clinical interpretability. This thesis systematically investigates the role of prompt engineering in enhancing the diagnostic performance of VLMs in histopathology. We propose a structured prompt engineering framework that modulates four critical dimensions: anatomical precision, information density, instructional framing, and output constraints, to evaluate their effect on model behavior. Using a clinically validated in-house dataset of 3,507 digestive system WSIs spanning multiple tissue types and pathological conditions, we conduct a comprehensive evaluation of four state of the art VLMs called Biomedical Contrastive Language-Image Pre-training (BioMedCLIP), Quilt-Net, Quilt-Large Language and Vision Assistant (Quilt-LLAVA), and Contrastive Learning from captions for Histopathology (CONCH). Our methodology includes a combination of quantitative assessments using Area Under the Curve (AUC) analysis , Receiver Operating Characteristic (ROC) and qualitative analyses to understand how prompt design influences diagnostic inference, interpretability, and generalization across tissues.

Our results demonstrate that prompt formulation significantly affects model performance across the full dataset. In particular, prompts that incorporate high anatomical specificity and clear instructional framing yield consistent improvements in classification accuracy across multiple tissue sites. The study further reveals that domain aligned prompting strategies are often more effective than

increase in architectural complexity, highlighting the centrality of human and AI communication in medical vision-language tasks. In addition to empirical findings, we contribute actionable guidelines for implementing VLMs in clinical computational pathology workflows, emphasizing prompt standardization and interpretability. This work shifts the emphasis from purely architectural innovation to optimizing the language mediated interface between human expertise and AI systems, thereby enhancing both diagnostic performance and clinical utility in zero-shot medical image analysis.

Acknowledgments

As I reach this milestone in my academic journey, I find myself reflecting on all the people who made this thesis possible. This work represents not just my efforts, but the collective support of many incredible individuals who stood by me throughout this challenging process.

I express my deepest gratitude to my supervisor, Dr. Mahdi S. Hosseini, whose guidance and expertise were instrumental in shaping this research. Your patience when I struggled with complex concepts, your insightful feedback that constantly pushed me to think deeper, and your unwavering support even when results weren't coming easily all of these made a world of difference. Thank you for believing in my capabilities and helping me navigate the fascinating world of research.

I am also incredibly grateful to Dr. Ahmed Alagha for his valuable input and assistance throughout this project. His thoughtful suggestions and willingness to share his knowledge significantly enhanced the quality of this work and broadened my understanding of the field. To my amazing peers Denisha Thakar, Damien Martins Gomes, Ali Nasiri Sarvi, Ali Affan and Guntas Singh , thank you for making this journey so much more than just academic work.

I owe everything to my family, who supported me in ways that cannot be fully expressed in words. My mother Aradhna Sharma, your daily check-ins and constant encouragement kept me going even when I doubted myself. Your ability to listen patiently to my technical rambles, despite not understanding much of it, shows your incredible love and support. My father Sumeet Sharma, inculcated in me the work ethic and perseverance that were essential for completing this thesis. His quiet confidence in my abilities never wavered, even when mine did. Thank you both for the sacrifices you have made to support my education and dreams.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Target Search and Localization: An Example

Computational pathology represents a transformative frontier in the integration of artificial intelligence (AI) with medical diagnostics, particularly in the analysis of histopathological images. This emerging field leverages powerful algorithmic approaches to analyze microscopic tissue specimens traditionally examined by human pathologists. The evolution of computational methods in pathology mirrors the broader trajectory of medical image analysis from rule-based systems to sophisticated deep learning architectures capable of identifying complex morphological patterns indicative of disease states (Gurcan et al., 2009; Hosseini et al., 2024). Consider the diagnostic process of identifying invasive colorectal adenocarcinoma in a WSI. A pathologist examines tissue at multiple magnifications, recognizes cellular atypia, identifies aberrant glandular structures permeating through the muscularis mucosa, and integrates these observations with the anatomical context to render a diagnosis. This complex cognitive process has traditionally been difficult to replicate algorithmically. Early computational approaches using convolutional neural networks (CNNs) achieved modest success in isolated classification tasks (K. He, Zhang, Ren, & Sun, 2015; Krizhevsky, Sutskever, & Hinton, 2012) but struggled with the integrative reasoning that characterizes expert diagnosis.

The recent emergence of large language models (LLMs) and VLMs represents a paradigm shift

in computational pathology by enabling cross-modal reasoning between visual patterns and natural language descriptions. Models such as CLIP (Radford et al., 2021a), Florence (Yuan et al., 2021), DALL-E (Ramesh et al., 2021), LLaVa (H. Liu, Li, Wu, & Lee, 2023) along with specialized histopathology models like Med-PaLM (Singhal et al., 2023) ,Quilt-LLAVA (Seyfioglu, Ikezogwo, Ghezloo, Krishna, & Shapiro, 2025), LLaVA-Med (C. Li et al., 2023) and CONCH (Lu, Chen, Williamson, et al., 2023) established meaningful connections between visual inputs and textual instructions or descriptions. This capability is particularly valuable in histopathology, where expert knowledge is encoded in natural language while diagnostic evidence manifests as visual patterns across multiple biological scales.

Returning to our colorectal adenocarcinoma example, these advanced VLMs can now process instructions like ”Identify regions of invasive adenocarcinoma infiltrating the muscularis mucosa” and locate relevant visual patterns within gigapixel WSIs. This represents a fundamental advancement in creating more intuitive interfaces for computational pathology systems. However, the effectiveness of these models depends critically on how the diagnostic task is linguistically framed a relationship that remains poorly characterized despite its central importance to clinical implementation.

The sensitivity of VLMs to prompt formulation presents both opportunities and challenges. On one hand, careful prompt engineering could potentially enhance model accuracy by directing attention to diagnostically relevant features and incorporating appropriate anatomical and histological context. On the other hand, this sensitivity introduces variability that may undermine reproducibility in clinical settings if not systematically addressed. This tension motivates our research into optimal prompt engineering strategies for computational pathology. Our study focuses specifically on digestive system pathology for several compelling reasons. First, digestive cancers represent a significant global health burden, with colorectal cancer alone ranking as the third most common malignancy worldwide (Hossain et al., 2022). Second, the morphological manifestations of invasive cancer across different digestive tissues (from esophagus to colon) exhibit substantial variation, providing an ideal testbed for evaluating the generalization capabilities of VLMs. Third, the digestive system presents diverse histological patterns across normal, inflammatory, pre-malignant, and malignant states, requiring models to distinguish subtle differences in tissue architecture that have

significant diagnostic implications.

By systematically analyzing how prompt construction affects diagnostic accuracy across tissue types, we establish evidence based guidelines for integrating VLMs into clinical pathology workflows. This research addresses the underexplored human and AI interface in computational pathology, where existing literature has emphasized architectural innovations while overlooking how diagnostic tasks are communicated to models. Our work demonstrates that optimizing anatomical precision, information density, and instructional framing in prompts significantly enhances zero-shot diagnostic capabilities for both cancer and dysplasia detection.

1.2 Problem Statement and Research Questions

Despite significant advances in computational pathology, VLMs face critical challenges when applied to zero-shot diagnostic tasks in histopathology. The fundamental problem addressed in this thesis is the insufficient understanding of how prompt design influences diagnostic performance in VLM based computational pathology systems. This knowledge gap manifests in several interconnected dimensions that impede the reliable deployment of these models in clinical settings.

(1) VLMs demonstrate remarkable sensitivity to the specific wording, structure, and framing of prompts (J. Gu et al., 2023; Radford et al., 2021a; Sahoo et al., 2025). In clinical diagnostic contexts, where precise interpretation is paramount, this sensitivity creates significant challenges. Minor variations in prompt formulation can lead to substantial differences in diagnostic output (J. Wang et al., 2024). The optimal linguistic strategies for prompting medical VLMs remain poorly characterized compared to general domain applications, and standardization of prompt methodologies across different pathological contexts is lacking, limiting reproducibility.

(2) Histopathological WSIs present unique computational challenges due to their gigapixel dimensions, which typically exceed the input resolution capabilities of current VLM architectures (R. J. Chen et al., 2022; Tellez et al., 2019). Additionally, these images contain relevant information at multiple scales, from tissue architecture to cellular and subcellular details,

requiring models to effectively integrate information across magnification levels. The complex spatial relationships between tissue structures carry significant diagnostic importance, necessitating specialized approaches to image processing and prompt design that can effectively guide model attention across multiple scales while maintaining diagnostically relevant context.

(3) Pathological diagnosis relies heavily on specialized anatomical knowledge and contextual information that general-purpose VLMs may not adequately capture. Normal histological variations across different anatomical sites must be distinguished from pathological changes, tissue-specific diagnostic criteria require precise anatomical referencing in prompts, and the relationship between anatomical precision in prompts and diagnostic accuracy remains poorly characterized. The optimal level of anatomical specificity, information density, and instructional framing required for reliable diagnosis has not been systematically investigated, creating uncertainty in how to effectively prompt VLMs for specialized medical diagnostic tasks.

Based on these interconnected challenges, we formulate the following specific research questions:

(1) How does the anatomical precision in prompts (ranging from tissue-agnostic to highly specific anatomical referencing) affect diagnostic accuracy across different tissue types and VLM architectures.

(2) What is the relationship between information density in prompts and model performance, and is there an optimal level of detail that maximizes diagnostic accuracy.

(3) How does instructional framing (including expert role assignment, task formulation, and query structure) influence model attention to diagnostically relevant features.

(4) To what extent do output constraints affect the consistency and reliability of diagnostic predictions across different VLM architectures.

(5) Can systematic prompt engineering approaches improve zero-shot diagnostic performance on rare or unusual pathological presentations where annotated training data is limited.

(6) How does model complexity (in terms of architecture, parameter count, and training approach) interact with prompt engineering strategies to influence diagnostic performance across different pathological conditions.

Our investigation of these questions employs a systematic approach that varies key dimensions of prompt design across multiple state of the art VLM architectures specifically (BioMedCLIP (Zhang et al., 2025), Quilt-Net (Ikezogwo et al., 2025), Quilt-LLAVA (Seyfioglu et al., 2025), and CONCH (Lu, Chen, Williamson, et al., 2023)), evaluating performance on a large, clinically validated dataset of digestive system WSIs spanning multiple tissue types and pathological conditions.

This comprehensive approach addresses a critical gap in current research, which has predominantly focused on architectural innovations rather than optimization of the human and AI interaction interface represented by prompt engineering. By systematically characterizing the relationship between prompt formulation and diagnostic accuracy, we aim to establish a foundation for more robust, reproducible, and clinically viable computational pathology systems.

1.3 Research Objectives and Contributions

This thesis aims to systematically evaluate and optimize prompt engineering strategies for vision-language models in computational pathology through a structured investigation that addresses the complex interplay between linguistic formulation and diagnostic performance. Our research is guided by the following primary objectives:

(1) We seek to quantify the impact of prompt variations along four critical dimensions called anatomical precision, information density, instructional framing, and output constraints on diagnostic accuracy across different VLM architectures. This objective involves the development of a comprehensive prompt engineering framework that systematically varies these dimensions and evaluates their influence on model performance across diverse tissue types and pathological conditions. By establishing the relative importance of these factors and their interactions, we aim to provide a robust empirical foundation for prompt design in computational pathology.

(2) We aim to identify optimal prompt structures for different tissue types and pathological conditions within digestive system pathology. This objective recognizes that the optimal prompt formulation may vary based on anatomical context, pathological entity, and specific diagnostic task. By analyzing how prompt effectiveness varies across these contexts, we seek to develop tissue-specific and task-specific guidelines that maximize diagnostic accuracy in diverse clinical scenarios.

(3) We seek to analyze attention map patterns from different VLM architectures to assess diagnostic relevance and model interpretability across varying prompt conditions. This objective addresses not only what prompt structures optimize performance but also how they influence the model's focus on diagnostically relevant regions within the image. By correlating attention patterns with expert-annotated regions of interest, we aim to enhance model interpretability and provide insights into the mechanisms by which prompt variations influence model behavior.

(4) We aim to develop practical guidelines for implementing VLMs in clinical diagnostic workflows based on empirical evidence from our systematic investigation. This objective translates our technical findings into actionable recommendations for researchers, developers, and potentially clinicians integrating VLM-based systems into pathology practice. These guidelines will address considerations for prompt design, model selection, and implementation strategies that optimize diagnostic performance while maintaining reproducibility and reliability.

The successful execution of these objectives contributes several significant advances to the field of computational pathology:

(1) We provide a comprehensive prompt engineering framework specifically designed for computational pathology that systematically addresses anatomical precision, information density, instructional framing, and output constraints. This framework fills a critical gap in current research by establishing a structured approach to prompt design in medical vision-language applications. Unlike ad hoc approaches that have characterized much of the existing work in this domain, our framework enables systematic optimization and reproducible implementation of prompt engineering strategies.

(2) We offer robust empirical evidence demonstrating the relationship between prompt formulation and diagnostic accuracy across diverse tissue types within the digestive system. This evidence illuminates how subtle variations in prompt structure can significantly impact model performance, providing a quantitative basis for understanding the sensitivity of VLMs to linguistic formulation in medical applications. Our findings help establish which aspects of prompt design are most critical for optimizing performance in specific diagnostic contexts.

(3) We provide novel insights into the attention mechanisms of different VLM architectures when processing histopathological images under various prompt conditions. By analyzing how attention patterns correlate with diagnostically relevant regions and how they change in response to prompt variations, we enhance understanding of model behavior and interpretability. These insights contribute to the development of more transparent and trustworthy AI systems for clinical applications.

(4) We establish practical guidelines for optimizing zero-shot diagnostic performance in computational pathology applications. These guidelines translate our experimental findings into actionable recommendations for researchers and developers implementing VLM based systems in pathology workflows. By providing evidence-based strategies for prompt design across different anatomical contexts and diagnostic tasks, we enhance the clinical utility and reliability of computational pathology systems.

The significance of these contributions extends beyond technical optimization to address fundamental challenges in the clinical implementation of AI systems in pathology. By enhancing the accuracy, reproducibility, and interpretability of VLM based diagnostic approaches, our research supports the development of more reliable and clinically relevant tools that can potentially improve diagnostic efficiency and accuracy in both common and rare pathological conditions. As computational pathology continues to evolve toward clinical integration, understanding how to effectively communicate diagnostic tasks through natural language prompts becomes increasingly important for human and AI collaboration in diagnostic workflows.

1.4 Thesis Organization

This thesis is organized into four chapters that progressively build a comprehensive understanding of vision-language models in computational pathology, with particular emphasis on prompt engineering for zero-shot diagnostic applications. The structure reflects the logical progression from foundational concepts to experimental methodology, results, and broader implications.

Chapter 1 establishes the research context and motivation through an illustrative example of target search in computational pathology. It introduces the fundamental problem addressed in this thesis, the insufficient understanding of how prompt design influences diagnostic performance in VLM based pathology systems and articulates the specific research questions that guide our investigation. The chapter outlines the primary objectives and contributions of the research, highlighting the development of a comprehensive prompt engineering framework, empirical evidence on prompt performance relationships, insights into attention mechanisms, and practical guidelines for clinical implementation.

Chapter 2 provides a comprehensive review of the technological and conceptual foundations of computational pathology, with emphasis on the evolution towards vision-language integration. The chapter begins with an overview of computational pathology as a discipline, including its historical development, key applications, and technical challenges. It then examines self-supervised learning approaches in computational pathology, including contrastive learning methods like CLIP (Radford et al., 2021a) and non-contrastive techniques (K. He et al., 2021). The chapter explores the fundamental architecture of transformer models (Vaswani et al., 2023) and their adaptation to vision tasks through Vision Transformers (ViT) (Dosovitskiy et al., 2021), with particular attention to the challenges of processing gigapixel histopathological images. It then discusses the emergence of foundation models in the biomedical domain and their impact on computational pathology, followed by an in depth examination of VLMs, including their architecture, training methodologies, and applications to histopathology. The chapter concludes with a detailed review of prompt engineering approaches and their current applications in computational pathology, identifying key knowledge gaps that motivate our research.

Chapter 3 presents the core experimental methodology and results of our investigation into zero-shot diagnostic pathology with vision-language models. The chapter begins with a detailed description of the four state of the art VLM architectures evaluated in our study: BioMedCLIP (Zhang et al., 2025), Quilt-Net (Ikezogwo et al., 2025), Quilt-LLAVA (Seyfioglu et al., 2025), and CONCH (Lu, Chen, Williamson, et al., 2023). It then characterizes our in house dataset of 3,507 digestive system WSIs, describing the distribution of tissue types, pathological conditions, and annotation methodology. The chapter details our prompt engineering framework, which systematically varies anatomical precision, information density, instructional framing, and output constraints across nine template structures. The results section presents our findings through quantitative performance metrics (AUC scores) across different prompt configurations and model architectures, complemented by qualitative analysis of attention maps that visualize how different prompts influence model focus on diagnostically relevant regions. The chapter includes a comprehensive ablative study that isolates the impact of each prompt dimension on model performance, as well as an analysis of how model performance varies across different tissue types and magnification levels. The discussion interprets these findings in the context of current knowledge and their implications for clinical implementation.

Chapter 4 synthesizes the key findings of our research and explores their broader implications for computational pathology and medical AI. The chapter begins by summarizing the major empirical results regarding the relationship between prompt design and diagnostic performance, highlighting the critical importance of anatomical precision and appropriate information density across different VLM architectures. It then discusses the theoretical implications of these findings for understanding vision-language integration in specialized medical domains. The chapter outlines practical guidelines for implementing prompt engineering in clinical computational pathology workflows, addressing considerations for different tissue types, diagnostic tasks, and model architectures. It acknowledges the limitations of the current study and identifies promising directions for future research, including extensions to other organ systems, more complex diagnostic tasks, and integration with multimodal data sources. The chapter concludes by situating our contributions within the broader evolution of computational pathology toward more interpretable, trustworthy, and clinically relevant AI systems that can enhance diagnostic accuracy and efficiency.

Throughout this organizational structure, we maintain a focus on the central research questions while providing sufficient context, methodological details, and interpretive discussion to establish the significance of our findings for both technical research and clinical practice. Each chapter builds upon the previous ones to create a cohesive narrative that advances understanding of how prompt engineering can optimize VLMs for computational pathology applications.

Chapter 2

Background and Literature Review

2.1 Background

This section establishes the technical foundations underlying computational pathology and AI driven approaches to histopathological image analysis. We examine the core principles, methodological frameworks, and architectural innovations that have enabled significant advances in this field, from basic image processing to sophisticated multimodal reasoning systems. The discussion progresses from computational pathology fundamentals through self-supervised learning methods, transformer architectures, foundation models, and vision-language integration, providing the necessary context for understanding our research on prompt engineering in diagnostic applications.

2.1.1 Computational Pathology

Computational Pathology represents an interdisciplinary field that integrates traditional pathology with advanced computational methods to extract meaningful information from pathological data. This emerging discipline extends beyond machine learning applications to encompass the entire digital transformation of pathology practice (Hosseini et al., 2024). By combining expertise from pathology, computer science, and biomedical engineering, computational pathology aims to develop tools and methodologies that enhance diagnostic accuracy, improve prognostication, and guide personalized treatment decisions. The foundation of computational pathology lies in the digitization of pathology specimens through WSI technology. Modern WSI scanners convert

glass slides into high resolution digital images, generating gigapixel sized representations that can be stored, shared, and analyzed computationally (Ehteshami Bejnordi et al., 2017). This digital transformation enables pathologists to access slides remotely, collaborate across institutions, and implement quality assurance measures that were previously challenging with physical slides. However, the digitization process introduces technical challenges related to storage infrastructure, color standardization, and image quality control (Tellez et al., 2019). Beyond simple digitization, computational pathology encompasses diverse analytical approaches. Image analysis algorithms, ranging from traditional computer vision techniques to modern deep learning models, can segment tissue components, detect cellular structures, and quantify morphological features with high precision and reproducibility (Litjens et al., 2017). These methods standardize measurements that were historically subjective, such as nuclear pleomorphism assessment, mitotic counting, and tumor infiltrating lymphocyte quantification (Veta et al., 2019). The resulting quantitative data provides objective metrics that can supplement pathologists qualitative assessments and reduce inter-observer variability.

A key strength of computational pathology is its capacity for multimodal data integration. Modern approaches combine histopathological images with molecular data (genomics, proteomics, transcriptomics), radiological findings, and clinical information to develop comprehensive models of disease (R. J. Chen et al., 2021). This integration reveals connections between morphological patterns and underlying molecular mechanisms, enabling more precise disease classification and treatment selection. For example, recent studies have demonstrated that computational analysis of Hematoxylin and Eosin (H&E) slides can predict genetic mutations and molecular subtypes that traditionally require expensive specialized testing (B. He et al., 2020). Workflow enhancement represents another significant benefit of computational pathology. Automated screening and triaging systems can prioritize cases requiring urgent review, while specialized algorithms can assist with time-consuming tasks like microorganism detection and immunohistochemistry scoring (Fuchs & Buhmann, 2011). These workflow improvements address practical challenges in pathology practice, including increasing case complexity, growing subspecialty requirements, and declining pathologist workforce numbers in many regions. By automating routine aspects of slide review, computational

tools allow pathologists to focus their expertise on complex diagnostic decisions and clinically relevant interpretations. The clinical impact of computational pathology has been demonstrated in several domains. The CAMELYON challenge series showed that algorithmic detection of breast cancer metastases in lymph nodes can match or exceed pathologist performance under certain conditions (Ehteshami Bejnordi et al., 2017). Similar successes have been reported in prostate cancer grading, dermatopathology, and hematopathology applications. These achievements highlight the potential for computational methods to serve as diagnostic decision support tools, particularly for standardized tasks with well-defined endpoints (Campanella et al., 2019).

Despite these advances, computational pathology faces several implementation challenges. Technical barriers include the need for standardized image acquisition protocols, robust validation methodologies, and interoperable software systems (Fuchs & Buhmann, 2011). Regulatory frameworks for computational pathology tools are still evolving, with regulatory bodies developing guidelines for clinical validation and quality assurance. Additionally, integration into existing laboratory information systems and electronic health records remains complex (Tizhoosh & Pantanowitz, 2018). Financial considerations also influence adoption, as laboratories must invest in digital infrastructure while navigating uncertain reimbursement models. Perhaps most importantly, cultural adaptation requires pathologists and laboratory professionals to develop new skills and adjust established workflows. Educational initiatives and collaborative research programs play crucial roles in addressing these challenges by building capacity and demonstrating practical benefits in diverse clinical settings (Tizhoosh & Pantanowitz, 2018). Looking forward, computational pathology is expanding beyond traditional histopathology to encompass spatial omics technologies, digital cytology, electron microscopy, and other specialized domains. Federated learning approaches are addressing data sharing concerns by enabling model training across institutions without transferring sensitive information (Lu, Kong, et al., 2020). Explainable AI methods are improving transparency and interpretability, which are essential for clinical adoption and regulatory approval (M., V., S., & H., 2020). Computational pathology is evolving from research into clinical practice, transforming traditional microscopy into a quantitative, data-enriched field. Rather than replacing pathologists, computational tools augment human expertise by providing consistent measurements, revealing subtle patterns,

and integrating complex multimodal data. This human-AI partnership forms the foundation of precision pathology, optimizing diagnostic accuracy and personalized treatment selection.

Computational pathology is evolving alongside broader healthcare technology trends, requiring advanced computational methods to handle emerging imaging technologies like multiplex immunofluorescence and spatial transcriptomics. Its applications extend to underserved regions with pathologist shortages and are being integrated into educational programs. These developments represent a fundamental paradigm shift in pathology practice rather than merely a technological addition.

2.1.2 Self-Supervised Learning in Computational Pathology

Self-supervised learning represents a fundamental paradigm shift in how computational systems learn from medical imaging data, particularly in the context of digital pathology (Koohbanani, Unnikrishnan, Khurram, Krishnaswamy, & Khalifa, 2021). Rather than relying on human-annotated labels, which are time-consuming and expensive to obtain, self-supervised approaches extract meaningful patterns directly from the data itself. This concept mirrors the natural learning process of pathology trainees, who develop pattern recognition abilities by examining numerous tissue specimens before focusing on specific diagnostic criteria (L. Chen et al., 2022). At its core, self-supervised learning creates artificial learning tasks from unlabeled data, generating supervisory signals that encourage models to understand the underlying structure of histopathological images (R. J. Chen & Krishnan, 2022). These pretext tasks might involve reconstructing masked portions of tissue images (K. He et al., 2021), predicting the spatial relationships between tissue regions, or identifying matching augmented views of the same specimen (Ciga, Xu, & Martel, 2021). Through solving these tasks, models develop rich representations of tissue morphology, cellular architecture, and pathological features without explicit annotation. The relevance of self-supervised learning to computational pathology stems from several inherent characteristics of histopathological data. First, digital pathology generates vast repositories of unlabeled whole slide images that contain valuable information but lack detailed annotations (Lu, Williamson, et al., 2020). Second, histopathological images exhibit complex hierarchical organization across multiple scales, from subcellular structures to tissue architecture. Self-supervised approaches naturally accommodate this multi-scale nature by

learning representations at different levels of magnification (R. J. Chen et al., 2022). Third, pathology images display significant variability in staining, preparation, and scanning protocols across institutions. Self-supervised learning methods develop more robust representations that can generalize across these technical variations (Hou et al., 2019). Several conceptual frameworks have emerged within self-supervised learning for pathology image analysis. Contrastive learning approaches build representations by comparing different views of tissue, encouraging similar embeddings for augmented versions of the same tissue while pushing apart representations of different tissues (Ciga et al., 2021). Non-contrastive methods focus on internal consistency within images without relying on explicit negative examples (Zbontar, Jing, Misra, LeCun, & Deny, 2021). Masked modeling techniques randomly obscure portions of images and train models to predict the missing content, forcing them to understand the contextual relationships within tissue structures (Xie et al., 2022). Knowledge distillation frameworks transfer information between different model components to enhance representation quality (Caron et al., 2021).

The hierarchical organization of pathology data has inspired multi-scale self-supervised approaches that simultaneously capture information at cellular, local tissue pattern, and global contextual levels (R. J. Chen et al., 2022). This capability aligns with diagnostic workflows in which pathologists examine specimens at progressively higher magnifications to integrate information across scales. Models that preserve this hierarchical perspective can maintain both the fine details necessary for cellular classification and the broader context essential for disease characterization (X. Wang et al., 2022). Integration of self-supervised learning with other methodological approaches creates powerful hybrid frameworks for computational pathology. Combining self-supervised representations with weakly-supervised methods enables more effective learning from slide-level labels (Lu, Williamson, et al., 2020). Graph based approaches incorporate spatial relationships between tissue regions, maintaining the topological information critical for accurate diagnosis (Gadiya, Anand, & Sethi, 2019; Jaume et al., 2021; Zhou et al., 2019). Multi-modal techniques integrate histopathological images with complementary data sources to provide more comprehensive characterization of tissue specimens (R. J. Chen et al., 2021). Recent innovations in self-supervised learning include multi-instance contrastive approaches that leverage slide-level

heterogeneity (Lu, Williamson, et al., 2020), patch-based pretraining strategies that preserve spatial relationships (X. Wang et al., 2022), and domain-adaptive frameworks that mitigate staining variation across institutions (Hou et al., 2019). These techniques address pathology specific challenges like sparse annotations and technical variability. Federated self-supervised frameworks have emerged to enable collaborative model training across institutions while preserving patient privacy and data sovereignty (Lu, Kong, et al., 2020). Integration with explainable AI methods further enhances these models clinical utility by providing interpretable visual evidence for their predictions (Jaume et al., 2021). As self-supervised methodologies mature, it becomes more instrumental in developing robust computational pathology systems .

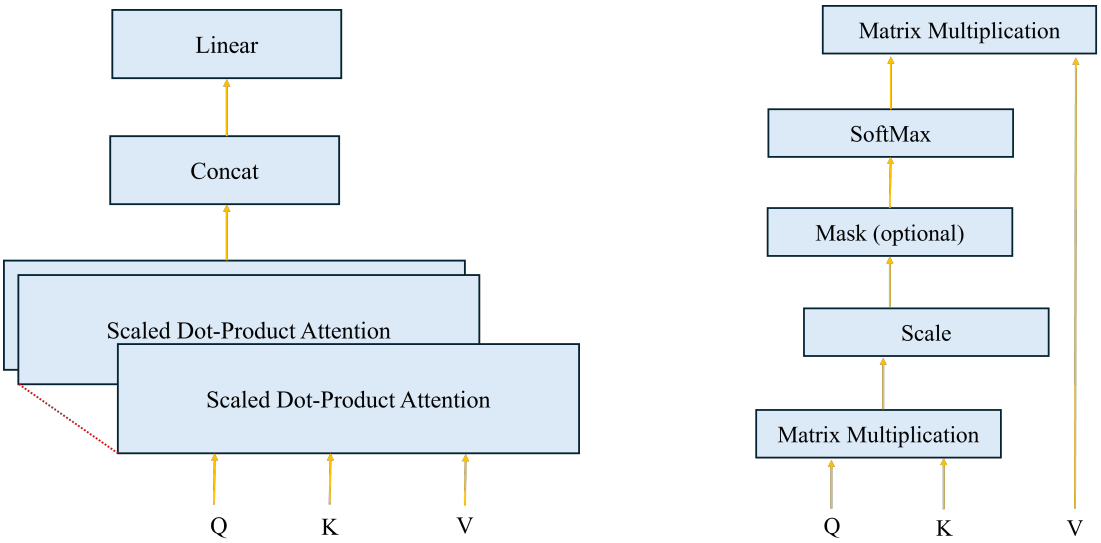
2.1.3 Transformers

Transformer architectures represent a pivotal advancement in neural network design that has fundamentally altered the landscape of machine learning research and applications across domains. Initially introduced by (Vaswani et al., 2023) for natural language processing tasks, these architectures have subsequently demonstrated remarkable efficacy in computer vision, multimodal learning, and computational pathology. The fundamental innovation of the transformer lies in its exclusive reliance on attention mechanisms to model relationships between elements in a sequence, eschewing the recurrent and convolutional operations that characterized previous architectural paradigms. The core architectural innovation of the transformer is the self-attention mechanism, which enables each element in a sequence to attend to all other elements, thereby capturing long-range dependencies with constant path length between any two positions. This contrasts with recurrent neural networks (RNNs), which process sequences serially and suffer from vanishing gradient problems when modeling long range dependencies, and CNNs, which capture local dependencies through spatially limited receptive fields. The multi-head attention design enables specialized pathology feature extraction across diverse tissue morphologies, capturing both local cellular patterns and long-range architectural dependencies critical for accurate histopathological analysis (R. J. Chen et al., 2022; Kirilenko, Andreychuk, Panov, & Yakovlev, 2022). Formally, the self-attention mechanism computes attention scores between query vectors Q and key vectors K, which determine the influence of each key on the output, followed by a weighted aggregation of value vectors V. This operation

is expressed mathematically as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where d_k represents the dimensionality of the key vectors and serves as a scaling factor to prevent excessively large attention scores in high-dimensional spaces. The original transformer enhances this mechanism through multi-head attention, which projects the queries, keys, and values into multiple subspaces, computes attention within each subspace, and subsequently concatenates the results as seen in Fig. 2.1. This approach enables the model to attend to information from different representation subspaces, thereby capturing diverse relationships within the data. The computational efficiency of the transformer derives from its parallelization during training.



(a) Multi-Head attention mechanism (b) Attention mechanism

Figure 2.1: Detailed overview of the multi-head attention mechanism

Unlike RNNs, which process sequences serially, transformers compute attention scores between all elements simultaneously, enabling efficient parallelization on modern hardware accelerators. This parallelization, coupled with the architecture's expressiveness, has facilitated the development of increasingly large-scale models with millions or billions of parameters, contributing to the emergence of foundation models. The quadratic computational complexity of self-attention with respect to sequence length ($O(n^2)$ for a sequence of length n) initially posed challenges for processing

long sequences. However, subsequent architectural innovations, including sparse attention patterns (Child, Gray, Radford, & Sutskever, 2019), linear attention mechanisms (Katharopoulos, Vyas, Pappas, & Fleuret, 2020), and other approximation techniques (Kitaev, Łukasz Kaiser, & Levskaya, 2020), have addressed this limitation while maintaining model performance.

The original transformer architecture comprises of an encoder-decoder structure designed for sequence to sequence tasks such as machine translation. The encoder transforms an input sequence into a continuous representation, while the decoder generates an output sequence based on this representation. Each encoder and decoder block combines self-attention mechanisms with position-wise feed-forward networks, normalization layers, and residual connections. Position-wise feed-forward networks apply identical transformations to each position independently, incorporating non-linearities and enabling the model to process the contextual representations generated by attention mechanisms. Layer normalization stabilizes training by normalizing activations across the feature dimension, while residual connections facilitate gradient flow through deep architectures. The transformer incorporates positional encodings to provide information about the relative or absolute positions of elements in the sequence. Various architectural variants have emerged from the original transformer design, each optimized for specific applications or addressing particular limitations. The transformer encoder alone forms the basis for bidirectional models like BERT (Bidirectional Encoder Representations from Transformers) (Devlin, Chang, Lee, & Toutanova, 2019), which develop contextual representations by attending to all tokens in a sequence. Conversely, the transformer decoder underlies autoregressive models like GPT (Generative Pre-trained Transformer) (Radford, Narasimhan, Salimans, & Sutskever, 2018), which generate sequences by predicting each token conditioned on all previous tokens. Hybrid approaches incorporate both encoder and decoder components for tasks requiring both bidirectional understanding and autoregressive generation (Raffel et al., 2023).

In the visual domain, ViT (Dosovitskiy et al., 2021) adapt the transformer architecture to image processing by decomposing images into sequences of fixed-size patches, linearly embedding these patches, and processing them with standard transformer encoders. This approach contrasts with the convolutional architectures, demonstrating that locality bias is not an essential inductive bias for visual processing when sufficient data and computational resources are available. Subsequent

developments have integrated convolutional operations into transformer architectures (Z. Liu et al., 2021), incorporated hierarchical processing (W. Wang et al., 2021), and developed efficient attention mechanisms for high-resolution images (Z. Liu et al., 2021). The application of transformers to computational pathology presents both significant opportunities and unique challenges. The gigapixel dimensions of histopathological images exceed the context length limitations of standard transformer architectures, necessitating adaptations to accommodate these large-scale images. Several strategies have emerged to address this constraint. Hierarchical transformers process images at multiple levels of granularity, from patches to regions to whole slides, enabling efficient modeling of long range dependencies while maintaining tractable computational complexity (R. J. Chen et al., 2022). Patch based approaches (X. Wang et al., 2022) segment whole slide images into manageable patches, process these patches independently with transformer encoders, and subsequently aggregate the patch level representations to derive slide level predictions. Attention mechanism modifications, including sparse attention patterns, linear attention variants, and local to global attention combinations, reduce the complexity of processing large histopathological images while preserving the ability to capture dependencies (X. Wang et al., 2022).

Beyond architectural adaptations, the future of transformers in computational pathology points toward specialized foundation models pre-trained on histopathological datasets (Filiot et al., 2023), increasingly sophisticated multi-modal integration with genomic and clinical data (R. J. Chen et al., 2021), and enhanced interpretability mechanisms aligned with diagnostic workflows. These developments suggest transformers will become the architectural backbone for next generation pathology systems, potentially revolutionizing diagnosis while bridging visual patterns with underlying disease mechanisms (Ciga et al., 2021).

2.1.4 Foundation Models

Foundation models represent a paradigm shift in artificial intelligence, characterized by large-scale neural networks trained on vast, diverse datasets that can be adapted to numerous downstream tasks with minimal additional training (Bommasani et al., 2022). Unlike conventional task specific models that require extensive labeled data for each new application, foundation models learn generalizable representations through self-supervised learning on unlabeled data, enabling efficient

knowledge transfer to specialized domains. This transfer learning capability, where knowledge from one domain or task transfers to another forms the cornerstone of foundation models' flexibility and efficiency.

The concept of foundation models coalesced around several pioneering systems, most notably LLMs like GPT developed by OpenAI. The GPT series demonstrated that scale in parameters, compute, and data—could lead to qualitatively different capabilities, with GPT-3 showcasing remarkable zero-shot and few-shot learning abilities (Brown et al., 2020). Similarly, models like Large Language Model Meta AI (LLAMA) as seen in Fig. 2.2 illustrated how architectural optimization could create more efficient foundation models that maintain performance while reducing computational requirements. These models exhibit emergent capabilities that appear only at scale, including complex reasoning, instruction following, and in-context learning that were not explicitly programmed (Wei et al., 2022). Foundation models can be broadly categorized into generative and non-generative architectures, each with distinct capabilities and applications. Generative foundation models, exemplified by GPT, LLAMA, and diffusion models like Stable Diffusion and DALL-E, are designed to produce novel content whether text, images, or other modalities by learning the underlying distribution of training data and sampling from this distribution. These models excel at creative tasks, content generation, and reasoning. Non-generative foundation models, such as BERT and early versions of CLIP, focus on representation learning and understanding rather than generation. They excel at tasks like classification, retrieval, and feature extraction, mapping inputs to meaningful vector representations without necessarily producing new content. While this distinction was initially clear, recent foundation models increasingly blur these boundaries, with many systems incorporating both discriminative understanding and generative capabilities (Ramesh, Dhariwal, Nichol, Chu, & Chen, 2022; Touvron, Lavril, et al., 2023).

These diverse architectures rely on similar training methodologies centered on self-supervised objectives that derive supervision signals intrinsically from data structure, eliminating dependency on manual annotation. For language models, these objectives include autoregressive next-token prediction, which maximizes the probability of subsequent tokens given preceding context (Devlin et al., 2019). This approach enables models to learn linguistic patterns, factual knowledge,

and reasoning capabilities from text alone, without explicit supervision. Visual foundation models implement analogous techniques including contrastive learning between image pairs, masked image modeling where models reconstruct obscured image regions, and teacher-student distillation frameworks (Caron et al., 2021; T. Chen, Kornblith, Norouzi, & Hinton, 2020; K. He et al., 2021). While general-purpose foundation models demonstrate impressive transfer capabilities, domain-specific variants achieve superior performance in specialized contexts through additional pre-training or fine-tuning on domain-relevant data. In biomedicine, models like BioBERT and ClinicalBERT adapt to medical terminology through additional pre-training, while visual models pre-trained on medical imaging capture domain-specific patterns (Alsentzer et al., 2019; Lee et al., 2020; Sellergren et al., 2023). Adapting these to computational pathology requires addressing unique challenges of histopathological images: gigapixel dimensions, multi-scale content, and specialized semantics—through hierarchical architectures that process information across scales and efficient attention mechanisms for handling large-scale images (R. J. Chen et al., 2022; X. Wang et al., 2022).

For computational pathology specifically, three principal approaches have emerged: fine-tuning existing vision models, training pathology-native models, and integrating multimodal data, enabling natural language interactions, semantic search, and cross-domain knowledge transfer. Implementation challenges include computational demands for gigapixel images requiring model compression, ensuring interpretability through concept based explanations, domain adaptation across varied staining protocols, and addressing privacy concerns through federated learning approaches that preserve patient data security while enabling model training across institutions (Lu, Kong, et al., 2020).

2.1.5 Vision Language Models

VLMs represent a specialized category of foundation models that integrate visual and linguistic understanding to interpret and reason about visual content through language. As foundation models have revolutionized AI development across domains, VLMs specifically address the challenge of bridging visual perception with language comprehension, enabling systems to understand and communicate about visual information using natural language interfaces. This integration creates powerful systems capable of tasks ranging from image captioning and visual question answering

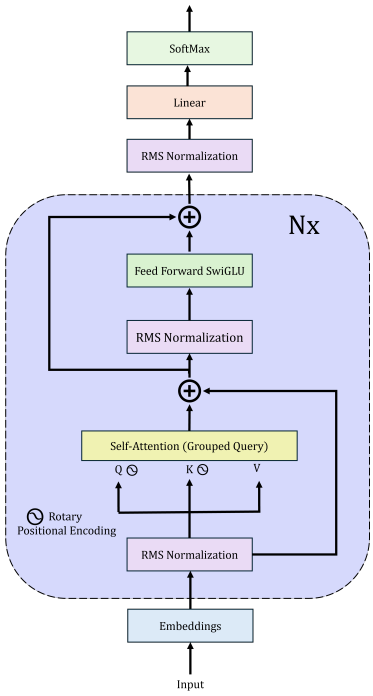


Figure 2.2: LLAMA architecture

to complex visual reasoning and cross-modal retrieval. The evolution of VLMs has been marked by several revolutionary conceptual advances. Initially, visual and linguistic understanding existed as separate capabilities, connected through simple alignment mechanisms (Agrawal et al., 2016; Vinyals, Toshev, Bengio, & Erhan, 2015). The breakthrough came with contrastive learning approaches, exemplified by CLIP, which trained models on millions of image-text pairs from the web without requiring curated annotations (Radford et al., 2021b). This methodology created a unified semantic space where visual and textual concepts could be directly compared, enabling remarkable zero-shot capabilities where models could recognize visual concepts simply by their textual description. This approach fundamentally altered how visual understanding systems could be developed and deployed.

The integration of VLMs with LLMs such as LLAMA as seen in Fig. 2.2 represents another revolutionary advance, connecting visual perception with the reasoning capabilities, world knowledge, and flexible output generation of modern language models.

Models like Bootstrapped Language-Image Pretraining (BLIP), Flamingo, and LLaVA demonstrated that visual information could serve as context for language models, enabling complex reasoning about visual content through natural language interaction (H. Liu et al., 2023). This integration allows systems to not merely describe what they see but to analyze, infer, and reason about visual information in ways that more closely resemble human cognitive processes. In computational pathology, VLMs address domain-specific challenges that traditional computer vision approaches struggle with. Histopathological images present unique properties including gigapixel dimensions, specialized visual features, and hierarchical tissue organization spanning multiple biological scales (R. J. Chen et al., 2024; Lu, Chen, Williamson, et al., 2023; Song et al., 2023). VLMs offer transformative capabilities for computational pathology through several key mechanisms. They enable intuitive natural language interfaces for examining complex histopathological data, allowing pathologists to query whole slide images using domain-specific terminology that aligns with clinical practice. By leveraging representations learned from general image-text pairs, these models facilitate zero-shot and few-shot learning, recognizing pathological patterns from descriptions alone without requiring extensive labeled examples (Lu, Chen, Zhang, et al., 2023). Their hierarchical architectures support multi-scale analysis that mirrors pathologists' workflow integrating information from tissue architecture to cellular details (R. J. Chen et al., 2022). Pathology specific VLMs connect visual findings with medical knowledge, bridging visual patterns and concepts from literature and clinical guidelines through domain adapted encoders paired with biomedically enhanced language models (Lu et al., 2024; Sun et al., 2024; Zhang et al., 2025). This integration enables sophisticated diagnostic reasoning that combines visual evidence with medical knowledge in ways that conventional computer vision approaches cannot achieve.

2.2 Literature Survey

This section synthesizes current research in computational pathology, examining how different AI approaches address the unique challenges of histopathological image analysis. We critically analyze the evolution from basic convolutional methods to advanced VLMs, comparing contrastive

and non-contrastive learning paradigms, knowledge distillation techniques, multiple instance frameworks, state space models, and prompt engineering strategies. This comprehensive review identifies key trends, remaining challenges, and promising directions that inform our investigation of prompt engineering for zero-shot diagnostic applications in computational pathology.

2.2.1 AI in Computational Pathology

Artificial intelligence has revolutionized computational pathology by addressing the unique challenges of analyzing gigapixel whole slide images with multi-scale tissue organization. The progression of AI approaches in this domain demonstrates increasingly sophisticated adaptations to these challenges. CNNs initially proved viable for cancer detection tasks but were limited by their inability to process entire WSIs and dependence on extensive annotations (Coudray et al., 2018). Multiple Instance Learning frameworks subsequently enabled training on slide-level diagnoses without requiring patch-level annotations, significantly reducing the annotation burden while maintaining diagnostic accuracy (Campanella et al., 2019; Lu, Williamson, et al., 2020). These methods conceptualize slides as collections of patches, employing attention mechanisms to focus computational resources on diagnostically relevant regions. Self-supervised learning techniques further reduced annotation requirements by leveraging unlabeled histopathology data to learn meaningful representations before fine-tuning on specific tasks. Domain-specific adaptations, particularly H&E color augmentations simulating staining variations, proved essential for representation quality (Kang, Song, Park, Yoo, & Pereira, 2023). ViT then revolutionized the field by effectively capturing long range dependencies crucial for assessing tissue architecture. The hierarchical image pyramid transformer (HIPT) specifically addressed histopathology's multi scale nature by pretraining at multiple magnification levels, mirroring pathologists diagnostic workflow of examining specimens across different magnifications (R. J. Chen et al., 2022). Recent innovations focus on multimodal integration combining histopathological images with molecular data to reveal morpho-molecular correlations that enhance disease characterization (R. J. Chen et al., 2021). These approaches enable computational staining that highlights morphological features predictive of molecular subtypes, democratizing access to precision medicine. VLMs now create natural language interfaces to

histopathological images, enabling interactive exploration through queries, automated report generation, and semantic search for similar cases (Lu, Chen, Williamson, et al., 2023).

Despite significant progress, implementation challenges persist, including domain shift between institutions, interpretability concerns, and data privacy issues (M. et al., 2020; Tellez et al., 2019). Federated learning approaches address some of these challenges by enabling model training across institutions without direct data sharing (Lu, Kong, et al., 2020). Looking forward, AI in computational pathology is evolving toward human and AI partnerships that augment pathologist capabilities through consistent quantification, region of interest highlighting, and multimodal data integration, positioning pathology at the forefront of precision medicine initiatives.

2.2.2 Self Supervised Learning for Computational Pathology

Contrastive learning has emerged as a dominant paradigm in computational pathology, enabling effective representation learning from unlabeled histopathology data. The core principle involves maximizing agreement between differently augmented views of the same image while minimizing similarity between different images. For a positive pair (i, j), the standard contrastive loss from Simple Framework for Contrastive Learning of Visual Representations (SimCLR) (T. Chen et al., 2020) can be expressed as:

$$\mathcal{L}_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(z_i, z_k)/\tau)} \tag{2}$$

where τ is a temperature parameter, and $\text{sim}(z_i, z_j)$ denotes cosine similarity between normalized embeddings. Histopathology presents unique challenges that shape contrastive learning approaches: extensive tissue heterogeneity within single slides, significant stain variability across laboratories, and the multi-scale nature of diagnostic patterns (Kang et al., 2023). These domain-specific challenges have driven the development of specialized contrastive frameworks. (Tellez, Litjens, van der Laak, & Ciompi, 2021) pioneered self-supervised learning for computational pathology through neural image compression, demonstrating that networks trained to reconstruct compressed histopathology images developed useful representations for tumor classification while requiring only a small percentage of labeled data. This established the viability of self-supervised

approaches for pathology applications where annotations are scarce. A critical advancement came when SimCLR was adapted for pathology (Ciga et al., 2021), revealing that domain-specific color augmentations capturing staining variations were substantially more effective than standard augmentations used for natural images. These H&E color perturbations, which simulate routine staining variations seen in clinical practice, significantly improved representation quality and have become a foundational principle for subsequent self-supervised methods in computational pathology (Ciga et al., 2021). Momentum Contrast for Unsupervised Visual Representation Learning (MoCo) (K. He, Fan, Wu, Xie, & Girshick, 2020) has been particularly influential in computational pathology. By maintaining a dynamic dictionary of encoded representations with a momentum-updated encoder, MoCo enables more consistent feature learning across large and diverse pathology datasets. The InfoNCE loss used in MoCo is formulated as:

$$\mathcal{L}_q = -\log \frac{\exp(q \cdot k_+/\tau)}{\sum_{i=0}^{K} \exp(q \cdot k_i/\tau)} \tag{3}$$

where q is the query encoding, k_+ is the positive key encoding, and $\{k_i\}_{i=0}^{K}$ are the encodings in the queue. This memory bank mechanism has proven particularly effective for rare histological patterns by providing a larger and more diverse set of negative samples, addressing the long-tailed distribution of tissue appearances in pathology datasets. Knowledge distillation techniques like DINO (Distillation with No Labels) (Caron et al., 2021) represent another important advancement in contrastive learning. The DINO loss is defined as:

$$\mathcal{L} = H(P_t, P_s) = -P_t \log P_s \tag{4}$$

where P_t and P_s are the teacher and student probability distributions, respectively. In pathology applications, DINO enables vision transformers to capture histological structures, with attention heads learning to localize diagnostically relevant regions without explicit supervision (R. J. Chen & Krishnan, 2022). The Self-Path framework (Koohbanani et al., 2021) introduced pathology-specific pretext tasks including magnification prediction, jigsaw puzzles of tissue regions, and rotation prediction. These tasks leverage domain knowledge that diagnostically relevant features appear at

different magnification levels and in specific spatial arrangements. Comparative evaluations demonstrated that these pathology-specific tasks outperformed generic contrastive methods for tumor classification. The contrastive learning framework has been extended to address the heterogeneity inherent in whole slide images through multiple instance contrastive learning (B. Li, Li, & Eliceiri, 2021). This approach considers the multiple instance learning paradigm where a slide contains numerous tissue patches (instances). The dual stream contrastive learning combines instance level and bag-level contrastive objectives. This approach significantly improved rare pattern detection in breast and colon histopathology by ensuring that the model learned discriminative features at both the cellular and architectural levels. The evolution of contrastive learning in computational pathology extends to applications in survival prediction (Abbet, Zlobec, Bozorgtabar, & Thiran, 2020), tumor segmentation (X. Wang et al., 2022), and rare disease identification (Azizi et al., 2022). These approaches demonstrate how foundational contrastive techniques can be adapted to specific clinical tasks. A significant development in contrastive learning came with the introduction of supervised contrastive learning (Khosla et al., 2021), which extended the self-supervised contrastive approach to the supervised setting. The supervised contrastive loss is defined as:

$$\mathcal{L}_{\text{sup}} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i \cdot z_p / \tau)}{\sum_{a \in A(i)} \exp(z_i \cdot z_a / \tau)} \tag{5}$$

where $P(i)$ is the set of indices of samples with the same class as i, and $A(i)$ is the set of all indices except i. Collectively, these advancements demonstrate a clear trend toward domain-specific adaptations of foundational contrastive learning techniques.

Non-contrastive self-supervised learning methods have emerged as compelling alternatives in computational pathology, offering distinct advantages by eliminating dependence on negative pairs and addressing limitations related to batch size and sampling strategies. These methods maintain representation distinctiveness through mechanisms other than explicit contrast between positive and negative pairs, often employing redundancy reduction, clustering, or reconstruction objectives. Barlow Twins (Zbontar et al., 2021) exemplifies the redundancy reduction approach by producing similar representations for distorted versions of the same image while simultaneously minimizing redundancy in the embedding space:

$$\mathcal{L}_{\text{BT}} = \sum_i (1 - \mathcal{C}_{ii})^2 + \lambda \sum_i \sum_{j \neq i} \mathcal{C}_{ij}^2 \tag{6}$$

where $\mathcal{C}$ is the cross-correlation matrix between network outputs. The first term encourages diagonal elements to be 1, ensuring perfect correlation for positive pairs, while the second term encourages off-diagonal elements to be 0, reducing redundancy by decorrelating dimensions (Lu, Williamson, et al., 2020). In computational pathology, barlow twins has demonstrated superior performance on smaller datasets compared to contrastive methods (Kang et al., 2023), making it particularly valuable for institutions with limited data repositories. VICReg (Variance-Invariance-Covariance Regularization) (Bardes, Ponce, & LeCun, 2022) maintains representation variance above a threshold while ensuring invariance between augmented views and minimizing covariance between different representation dimensions.

In pathology applications, VICReg has demonstrated robust performance with particular strength in preserving local structural information critical for diagnostic assessment (Bardes et al., 2022), better preserving local tissue morphology characteristics than contrastive methods that may focus on globally distinctive features. The integration of masked autoencoding with vision transformers has shown remarkable promise in computational pathology. Masked autoencoders randomly mask patches and attempt to reconstruct the original content:

$$\mathcal{L}_{\text{MAE}} = \frac{1}{|M|} \sum_{i \in M} ||x_i - \hat{x}_i||_2^2 \tag{7}$$

where M represents masked patch indices and $\hat{x}_i$ the reconstructed patches (K. He et al., 2021). TransPath (X. Wang et al., 2022) pioneered region-aware masking based on tissue importance. This approach demonstrated significant improvements in tumor classification and segmentation by focusing reconstruction efforts on diagnostically significant regions.

This multi-scale approach has demonstrated superior performance for survival prediction across multiple cancer types (R. J. Chen et al., 2022; Vorontsov et al., 2024), capturing both cellular morphology and tissue architecture in alignment with diagnostic paradigms in pathology. Simple Masked Image Modeling (SimMIM) (Xie et al., 2022) has been adapted for computational pathology through tissue density-guided masking that prioritizes regions with higher cellular content. This

approach ensures focus on diagnostically relevant regions while paying less attention to background areas (Xie et al., 2022). A comprehensive benchmark of non-contrastive self-supervised frameworks for computational pathology (Kang et al., 2023) revealed that while barlow twins performed better on smaller datasets, transformer-based masked autoencoders excelled with larger pre-training data. The benchmark also highlighted the critical importance of domain-specific augmentations for all non-contrastive methods, with H&E color perturbations providing substantial benefits across different architectural approaches. The integration of non-contrastive self -supervised with ViT has demonstrated the emergence of interpretable attention maps that correspond to diagnostically relevant regions without explicit supervision (L. Chen et al., 2022). This implicit localization capability offers visual explanation that aligns with pathologists diagnostic reasoning, potentially enhancing model trustworthiness in clinical settings. Together, these non-contrastive approaches offer complementary strengths to contrastive methods in computational pathology. Their reduced dependency on negative samples, superior performance on smaller datasets, and explicit modeling of multi-scale tissue characteristics address specific challenges in pathology image analysis. As these methods continue to evolve with domain-specific adaptations, they promise to further reduce annotation requirements while preserving interpretability and diagnostic accuracy.

Knowledge distillation represents a powerful paradigm in deep learning that enables the transfer of knowledge from complex, high capacity models to more compact architectures. Originally formalized by (Hinton, Vinyals, & Dean, 2015), knowledge distillation employs a teacher-student framework where a large, pre-trained teacher model guides the training of a smaller student model through soft targets. The fundamental knowledge distillation objective can be expressed as a weighted combination of task-specific and distillation losses:

$$\mathcal{L} = \alpha \mathcal{L}_{task} + (1 - \alpha)\mathcal{L}_{KD} \tag{8}$$

where $\mathcal{L}_{KD}$ typically employs Kullback-Leibler divergence to measure differences between probability distributions from teacher and student models. The temperature parameter controls the softness of these distributions, revealing the teacher's confidence in secondary predictions and transferring this generalization capability to the student.

Subsequent work has expanded on this foundation in several directions. (Romero et al., 2015) introduced FitNets, which extends knowledge distillation to intermediate layers, enabling the transfer of not just outputs but also feature representations. This approach, termed hint-based training, allows deeper and thinner student networks to learn from wider teacher networks by matching intermediate feature maps. The student's hidden layers are guided to mimic the teacher's intermediate representations, providing richer supervision than output-only distillation.

(Park, Kim, Lu, & Cho, 2019) proposed Relational Knowledge Distillation (RKD), shifting focus from individual outputs to structural relationships between data examples. RKD preserves the relative distances and angles between data points in the feature space, ensuring that the student learns the same structural relationships captured by the teacher. This approach proves particularly effective when the absolute values of outputs are less important than their relative relationships, enhancing the student's generalization capabilities.

(Gou, Yu, Maybank, & Tao, 2021) conducted a comprehensive survey of knowledge distillation methods, categorizing the evolving landscape into response based, feature-based, and relation-based approaches. Their analysis reveals that while the original formulation focused on transferring softened logits, modern approaches increasingly emphasize the transfer of structural knowledge and representations. These advanced techniques have enabled significant compression of large models while maintaining performance across various domains including computer vision, natural language processing, and speech recognition.

Knowledge distillation continues to evolve as model architectures grow in complexity. The technique offers substantial practical benefits, enabling the deployment of high performing models in resource constrained environments, reducing computational costs and energy consumption, and potentially improving generalization through the regularization effect of soft targets. As foundation models continue to grow in size and capability, knowledge distillation will likely play an increasingly crucial role in making these advancements accessible across diverse computational environments.

2.2.3 Multiple Instance Learning

Multiple Instance Learning (MIL) addresses a distinctive form of weakly supervised learning where labels are associated with groups of instances (bags) rather than individual instances. In the standard MIL formulation, a bag is labeled positive if at least one instance within it is positive, while negative bags contain exclusively negative instances. This framework elegantly accommodates scenarios where fine-grained instance-level annotations are unavailable or prohibitively expensive to obtain, making it applicable across diverse domains including computer vision, drug discovery, and document classification (Dietterich & Bakiri, 1995).

The seminal work by (Ilse, Tomczak, & Welling, 2018) introduced attention-based pooling for MIL, significantly advancing the field by enabling models to learn which instances are most relevant for bag-level classification. Their approach employs a trainable attention mechanism to dynamically weight instance contributions when forming bag-level predictions. The attention-based aggregation can be formulated as:

$$z = \sum_{k=1}^{K} a_k \cdot f_k, \quad \text{where} \quad a_k = \frac{\exp(w^T \tanh(V f_k^T))}{\sum_{j=1}^{K} \exp(w^T \tanh(V f_j^T))} \tag{9}$$

with V and w being trainable parameters. This formulation allows the network to determine which instances are most informative without requiring explicit instance-level annotations. The resulting attention weights provide interpretable insights into instance importance, revealing which elements most strongly influence the bag-level classification.

(X. Wang, Yan, Tang, Bai, & Liu, 2018) conducted a comprehensive analysis of MIL pooling strategies, comparing various instance aggregation approaches including max-pooling, mean-pooling, and attention-based pooling. Their findings demonstrate that different aggregation functions embody distinct assumptions about the relationship between instance-level and bag-level classifications. They showed that while max-pooling aligns with the standard MIL assumption (focusing exclusively on the most positive instance), attention mechanisms better capture complex scenarios where multiple instances contribute to the bag label with varying importance.

Recent innovations in MIL include transformer-based approaches that model relationships between instances while maintaining permutation invariance. These methods treat instances as tokens

and employ self-attention mechanisms to capture interactions, demonstrating the continued evolution of MIL toward more sophisticated modeling of bag structures. As weakly supervised learning continues to gain importance in scenarios where exhaustive annotation is impractical, MIL remains a fundamental paradigm with ongoing methodological advances across application domains.

2.2.4 State Space Models

State space models offer an alternative to attention-based architectures, with notable advantages in computational efficiency (A. Gu, Goel, & Ré, 2022). Inspired by classical linear state space systems, these models maintain linear complexity while effectively capturing long-range dependencies (Smith, Warrington, & Linderman, 2023). The fundamental state space transformation is expressed as:

$$\frac{d}{dt}h(t) = Ah(t) + Bx(t), \quad y(t) = Ch(t) \tag{10}$$

where A denotes the state matrix, B the input matrix, and C the output matrix (A. Gu et al., 2022). This formulation enables efficient processing of sequential data while preserving important temporal relationships. The Structured State Space Sequence Model (S4) extends this framework to handle multi-scale data through discretized S4 layers with learnable parameters (A. Gu et al., 2022). Similarly, various architectures combine the efficiency of state space models with domain-specific adaptations (Nguyen et al., 2022). Benchmark studies have demonstrated that state space architectures can process high-resolution images substantially faster than equivalent transformer models while using significantly less memory. Architectural innovations include hierarchical designs that operate at multiple scales and selective state space models, which incorporate learned gating mechanisms to dynamically adjust state parameters based on input characteristics (A. Gu & Dao, 2024). These approaches have proven effective for complex analysis tasks requiring the simultaneous assessment of multiple components. Training strategies for state space models often utilize self-supervised pretraining to learn meaningful representations without explicit labels (Smith et al., 2023), frequently employing contrastive objectives between different views of the same data. The

computational advantages of state space models linear complexity, reduced memory usage, and significant efficiency gains make them particularly well suited for resource-constrained environments and real-time analysis applications (A. Gu et al., 2022).

2.2.5 Prompt Engineering in Computational Pathology

Prompt engineering involves crafting inputs or input templates to elicit desired behavior from a pre-trained model. It has emerged as a critical component in leveraging LLMs and VLMs for computational pathology applications. At its core, prompt engineering entails designing input instructions that guide these models toward generating optimal outputs, with particular emphasis on domain-specific requirements in medical imaging analysis (Qu et al., 2024). The strategic formulation of prompts directly influences the quality, relevance, and accuracy of model responses, especially in diagnostic applications where precision is paramount.

Configuration parameters significantly affect model behavior and are essential elements of prompt engineering practice (J. Gu et al., 2023). These parameters are particularly crucial in medical settings, where diagnostic accuracy demands careful tuning to ensure consistency and reliability. Various prompting techniques have evolved to enhance model performance across different tasks (Sahoo et al., 2025). Zero-shot prompting represents the most basic form, while one-shot and few-shot prompting techniques introduce exemplars to guide the model through similar input-output patterns. These approaches are especially valuable in specialized domains such as pathology, where task-specific guidance helps models focus on relevant visual features and domain-specific terminology (J. Wang et al., 2024).

Advanced techniques, including Chain-of-Thought (CoT) prompting, have demonstrated significant improvements in complex tasks requiring multi-step logical inference (Wei et al., 2023). For VLMs analyzing pathology images, CoT enables explicit reasoning about visual features and their diagnostic significance. System prompting, role prompting, and contextual prompting represent complementary strategies that establish the operational frame for model responses, often positioning the model as a domain expert, such as a pathologist. The development of VLMs for computational pathology necessitates careful consideration of domain-specific prompt engineering

strategies (Qu et al., 2024). Domain-adapted prompts incorporating medical terminology, anatomical specificity, and diagnostic criteria have been shown to significantly enhance model performance compared to generic templates.

The effectiveness of prompt engineering strategies varies across different VLM architectures, with models exhibiting different levels of sensitivity to prompt formulation. Models with dedicated pathology pretraining generally achieve better performance with domain-specific prompts compared to general domain counterparts. As VLMs continue to evolve, prompt engineering techniques must adapt to exploit architecture-specific strengths while mitigating model limitations, particularly in high-stakes medical applications (J. Gu et al., 2023).

Recent developments have focused on dynamic prompt optimization, including interpretable prompt optimization and attribute-guided prompt tuning (Zhan, Zhang, Lin, Wang, & Wang, 2023). Reinforcement learning from human feedback (RLHF) shows promise for dynamically adjusting prompts to stabilize model responses and improve prediction reliability. The field is also progressing toward automated prompt optimization, where models autonomously generate and refine prompts based on feedback loops. These innovations suggest that prompt engineering will remain essential for the deployment of VLMs in computational pathology, with increasingly sophisticated techniques emerging to address domain-specific challenges and enhance diagnostic capabilities.

Chapter 3

Investigating Zero-Shot Diagnostic Pathology in Vision-Language Models with Efficient Prompt Design

3.1 Introduction

Building upon the foundation of VLMs in computational pathology as discussed in Chapter 1 and 2, this chapter conducts a systematic zero-shot evaluation of the four state of the art VLM architectures called Quilt-Net (Ikezogwo et al., 2025), Quilt-LLAVA (Seyfioglu et al., 2025), CONCH (Lu, Chen, Williamson, et al., 2023) and BioMedCLIP (Zhang et al., 2025). Each model represents a unique approach to integrating visual and linguistic information in the histopathology domain. Using an in-house dataset of 3,507 clinically validated digestive system WSIs, we assess these model's diagnostic performance while systematically varying prompt engineering dimensions including domain specificity, anatomical precision, instructional framing, and output constraints. This comparative analysis provides insights into how architectural differences interact with prompt formulation to influence diagnostic accuracy across different tissue types and pathological conditions.

3.2 Methods

This section presents the architectures and foundations of the four distinct VLM frameworks investigated as part of this research under a zero-shot setting. All hyperparameter values appearing in the training procedure discussion, including temperature parameters (τ, σ), learning rates (η), batch sizes (B), epoch counts (E), and loss weights (λ_{Con}, λ_{Cap}) are adopted directly from the original model implementations and publications for CONCH (Lu, Chen, Williamson, et al., 2023), BioMedCLIP (Zhang et al., 2025), Quilt-Net (Ikezogwo et al., 2025), and Quilt-LLAVA (Seyfioglu et al., 2025), unless stated otherwise.

3.2.1 CLIP based VLMs

Both BioMedCLIP and Quilt-Net adapt the CLIP architecture's dual-encoder paradigm for specialized domains, employing contrastive learning to create a joint embedding space between images and text descriptions. As shown in equations 11,12 and 13, these models employ a contrastive loss function to align visual and textual representations. Quilt-Net establishes a foundational framework for learning robust visual representations from WSIs while simultaneously aligning these representations with natural language descriptions (Ikezogwo et al., 2025; Lu, Chen, Williamson, et al., 2023). It combines the strengths of both contrastive learning and hierarchical feature extraction to address the unique challenges of computational pathology, building upon recent advancements in self-supervised learning for gigapixel histopathology images (R. J. Chen et al., 2022). Quilt-Net is trained by finetuning the pre-trained CLIP model from OpenAI (Radford et al., 2021b) on Quilt-1M . Quilt-1M is a large-scale dataset of 1 million image-text pairs for histopathology, created by combining data from YouTube educational videos, Twitter, research papers, and the internet. It was developed to enable representation learning for histopathology.

BioMedCLIP takes a similar approach but is specifically tailored for broader biomedical applications, being pre-trained on PMC-15M, a 15 million biomedical image and text pairs sourced from scientific literature and clinical repositories. PMC-15M, a dataset collected from 4.4 million scientific articles in PubMed Central. The dataset spans diverse biomedical image types, including radiography, microscopy, and pathology. While Quilt-Net focuses specifically on pathology,

BioMedCLIP's domain specialization allows it to recognize subtle tissue patterns across various biomedical imaging modalities and correlate them with precise medical descriptions. This specialization enables BioMedCLIP to capture nuanced relationships between pathological features and corresponding medical terminology that might be overlooked by less specialized models. Both models demonstrate superior performance compared to those pre-trained on natural images (Deng et al., 2009). For a batch size of N (image, text) pairs, the contrastive loss can be formulated as:

$$\mathcal{L}_{\text{contrastive}} = \frac{1}{2N} \sum_{i=1}^{N} \left(\mathcal{L}_{\text{v2t}}^{i} + \mathcal{L}_{\text{t2v}}^{i} \right) \tag{11}$$

where the vision-to-text and text-to-vision losses are defined as:

$$\mathcal{L}_{\text{v2t}}^{i} = -\log \frac{\exp(f_i^I \cdot f_i^T / \tau)}{\sum_{j=1}^{N} \exp(f_i^I \cdot f_j^T / \tau)} \tag{12}$$

$$\mathcal{L}_{\text{t2v}}^{i} = -\log \frac{\exp(f_i^T \cdot f_i^I / \tau)}{\sum_{j=1}^{N} \exp(f_i^T \cdot f_j^I / \tau)} \tag{13}$$

where f_i^I and f_i^T are the normalized image and text embeddings respectively, and τ is a temperature parameter controlling the sharpness of the probability distribution. The training procedure for Quilt-Net and BioMedCLIP can be formalized as shown in Algorithm 1.

Algorithm 1 Quilt-Net and BioMedCLIP Training Procedure

procedure TRAINCLIP($\mathcal{D}$, B, τ, η, E)
  Input: Dataset $\mathcal{D}$, batch size B, temperature τ, learning rate η, epochs E
  Initialize image encoder E_I (ViT-B/32) and text encoder E_T (GPT-2) with CLIP weights
  for epoch $= 1$ to E **do**
    for each batch $\{(I_1, T_1), \ldots, (I_B, T_B)\} \subseteq \mathcal{D}$ **do**
      for $i = 1$ to B **do**
        $f_i^I \leftarrow E_I(I_i) / \|E_I(I_i)\|_2$; $f_i^T \leftarrow E_T(T_i) / \|E_T(T_i)\|_2$ ▷ Normalized embeddings
      end for
      $L_{\text{I}\to\text{T}} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(\text{Sim}(f_i^I, f_i^T)/\tau)}{\sum_{j=1}^{B} \exp(\text{Sim}(f_i^I, f_j^T)/\tau)}$ ▷ Sim: cosine similarity
      $L_{\text{T}\to\text{I}} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(\text{Sim}(f_i^T, f_i^I)/\tau)}{\sum_{j=1}^{B} \exp(\text{Sim}(f_i^T, f_j^I)/\tau)}$
      $L = (L_{\text{I}\to\text{T}} + L_{\text{T}\to\text{I}})/2$
      Update E_I and E_T using L and η
    end for
  end for
  return E_I, E_T
end procedure

During inference for a classification task, as summarized in Fig. 3.1a, an input image is fed to the image encoder and the class labels are fed to the text encoder. The image and text embeddings then undergo cosine similarity, where the image-text combination with the highest similarity is selected as the class label. The image encoder is based on the ViT-B/32 architecture (Dosovitskiy et al., 2021), while the text encoder is based on GPT-2 (Radford et al., 2019).

3.2.2 Quilt-LLAVA

Quilt-LLAVA extends beyond the dual encoder approach of BioMedCLIP and Quilt-Net by adopting the LLAVA framework (H. Liu et al., 2023; Wu et al., 2023), which integrates an LLM based on LLAMA-2 (Touvron, Martin, et al., 2023) for enhanced vision-language capabilities in computational pathology. This architectural approach enables sophisticated interaction between visual histopathological data and medical textual descriptions, addressing limitations in previous models that lacked generative capabilities (H. Liu et al., 2023). In the Quilt-LLAVA architecture, generally described in Fig. 3.1b, the input image goes through a visual encoder (i.e., pre-trained Quilt-Net (Ikezogwo et al., 2025)) to extract features that are then projected into embeddings. The projection layer received the input from the visual encoder and maps the visual features into the language model's embedding space:

$$h_v = W \cdot z_v \tag{14}$$

where W are the weight of the project layer, $z_v = g(x_v)$ represents the output of the CLIP vision encoder g applied to the input image x_v, and h_v is the projected visual embedding compatible with the language model's embedding space. The LLAMA-2 architecture incorporates three key innovations that substantially enhance performance beyond traditional transformer models:

(1) **Pre-Normalization (RMS Norm):** LLAMA dramatically improves training stability by applying normalization to the input of each transformer sub-layer, rather than the output as in

conventional transformers. This RMS normalization is defined as:

$$\text{RMS}(x) = \sqrt{\frac{1}{n}\sum_{1}^{n}(x^2)} \tag{15}$$

$$\bar{x} = \frac{x}{\text{RMS}(x) + \epsilon} \tag{16}$$

where ϵ is a small value to prevent division by zero. This approach effectively eliminates training instabilities.

(2) **SwiGLU Activation Function:** LLAMA uses the SwiGLU activation function, which provides superior gradient flow and expressiveness:

$$\text{swiGLU}(x) = \text{Swish}_\beta(xW + b) \otimes (xV + c) \tag{17}$$

$$\text{Swish}_\beta(xW + b) = (xW + b) \otimes \sigma(\beta(xW + b)) \tag{18}$$

$$\sigma(\beta(xW + b)) = \frac{1}{1 + e^{-(\beta(xW+b))}} \tag{19}$$

(3) **Rotary Positional Embeddings (RoPE):** LLAMA's revolutionary approach to position encoding uses rotary embeddings instead of the static positional encodings in vanilla transformers. This ingenious encoding method vastly improves the model's ability to handle long-range dependencies and generalizes to sequence lengths beyond those seen during training.

Quilt-LLaVA uses a two-stage training approach :

(1) It is aligned with the histopathology domain using 723K image-text pairs from QUILT-1m dataset, with only the MLP projection layer trained while the vision encoder and language model are frozen.

(2) Then only the language model and MLP are instruction tuned on QUILT-INSTRUCT, a dataset of 107K histopathology-specific question-answer pairs extracted from educational videos with spatially localized medical concepts.

This enables Quilt-LLaVA to analyze histopathology images in detail, localize medical concepts,

reason beyond single image patches, and significantly outperform models like LLAVA and LLAVA-MED on histopathology visual question answering tasks. The entire training procedure for it can be seen in Algorithm 2.

Algorithm 2 Quilt-LLAVA Two-Stage Training Procedure

procedure TRAINQUILTLLAVA($\mathcal{D}_{\text{pre}}$, $\mathcal{D}_{\text{inst}}$, B, η_1, η_2, E_1, E_2)
  Input: Pre-training dataset $\mathcal{D}_{\text{pre}}$, instruction dataset $\mathcal{D}_{\text{inst}}$, batch size B
  Freeze vision encoder $g(\cdot)$ (CLIP ViT-L/32) and LLM parameters ϕ
  Initialize projection layer W randomly with optimizer learning rate η_1
  Stage 1: Pre-training projection layer only
  for epoch $= 1$ to E_1 **do**
    for each batch $\{(X_v^1, X_c^1), \ldots, (X_v^B, X_c^B)\} \subseteq \mathcal{D}_{\text{pre}}$ **do**
      for $j = 1$ to B **do**
        $Z_v^j \leftarrow g(X_v^j)$ ▷ Extract frozen visual features
        $H_v^j \leftarrow W \cdot Z_v^j$ ▷ Project to word embedding space
        $\mathcal{L}_j^{\text{pre}} \leftarrow \sum_{t=1}^{|X_c^j|} -\log p\Big(X_{c,t}^j \mid H_v^j, X_{c,<t}^j\Big)$
      end for
      Update W using gradients of $\mathcal{L}^{\text{pre}} = \frac{1}{B}\sum_{j=1}^{B} \mathcal{L}_j^{\text{pre}}$
    end for
  end for
  Stage 2: Instruction tuning
  Keep $g(\cdot)$ frozen, initialize optimizer for W and ϕ with learning rate η_2
  for epoch $= 1$ to E_2 **do**
    for each batch $\{(X_v^1, X_q^1, X_a^1), \ldots, (X_v^B, X_q^B, X_a^B)\} \subseteq \mathcal{D}_{\text{inst}}$ **do**
      for $j = 1$ to B **do**
        $Z_v^j \leftarrow g(X_v^j)$ ▷ Extract frozen visual features
        $H_v^j \leftarrow W \cdot Z_v^j$ ▷ Project to word embedding space
        $X_{\text{instruct}}^j \leftarrow$ Format instruction according to Eq. (2) in paper
        $\mathcal{L}_j^{\text{inst}} \leftarrow \sum_{t=1}^{|X_a^j|} -\log p\Big(X_{a,t}^j \mid H_v^j, X_{\text{instruct}}^j, X_{a,<t}^j\Big)$
      end for
      Update W and ϕ using gradients of $\mathcal{L}^{\text{inst}} = \frac{1}{B}\sum_{j=1}^{B} \mathcal{L}_j^{\text{inst}}$
    end for
  end for
  return $g(\cdot)$, W, ϕ
end procedure

In Stage 1, the loss function $\mathcal{L}_j^{\text{pre}} = \sum_{t=1}^{|X_c^j|} -\log p\big(X_{c,t}^j \mid H_v^j, X_{c,<t}^j\big)$ is computed over caption tokens X_c^j, where $X_{c,<t}^j$ denotes all caption tokens before the current token $x_{c,t}^j$. In Stage 2, the loss function $\mathcal{L}_j^{\text{inst}} = \sum_{t=1}^{|X_a^j|} -\log p\big(X_{a,t}^j \mid H_v^j, X_{\text{instruct}}^j, X_{a,<t}^j\big)$ is computed over answer tokens X_a^j, with X_{instruct}^j representing the formatted instruction and $X_{a,<t}^j$ denoting all answer tokens preceding the current token $x_{a,t}^j$. X_v represents the image input that grounds the visual context.

3.2.3 CONCH

CONCH builds upon the foundations of BioMedCLIP, Quilt-Net and Quilt-LLAVA while introducing novel components for contextual reasoning and knowledge integration (Lu, Chen, Williamson, et al., 2023). Drawing inspiration from the Contrastive Captioners (CoCa) method and recent advances in VLMs (Yu et al., 2022), CONCH employs a decoupled decoder design that simultaneously supports contrastive and generative objectives. CONCH was trained on a large dataset of 1.17 million histopathology image–caption pairs, sourced from open-access articles and educational content. CONCH consists of an image encoder, a text encoder, and a multi-modal text decoder. The training procedure follows a unified approach combining contrastive and captioning objectives:

$$\mathcal{L}_{\text{CoCa}} = \lambda_{\text{Con}} \cdot \mathcal{L}_{\text{Con}} + \lambda_{\text{Cap}} \cdot \mathcal{L}_{\text{Cap}} \tag{20}$$

where λ_{Con} and λ_{Cap} are loss weighting hyper-parameters. The contrastive loss is formulated as a symmetric loss between image-to-text and text-to-image directions:

$$\mathcal{L}_{\text{Con}} = -\frac{1}{N}\left(\sum_{i=1}^{N} \log \frac{\exp(x_i^\top y_i/\sigma)}{\sum_{j=1}^{N} \exp(x_i^\top y_j/\sigma)} + \sum_{i=1}^{N} \log \frac{\exp(y_i^\top x_i/\sigma)}{\sum_{j=1}^{N} \exp(y_i^\top x_j/\sigma)}\right) \tag{21}$$

where x_i and y_j are normalized embeddings of the image in the i-th pair and that of the text in the j-th pair. N is the batch size, and σ is the temperature to scale the logits. The captioning loss uses the standard autoregressive language modeling objective:

$$\mathcal{L}_{\text{Cap}} = -\sum_{t=1}^{T} \log P_\theta(y_t | y_{<t}, x) \tag{22}$$

The training procedure for CONCH can be observed in the Algorithm 3.

BioMedCLIP, Quilt-Net, Quilt-LLAVA, and CONCH represent distinct vision-language architectures for computational pathology, each with different parameter scales and architectural approaches (Chanda, Aryal, Soltani, & Ganji, 2024; Lu, Chen, Williamson, et al., 2023). Quilt-Net and BioMedCLIP employ a CLIP-inspired dual-encoder approach with over 186M parameters (86M for the ViT-B/32 image encoder and over 100M for the text encoder), establishing effective contrastive learning between histopathological images and text. Quilt-LLAVA significantly expands

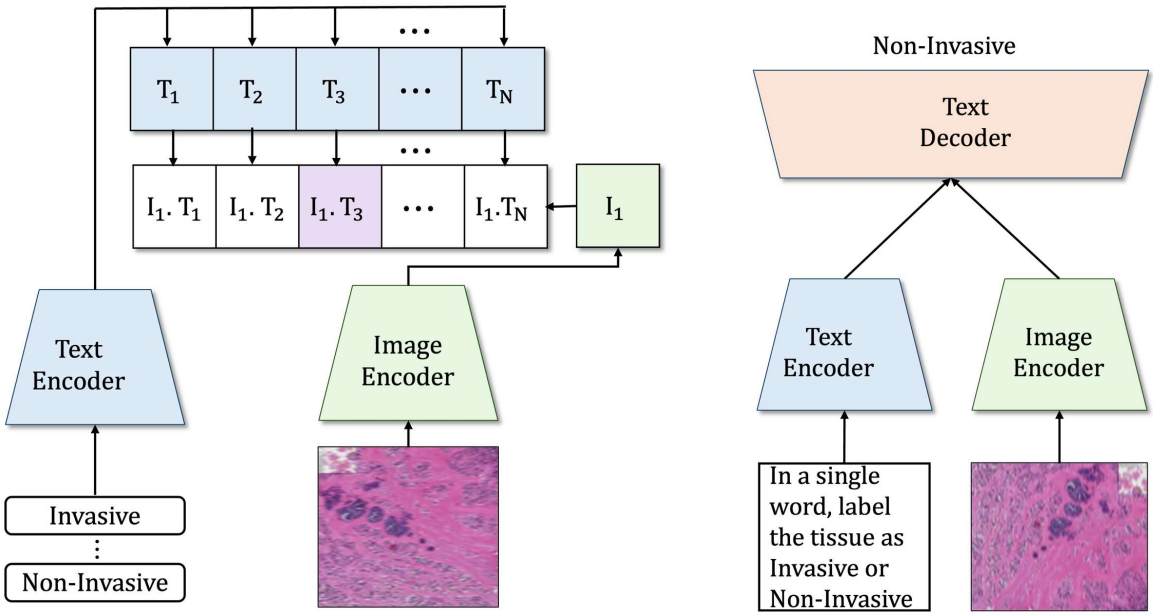
Algorithm 3 CONCH Training Procedure

procedure TRAINCONCH($\mathcal{D}$, λ_{Con}, λ_{Cap}, σ)
  Initialize model parameters $\Theta = \{\Theta_V, \Theta_T, \Theta_{\text{proj}}, \Theta_{\text{fusion}}\}$ ▷
  Θ_V: visual encoder weights,
  Θ_T: text encoder weights,
  Θ_{proj}: projection head for joint embedding,
  Θ_{fusion}: decoder for caption generation
  Initialize temperature parameter σ
  while not converged **do**
    Sample batch (I, T) from dataset $\mathcal{D}$
    $V = f_V(I; \Theta_V)$ ▷ Visual encoder
    $T = f_T(T; \Theta_T)$ ▷ Text encoder
    $Z_V, Z_T = \text{Project}(V, T; \Theta_{\text{proj}})$ ▷ Project to joint embedding space
    $\mathcal{L}_{\text{Con}} = \mathcal{L}_{\text{InfoNCE}}(Z_V, Z_T; \sigma)$ ▷ Contrastive loss
    $\mathcal{L}_{\text{Cap}} = \mathcal{L}_{\text{CE}}(T, V; \Theta_{\text{fusion}})$ ▷ Captioning loss
    $\mathcal{L} = \lambda_{\text{Con}}\mathcal{L}_{\text{Con}} + \lambda_{\text{Cap}}\mathcal{L}_{\text{Cap}}$ ▷ Combined loss
    Update Θ using gradient of $\mathcal{L}$
  end while
  return Trained model parameters Θ
end procedure

this capacity by integrating a large language model with the visual encoder, increasing the parameter count to approximately 7B parameters, enabling more sophisticated reasoning while maintaining a lightweight projection layer. CONCH, with approximately 200M parameters (110M for the language model and 90M for the ViT-B/16 vision encoder), introduces a CoCa-inspired decoupled decoder architecture that efficiently supports both contrastive objectives through a unified framework. This design offers computational advantages while still outperforming general-purpose VLMs on histopathology tasks, with significant performance gains on cancer subtyping and prognostic prediction compared to models without hierarchical visual processing capabilities (R. J. Chen et al., 2022; Vorontsov et al., 2024).

3.3 Benchmarking on Big-Data Cohort of Digestive Pathology

This study leverages a digestive computational pathology dataset obtained through secondary use of giga-pixel WSIs, generated during routine clinical care with ethics approval from Centre hospitalier de l'Université de Montréal (CHUM). The dataset comprises a comprehensive collection of 3,507 high-resolution WSIs in big-data form encompassing diverse tissue specimens from the digestive system with H&E staining. Each WSI is annotated on the slide level, providing rich



(a) Non-Generative VLMs (b) Generative VLMs

Figure 3.1: High level overview of the inference process for the four VLMs.

material for our vision-language modeling experiments. The dataset includes seven distinct tissue types with varying representation across classes. Colon wall (CW) specimens constitute the largest proportion (36.18%, $n = 1,269$), followed by lymph nodes (LN) (28.40%, $n = 996$), fibroadipose tissue (FT) (17.65%, $n = 619$), and small intestinal wall (SIW)(12.26%, $n = 430$). The remaining specimens include appendiceal wall (AW) (3.08%, $n = 108$), muscular colon wall (MCW) (1.28%, $n = 45$), and anastomotic or gastroduodenal junctions (GJ) between the colon and small intestine (1.14%, $n = 40$). Each WSI in the dataset is annotated in terms of the presence of invasive cancer. Table 3.1 summarizes the dataset statistics, while Fig. 3.2 presents sample WSI thumbnails from each tissue type.

Dysplasia is the abnormal growth or development of cells, tissues, or organs, typically characterized by altered size, shape, and organization. The majority of specimens in the dataset show no dysplasia (92.73%, $n = 3,252$). High-grade dysplasia is present in 5.19% ($n = 182$) of specimens, while low-grade dysplasia is detected in 2.08% ($n = 73$). This imbalance reflects real world clinical scenarios where pathological findings often represent a small subset of examined tissue. Importantly, the distribution of dysplasia varies considerably across tissue types. Dysplastic changes

Table 3.1: In-house digestive dataset statistics

Tissue Type	Distribution		Dysplasia Status			Invasiveness	
	Count	%	None	Low	High	Non-inv.	Inv.
Colon wall	1,269	36.18%	1,038	65	166	809	460
Lymph node	996	28.40%	996	0	0	845	151
Fibroadipose tissue	619	17.65%	619	0	0	521	98
Small intestinal wall	430	12.26%	413	7	10	364	66
Appendiceal wall	108	3.08%	106	0	2	104	4
Muscular colon wall	45	1.28%	44	0	1	29	16
Anastomotic junction	40	1.14%	36	1	3	21	19
Total	**3,507**	**100%**	**3,252**	**73**	**182**	**2,693**	**814**
Percentage	-	-	**92.73%**	**2.08%**	**5.19%**	**76.79%**	**23.21%**

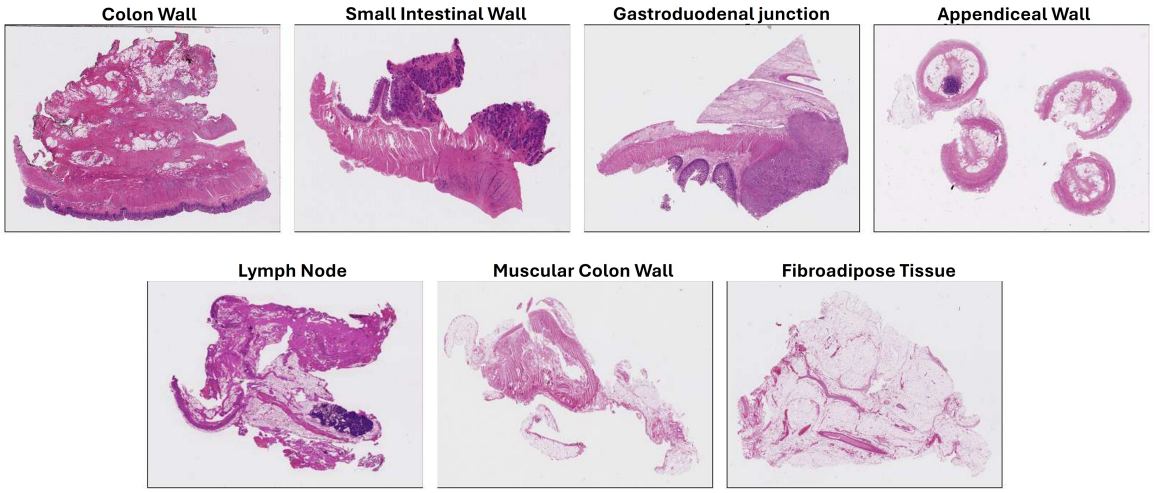


Figure 3.2: Sample images from the in-house dataset

are predominantly observed in CW specimens, where 18.20% of CW specimens exhibit some degree of dysplasia (13.08% high-grade and 5.12% low-grade). Dysplasia is also present to a lesser extent in SIW (3.95%) and GJ specimens (10.00%), while being rare or absent in LN and FT.

Invasiveness refers to the ability of abnormal cells, particularly cancer cells, to penetrate and infiltrate surrounding tissues, breaking through basement membranes and potentially spreading to distant sites. Approximately, one fourth of all specimens (23.21%, $n = 814$) exhibit invasive characteristics, while the majority (76.79%, $n = 2,693$) are non-invasive. The distribution of invasiveness varies significantly across tissue types, revealing valuable patterns for model learning. The highest rates of invasion are observed in GJ specimens (47.50%), followed by CW (36.25%), MCW (35.56%), and LN (15.16%). This finding is an important prognostic marker during routine

pathology diagnostics and is critical to clinical care as the degree of invasion into tissue layers and types are the most significant predictors of cancer aggressiveness. It presents a clinically relevant and critical task that can be addressed by VLMs to learning contextually relevant associations.

3.4 Evaluation Methodology

This section discusses the methodology followed to assess the aforementioned VLMs on the in-house digestive dataset. All models were evaluated in a zero-shot setting, using pre-trained weights from their respective base architectures without any task-specific fine-tuning. We investigate how prompt engineering affects the model's ability to identify invasiveness and cancer status across diverse digestive system tissue samples. We processed WSIs from our dataset by extracting patches at 10× and 5× magnification levels using a sliding window approach with a patch size of 512×512 pixels and 0% overlap. A small custom CNN consisting of 109,586 learnable parameters and three convolution layers is used to remove background patches from the patch dataset. This CNN was trained on images consisting of tissue and background patches collected from various dataset's such as our in-house dataset, NCT-CRC-HE (Ignatov & Malivenko, 2024) and PCAM (Veeling, Linmans, Winkens, Cohen, & Welling, 2018) so that it generalizes well and is more robust to outliers. The tissue patches selected from these datasets were random. The pipeline can be seen in Appendix Fig. A.2 .We developed a systematic prompt engineering framework based on information theory and clinical communication principles to evaluate how linguistic variations influence model performance in computational pathology tasks.

We refer to our strategy as efficient prompt design because it seeks to optimize diagnostic performance and generalization through carefully structured linguistic inputs, rather than relying on additional model fine-tuning or architecture specific engineering. Our framework explores four critical dimensions of prompt design that we hypothesize significantly impact model performance. We formalized a set of nine prompt templates by systematically varying four key dimensions: detail specificity (DS), anatomical precision (AP), instructional framing (IF), and output constraints (OC). In templates where organ-specific information is required, we use the variable O as a placeholder for the target organ being examined. DS refers to the level of granularity in the prompt,

Table 3.2: Prompt templates for histopathology Invasive classification

ID	Dimensions	Template
$P_1(O)$	***DS***: Medium ***AP***: High ***IF***: Minimal ***OC***: Explicit	The image is taken from the O using H&E staining, output only the label name which best fits the image out of the following Invasive or Non-Invasive
$P_2(O)$	***DS***: Medium ***AP***: High ***IF***: Minimal ***OC***: Explicit	The image is taken from the O using H&E staining, output only the label name which best fits the image out of the following Cancerous or Normal
$P_3(O)$	***DS***: High ***AP***: High ***IF***: Expert ***OC***: Explicit	You are an expert pathologist analyzing histopathology slides. Given an image of a tissue sample stained with Hematoxylin and Eosin (H&E) from the O and the question of classifying the presence of cancer, classify it as either 'Invasive' or 'Non-Invasive'. Provide only the single word label.
$P_4(O)$	***DS***: High ***AP***: High ***IF***: Task ***OC***: Explicit	Given an image of a tissue sample stained with Hematoxylin and Eosin (H&E) from the O, classify the existence of cancer as either 'Invasive' or 'Non-Invasive'. Provide only a single word label.
$P_5(O)$	***DS***: Medium ***AP***: High ***IF***: Task ***OC***: Explicit	Given an image of a tissue sample stained with hematoxylin and eosin from the O, identify whether the sample is cancerous or not. Provide only a single word label
P_6	***DS***: Medium ***AP***: Medium ***IF***: Task ***OC***: Explicit	Given an image of a tissue sample stained with hematoxylin and eosin from the gastrointestinal system, identify whether the sample is cancerous or not. Provide only a single word label
P_7	***DS***: Medium ***AP***: Medium ***IF***: Task ***OC***: Explicit	Given an image of a tissue sample stained with hematoxylin and eosin from the digestive system, identify whether the sample is cancerous or not. Provide only a single word label
P_8	***DS***: Medium ***AP***: Low ***IF***: Task ***OC***: Explicit	Given an image of a tissue sample stained with hematoxylin and eosin, identify whether the sample is cancerous or not. Provide only a single word label
P_9	***DS***: High ***AP***: Medium ***IF***: Expert ***OC***: Explicit	As a pathologist examining this H&E-stained digestive system tissue sample, provide your assessment of malignancy as a single word: either 'Invasive' or 'Non-Invasive'.

ranging from general to detailed instructions, taking values of LOW, MEDIUM, or HIGH. AP represents the extent to which the prompt includes precise anatomical details to make the prompt more focused, similarly taking values of LOW, MEDIUM, or HIGH. IF determines the structure of the prompt, such as posing a direct question versus providing a declarative statement, with values of EXPERT (positioning the model as a specialist), MINIMAL (providing basic instructions), or

TASK (focusing on specific objectives). OC controls the format and length of the model's response to ensure consistency, defined as either EXPLICIT (strictly defined output format) or IMPLICIT (loosely defined format). Table 3.2 presents each prompt template with its corresponding dimensional properties.

The prompt templates were strategically designed to address several research questions in medical vision-language interaction:

(1) **Information Theoretic Perspective**: We hypothesized that intermediate levels of information content in prompts (neither too sparse nor too detailed) would optimize model performance, following principles from communication theory and cognitive load theory (B. Wang, Liu, Karimnazarov, & Thompson, 2024). Prompts 3-5 were designed with varying information density to test this hypothesis.

(2) **Anatomical Specificity Gradient**: Prompts 5-8 implement a controlled degradation of anatomical specificity to quantify how precision of anatomical reference affects classification performance. This addresses a key question in medical AI regarding the importance of anatomical context in diagnostic reasoning.

(3) **Expert Role Framing**: Prompts 3 and 9 incorporate expert role assignment, a technique that has shown promise in general LLM task performance but remains under-explored in medical vision-language tasks. By positioning the model as a pathologist, we investigated whether role framing enhances performance on specialized medical tasks.

(4) **Output Constraint Consistency**: All prompts maintain explicit output constraints to isolate the effects of input prompt variations rather than confounding with output format variations.

This systematic approach to prompt design allows us to quantify the relationship between linguistic features of prompts and model performance, potentially yielding insights for optimal prompt engineering in medical vision-language applications. The decision to use binary or explicit responses in the model output is motivated by the nature of the annotations, which are provided at the slide level. As a result, the ground truth does not support more granular or localized predictions, making binary classification the most reliable and interpretable approach for this task.

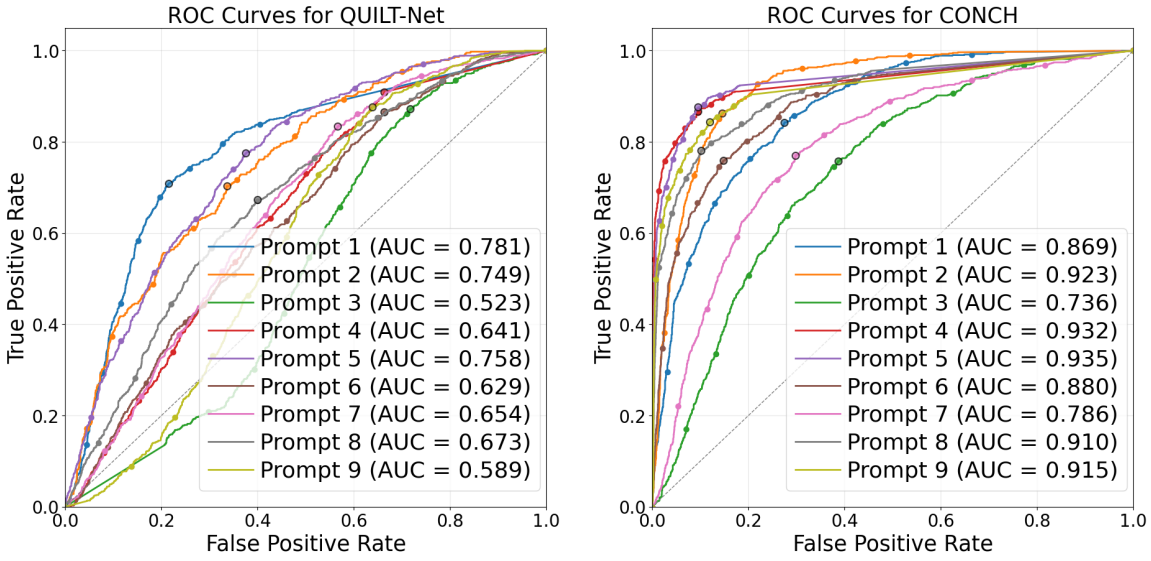
3.5 Results

This section summarizes the performance of the different VLM models under various prompts on our in-house digestive dataset. We first conduct an ablative study comparing the different performances using AUC scores. The scores for the tissue patches of a given WSI are aggregated to generate the WSI label. We analyze the effect of prompt design and model complexity on the obtained performance. We then analyze attention maps obtained by the different models on different WSIs and highlight relevant tissue regions, with feedback given by certified pathologists. All experiments were conducted on NVIDIA A100-SXM4-40GB GPUs to ensure consistent evaluation across all models. Quilt-LLAVA during inference uses a temperature of 0.1 for minimized hallucinations and less variability in output.

3.5.1 Ablative study

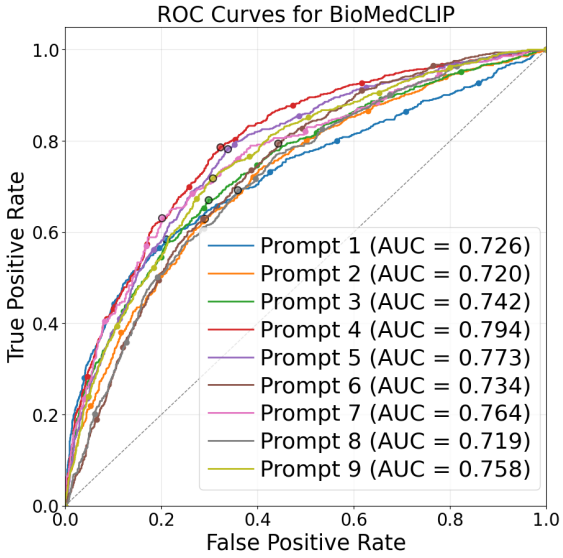
Our ablative study starts by investigating the impact of prompt formulations on model performance. Figures 3.3 and 3.4 present the AUC curves and AUC heatmap for different prompts used on Quilt-Net and CONCH, demonstrating significant performance variations based on architectural differences and prompt design choices. We evaluated model performance using AUC curves and the corresponding AUC metric, which plot true positive rates against false positive rates at various classification thresholds. The AUC metric ranges from 0 to 1, with higher values indicating superior discriminative ability.

When analyzing the model's performance, it can be seen that Quilt-Net generally shows more variance and pronounced drops in AUC scores with certain prompts. This is unlike CONCH, which is more robust to most of the changes. When comparing prompts 3 to 5, which vary in terms of information density, it is evident that more information leads to degraded performance. This can be seen in the AUC drop from 0.758 to 0.523 for Quilt-Net, and from 0.935 to 0.736 for CONCH, when going from $P_5(O)$ to $P_3(O)$. This supports our first hypothesis, presented in Section 3.4, stating that intermediate levels of general information optimize the model's performance better when compared to those that are too detailed.

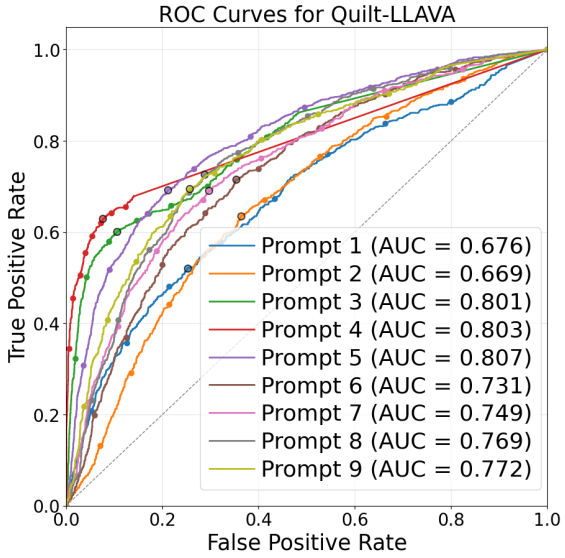


(a) Quilt-Net

(b) CONCH



(c) BioMedCLIP



(d) Quilt-LLAVA

Figure 3.3: Performance comparison of VLM models.

When comparing prompts 5 to 8, which vary in terms of anatomical specificity, the importance of precise anatomical context is evident in classification performance. For both Quilt-Net and CONCH, prompt 5 (which has high anatomical precision) achieves stronger performance (0.758 for Quilt-Net and 0.935 for CONCH) when compared to prompts 6, 7, and 8 that have lower anatomical specificity. In particular, prompt 8 results in the highest performance in this particular group (0.673

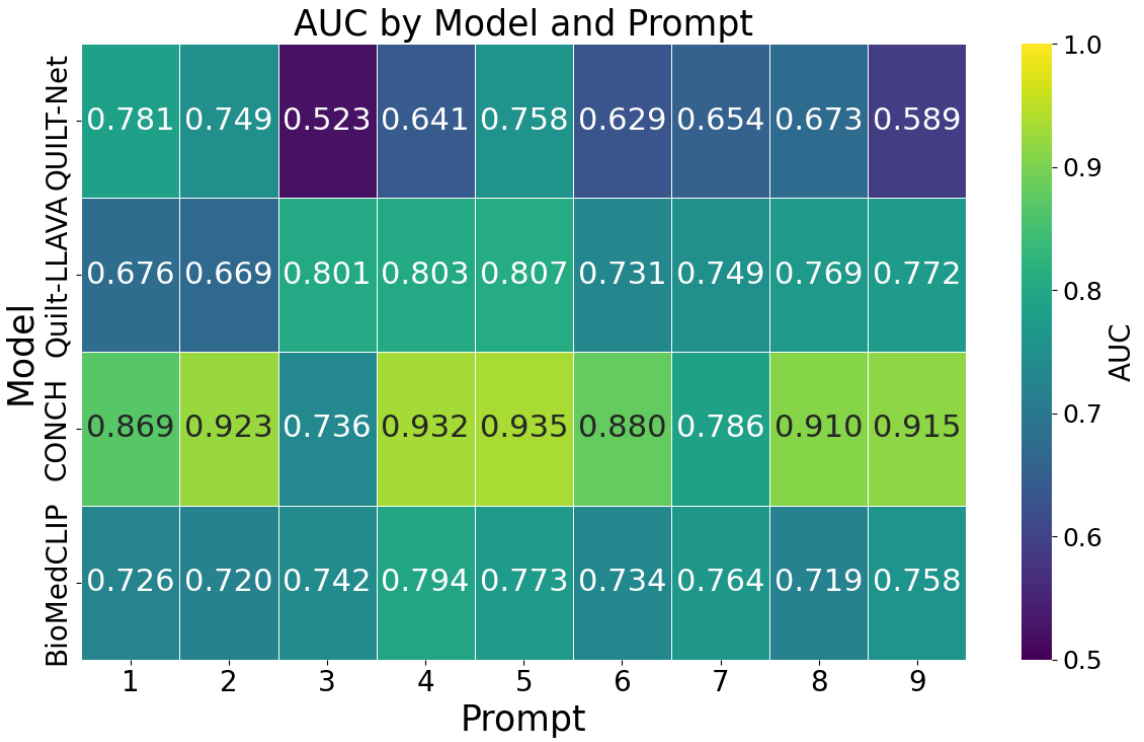


Figure 3.4: AUC heatmap displaying performance values by model and prompt configuration.

for Quilt-Net and 0.910 for CONCH). This answers our second research question regarding the importance of anatomical context in diagnostic reasoning. When evaluating the impact of expert role framing (prompts 3 and 9), we observe a negative or neutral effect on performance. Specifically, in Quilt-Net, prompts 3 and 9 show the worst performance, with AUC scores of 0.523 and 0.589, respectively. Similarly, prompt 3 gives the lowest performance in CONCH (AUC = 0.736), while prompt 9 shows neutral behavior with no improvement (AUC = 0.915). This suggests that framing the model as an expert does not necessarily enhance the model's performance and may even introduce unnecessary complexity that misleads the model's pre-trained embeddings.

When analyzing the model's performance, Quilt-LLAVA demonstrates moderate sensitivity to prompt variation, with AUC values ranging from 0.669 to 0.807. Compared to Quilt-Net, which exhibited severe performance drops under detailed or poorly framed prompts, Quilt-LLAVA maintains relatively stable performance, indicating a stronger capacity to parse complex linguistic input. When evaluating prompts 3 to 5, which differ in information density, Quilt-LLAVA performs best

with prompt $P_5(O)$ (AUC = 0.807), which features medium detail specificity, high anatomical precision, and task-oriented framing. Performance slightly declines with prompts $P_4(O)$ and $P_3(O)$ (AUC = 0.803 and 0.801, respectively), suggesting that overly detailed or expert-oriented framing introduces mild cognitive overhead, but not to the extent observed in Quilt-Net or CONCH. This behavior also supports our first hypothesis that intermediate information density (as in $P_5(O)$) strikes the best balance for this model, minimizing ambiguity without overloading the LLM with excessive detail. When comparing prompts $P_5(0)$ through P_8, which systematically degrade anatomical specificity, Quilt-LLAVA again exhibits patterns aligned with our second hypothesis. Prompt $P_5(O)$, containing explicit organ level context, produces the highest performance (AUC = 0.807), while prompts P_6, P_7, and P_8 progressively lower anatomical granularity and correspondingly result in declining performance (AUCs of 0.731, 0.749, and 0.769, respectively). The relatively contained drop in AUC (from 0.807 to 0.731) highlights Quilt-LLAVA's resilience but still underscores the value of anatomical precision in model comprehension and decision-making.

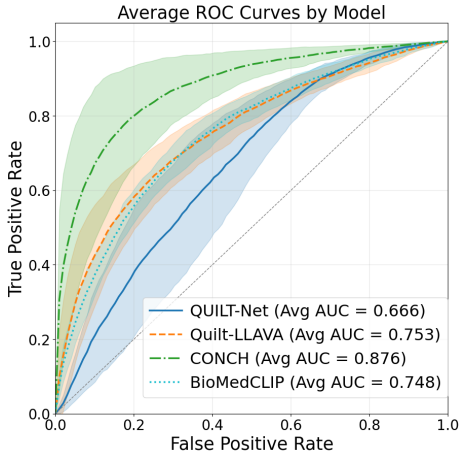
Regarding expert role framing, prompts $P_3(O)$ and $P_9(O)$ yield AUCs of 0.801 and 0.772, respectively, reflecting a minimal performance change relative to task based prompts. Unlike Quilt-Net or CONCH, which either deteriorate or remain neutral under expert framing, Quilt-LLAVA exhibits tolerance to this linguistic shift. However, it does not appear to benefit meaningfully from it. This suggests that while instruction tuning allows Quilt-LLAVA to handle more human-like instructions, the addition of expert persona framing does not further enhance its capacity for specialized tasks in computational pathology.

In contrast, BioMedCLIP exhibits the narrowest AUC range among all models, with scores spanning from 0.719 to 0.794, indicating low sensitivity to prompt variation. When comparing prompts $P_3(O)$ through $P_5(O)$, which differ in information density, we observe that prompt $P_4(O)$ achieves the best performance (AUC = 0.794), followed by $P_5(O)$ (AUC = 0.773), and $P_3(O)$ (AUC = 0.742). This ordering partially supports our first hypothesis, where an intermediate level of information (as in $P_4(O)$, which uses high detail specificity with task oriented framing) results in optimal performance. Unlike Quilt-Net and CONCH, however, BioMedCLIP does not experience drastic performance drops under more verbose or expert framed prompts, suggesting a more flattened sensitivity curve and weaker interaction between prompt verbosity and model performance. When

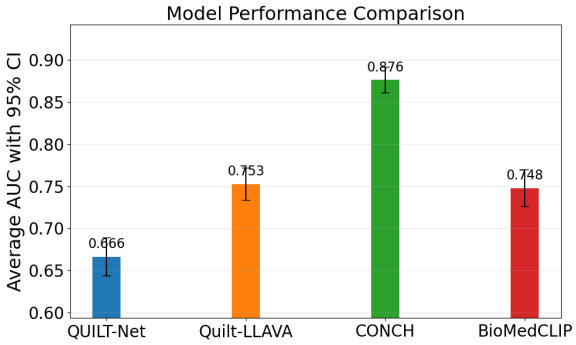
analyzing prompts $P_5(O)$ through P_8, which vary in anatomical specificity, we observe a gradual performance decline with decreasing anatomical precision. Prompt $P_5(O)$, which includes explicit organ level information, yields an AUC of 0.773, while prompts P_6, P_7, and P_8 result in slightly lower AUCs of 0.734, 0.764, and 0.719, respectively. The relatively shallow decline across this set indicates that although anatomical specificity contributes positively, the effect is less pronounced in BioMedCLIP compared to models with more tightly coupled vision-language alignment such as CONCH. This trend is consistent with our second hypothesis but highlights the model's overall insensitivity to changes in anatomical granularity.

With respect to expert role framing, prompts $P_3(O)$ and P_9 lead to modest gains over the least specific prompts, with AUCs of 0.742 and 0.758, respectively. This behavior contrasts with Quilt-Net, which exhibits a substantial drop under the same conditions. In BioMedCLIP, the addition of an expert persona appears to have a mildly stabilizing effect without significantly enhancing or diminishing performance. This neutrality suggests that BioMedCLIP's pre-training on broad biomedical data provides a general robustness but lacks the fine-tuned responsiveness to task framing found in instruction-tuned models. Overall, BioMedCLIP's consistent yet modest performance across prompt variants indicates a prompt-agnostic architecture that operates effectively within a narrow band of performance. While it benefits marginally from clearer anatomical context and moderate instruction design, it does not fully capitalize on optimized prompt structures. This behavior likely stems from its contrastive training paradigm on a large scale biomedical corpus, which confers general visual-text alignment without pathology-specific specialization. Consequently, BioMedCLIP remains stable across variations but does not exhibit the dynamic range seen in more domain adapted models.

Figure 3.5 analyzes the average performance (AUC) of the four VLMs. As seen in the figure, CONCH achieves the highest average AUC (0.876), followed by Quilt-LLAVA (0.753), BioMedCLIP (0.748), and Quilt-Net (0.666). The underperformance by Quilt-Net is expected, given it is the smallest model compared to the others. However, despite Quilt-LLAVA being the largest model , it does not outperform CONCH . This suggests that model scale alone is not a dominant factor in performance and that domain-specific training and vision-language alignment have crucial roles. While Quilt-LLAVA uses instruction tuning and a powerful LLM , it is constrained by suboptimal



(a) Average ROC curves by model with confidence intervals.



(b) Average AUC performance comparison.

Figure 3.5: Performance comparison of VLM models.

domain alignment between its visual encoder and the LLM for computational pathology. This is unlike CONCH, which uses a contrastive learning approach specifically tuned on histopathology image-text pairs, allowing for better generalizability.

Collectively, these observations reinforce our initial hypotheses: (1) intermediate information density optimizes model performance, (2) anatomical specificity significantly enhances classification accuracy, and (3) expert role framing does not reliably improve and may even degrade results. The varied prompt responsiveness across architectures also suggests that robust vision-language alignment and pathology-specific pre-training are more impactful than raw parameter count or instruction-following capacity alone.

Table 3.3 summarizes the AUC performance of all models across digestive system tissues for each prompt. CONCH consistently achieves strong results across organs, notably attaining perfect classification for GJ under prompt P_6 (AUC = 1.000). Quilt-LLAVA shows stable high performance on GJ and CW across prompts, while BioMedCLIP performs uniformly but with a lower ceiling, including a major drop on AW with prompt P_7 (AUC = 0.181). QUILT-Net exhibits high variability, performing well on MCW and CW under optimal prompts, but underperforming significantly on AW and LN. These trends highlight both the strengths and limitations of each model in organ-specific classification, especially for low-frequency or histologically complex tissues.

Table 3.3: VLM Performance on Digestive System Tissue Cancer Classification

Prompt	Model	Organs (AUC)						
		CW	**SIW**	**GJ**	**AW**	**LN**	**MCW**	**FT**
P_1(O)	QUILT-Net	0.944	0.814	0.970	0.257	0.690	0.813	0.802
	Quilt-LLAVA	0.657	0.507	0.982	0.594	0.682	0.651	0.544
	CONCH	0.930	0.894	0.937	0.483	0.830	0.903	0.797
	BioMedCLIP	0.877	0.553	0.579	0.542	0.538	0.565	0.444
P_2(O)	QUILT-Net	0.703	0.781	0.974	0.827	0.842	0.989	0.800
	Quilt-LLAVA	0.718	0.798	0.987	0.452	0.655	0.862	0.734
	CONCH	0.942	0.886	0.967	0.488	0.952	0.823	0.874
	BioMedCLIP	0.823	0.842	0.892	0.649	0.755	0.800	0.730
P_3(O)	QUILT-Net	0.789	0.724	0.815	0.778	0.533	0.547	0.759
	Quilt-LLAVA	0.863	0.622	0.898	0.547	0.750	0.902	0.760
	CONCH	0.818	0.806	0.920	0.445	0.800	0.713	0.614
	BioMedCLIP	0.817	0.733	0.860	0.712	0.553	0.888	0.602
P_4(O)	QUILT-Net	0.607	0.649	0.642	0.481	0.716	0.440	0.714
	Quilt-LLAVA	0.910	0.726	0.974	0.500	0.654	0.875	0.577
	CONCH	0.962	0.994	0.977	0.606	0.938	0.886	0.823
	BioMedCLIP	0.848	0.877	0.945	0.659	0.716	0.782	0.793
P_5(O)	QUILT-Net	0.872	0.759	0.947	0.550	0.633	0.782	0.840
	Quilt-LLAVA	0.814	0.905	0.995	0.447	0.721	0.882	0.820
	CONCH	0.974	0.972	0.980	0.635	0.921	0.897	0.855
	BioMedCLIP	0.830	0.737	0.925	0.469	0.730	0.807	0.799
P_6	QUILT-Net	0.579	0.548	0.759	0.613	0.596	0.694	0.813
	Quilt-LLAVA	0.752	0.667	0.952	0.517	0.689	0.780	0.782
	CONCH	0.954	0.961	1.000	0.477	0.761	0.944	0.842
	BioMedCLIP	0.744	0.483	0.846	0.736	0.788	0.591	0.768
P_7	QUILT-Net	0.807	0.617	0.947	0.502	0.509	0.843	0.629
	Quilt-LLAVA	0.798	0.721	0.989	0.427	0.700	0.827	0.776
	CONCH	0.890	0.750	0.995	0.443	0.639	0.856	0.793
	BioMedCLIP	0.858	0.845	0.974	0.181	0.620	0.893	0.587
P_8	QUILT-Net	0.685	0.572	0.912	0.556	0.603	0.781	0.776
	Quilt-LLAVA	0.828	0.740	0.985	0.511	0.678	0.824	0.830
	CONCH	0.937	0.845	0.992	0.611	0.910	0.968	0.845
	BioMedCLIP	0.767	0.566	0.892	0.550	0.669	0.759	0.767
P_9	QUILT-Net	0.851	0.685	0.962	0.337	0.397	0.836	0.579
	Quilt-LLAVA	0.798	0.828	0.957	0.448	0.699	0.879	0.850
	CONCH	0.971	0.971	0.992	0.466	0.831	0.878	0.871
	BioMedCLIP	0.841	0.749	0.957	0.435	0.651	0.802	0.745

Following our investigation of prompt engineering for cancer detection, we extended our ablative study to examine the more nuanced task of dysplasia classification. The subtle architectural and cytological alterations in dysplastic tissues often exist on a continuum with reactive changes, making their computational detection particularly challenging yet clinically crucial for early intervention.

To assess prompt wording effects on dysplasia detection, we conducted an ablative experiment using three prompt variants derived from the base prompt P_5(O) since it is the best performing

Table 3.4: Prompt templates for histopathology dysplasia classification

ID	Dimensions	Template
$D_1(O)$	***DS***: Medium ***AP***: High ***IF***: Task ***OC***: Explicit	Given an image of a tissue sample stained with hematoxylin and eosin from the O, identify whether the sample is Dysplasia or Benign. Provide only a single word label
$D_2(O)$	***DS***: Medium ***AP***: High ***IF***: Task ***OC***: Explicit	Given an image of a tissue sample stained with hematoxylin and eosin from the O, identify whether the sample is Atypia or Benign. Provide only a single word label
$D_3(O)$	***DS***: Medium ***AP***: High ***IF***: Task ***OC***: Explicit	Given an image of a tissue sample stained with hematoxylin and eosin from the O, identify whether the sample is Precancerous or Benign. Provide only a single word label

prompt as observed from previous results. In each variant, the key term for the target pathology was changed while keeping all other prompt aspects constant (medium detail specificity, high anatomical precision, task-oriented instruction, and explicit output constraints). Specifically, D_1(O) used the term dysplasia, D_2(O) replaced it with atypia, and D_3(O) used precancerous. These synonyms describe the same precancerous condition but differ in technical tone as seen in Table 3.4. The task for each vision-language model remained identifying dysplasia in images given the prompt. Performance was evaluated by AUC, summarized in the Fig. 3.6.

Each model exhibited a distinct sensitivity to the terminology. QUILT-Net performed best with the original term dysplasia, achieving an AUC of about 0.711 with D_1(O). Using the synonym atypia substantially lowered QUILT-Net's performance (AUC = 0.607 for D_2(O)), indicating this model is tuned to the exact pathological term. Its AUC with precancerous (D_3(O)) was intermediate (AUC = 0.664), suggesting that while QUILT-Net can partly understand the term, it still prefers the standard medical vocabulary. CONCH and BioMedCLIP, both domain pretrained models, similarly showed stronger results with pathology specific wording. CONCH attained its highest AUC (= 0.904, the highest among all models) when prompted with atypia, but its performance dropped when the prompt used dysplasia. BioMedCLIP likewise performed best with technical terminology, it favored atypia (AUC = 0.832) over dysplasia (AUC = 0.817) and saw a notable decline to about 0.684 AUC with the precancerous. In contrast, the Quilt-LLAVA model exhibited the opposite trend. Quilt-LLAVA struggled with the highly technical prompt, obtaining its lowest (AUC =

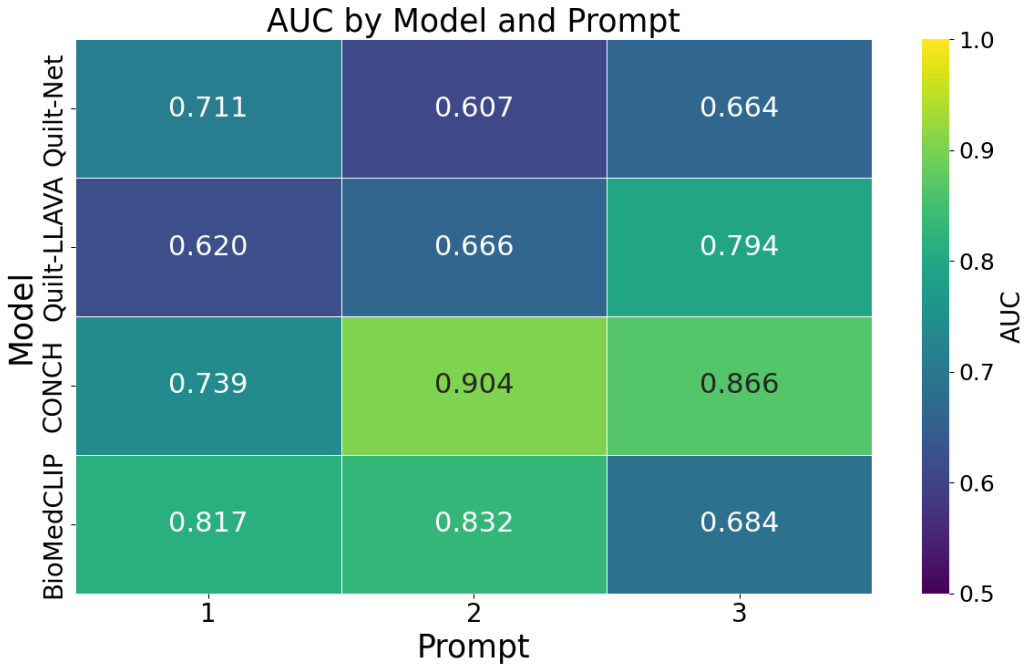
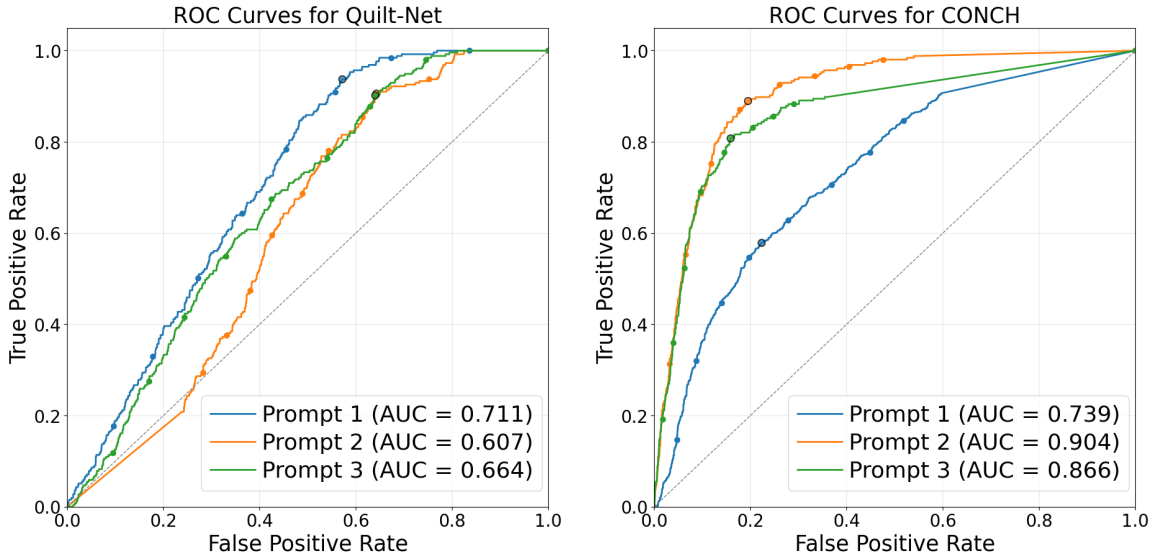


Figure 3.6: AUC heatmap displaying performance values by model and prompt configuration for dysplasia classification

0.620) with D_1(O) dysplasia, but its performance improved markedly with more colloquial phrasing. Using precancerous in the prompt boosted Quilt-LLAVA's AUC to approximately 0.794, the highest for this model. This gain highlights that the LLM-based vision model benefited from laymen terminology that aligned with its general language understanding. These AUC curves give a more detailed understanding of these models behave as observed in Fig. 3.7

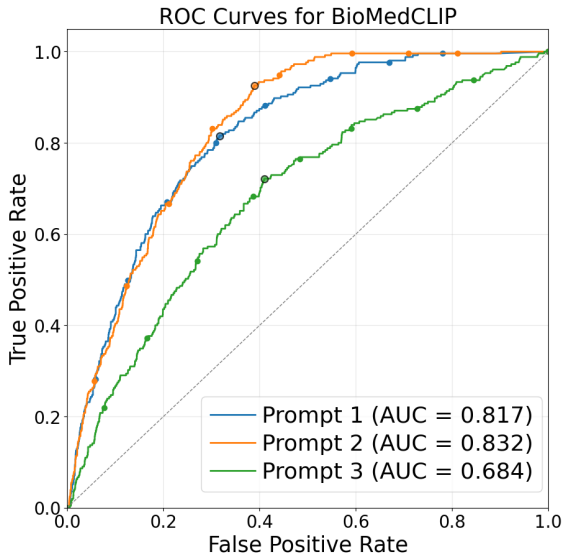
In summary, the dysplasia detection results show that prompt terminology can significantly influence model performance. Each model demonstrates a preferred vocabulary reflecting its training and design: the two contrastive models (CONCH and BioMedCLIP) and QUILT-Net perform optimally with domain-specific terms, whereas the multi-modal Quilt-LLAVA requires more accessible language for best results. Notably, all models handled the concept of dysplasia to some extent, but their AUCs varied by approximatley 0.20 across the three wording variants as observed from Fig. 3.8. These findings underscore the importance of prompt calibration for each model. The preferred terminology for dysplasia thus differs by model, and choosing the right prompt phrasing yields a measurable improvement in AUC for this task.

Table 3.5 presents tissue-level AUC performance for each model across the three dysplasia

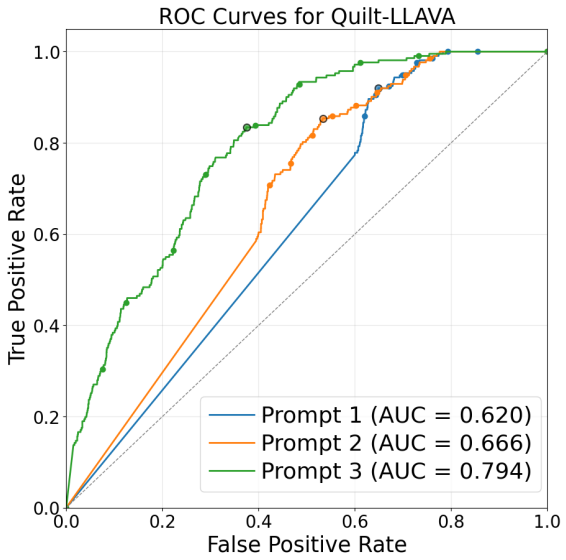


(a) Quilt-Net

(b) CONCH



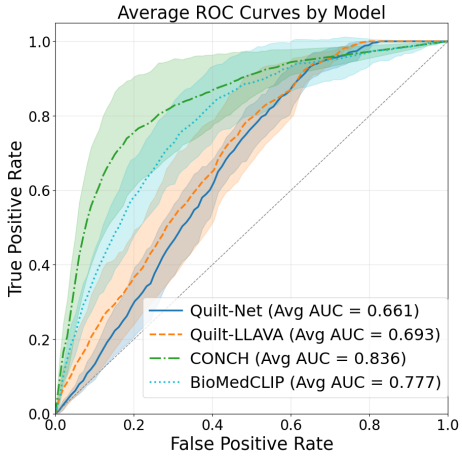
(c) BioMedCLIP



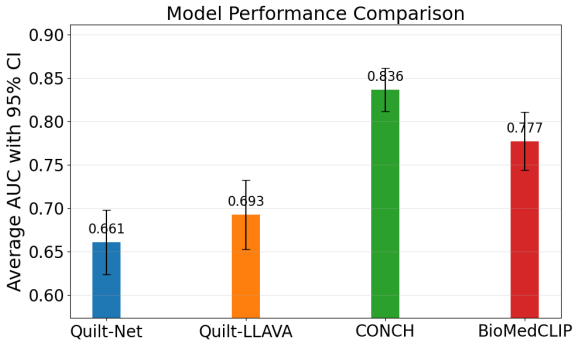
(d) Quilt-LLAVA

Figure 3.7: Performance comparison of VLM models on dysplasia.

prompts. CONCH consistently achieves the highest scores, particularly on GJ (AUC = 0.982 with $D_2(O)$) and MCW (AUC = 0.870 with $D_1(O)$). Quilt-LLAVA performs best on GJ under $D_3(O)$ (AUC = 0.887), while BioMedCLIP shows strong results for MCW (AUC = 0.860) and FT (AUC = 0.789) using $D_3(O)$ and $D_2(O)$, respectively. QUILT-Net's performance varies significantly across tissues and prompts, with its highest AUC (0.737) on GJ using $D_2(O)$. These results reinforce the importance of both anatomical context and terminology alignment in prompt design for dysplasia



(a) Average ROC curves by model with confidence intervals.



(b) Average AUC performance comparison

Figure 3.8: Performance comparison of VLM models on dysplasia.

Table 3.5: VLM performance on Digestive System Tissue Dysplasia Classification

Prompt	**Model**	**Organs (AUC)**						
		AW	**CW**	**GJ**	**MCW**	**LN**	**FT**	**SIW**
$D_1(O)$	QUILT-Net	0.724	0.307	0.148	0.263	0.642	0.578	0.309
	Quilt-LLAVA	0.413	0.630	0.729	0.586	0.521	0.593	0.573
	CONCH	0.815	0.736	0.947	0.870	0.757	0.776	0.607
	BioMedCLIP	0.695	0.585	0.702	0.526	0.675	0.723	0.302
$D_2(O)$	QUILT-Net	0.462	0.467	0.737	0.534	0.394	0.528	0.331
	Quilt-LLAVA	0.399	0.542	0.709	0.569	0.538	0.668	0.596
	CONCH	0.578	0.854	0.982	0.753	0.647	0.739	0.755
	BioMedCLIP	0.728	0.671	0.692	0.478	0.691	0.789	0.448
$D_3(O)$	QUILT-Net	0.649	0.406	0.396	0.552	0.600	0.623	0.308
	Quilt-LLAVA	0.522	0.621	0.887	0.668	0.698	0.784	0.582
	CONCH	0.550	0.906	0.947	0.845	0.858	0.734	0.792
	BioMedCLIP	0.498	0.714	0.617	0.860	0.703	0.658	0.558

detection. Another important factor is that the prompts fed to the non-generative VLMs models being used for our experiments expect each input prompt to represent a different class; hence the prompt template has to be decomposed such that each prompt represents only one class and the number of input prompts to these non-generative VLMs is equal to the number of classes, whereas generative VLMs expect all the class names to be within the same prompt.

Figure 3.9 analyzes the effect of magnification levels on the model's performance. Generally, it can be seen that better performance is achieved with higher magnification, as it aids in capturing finer morphological structures, which could be needed for accurate classifications. It is also evident

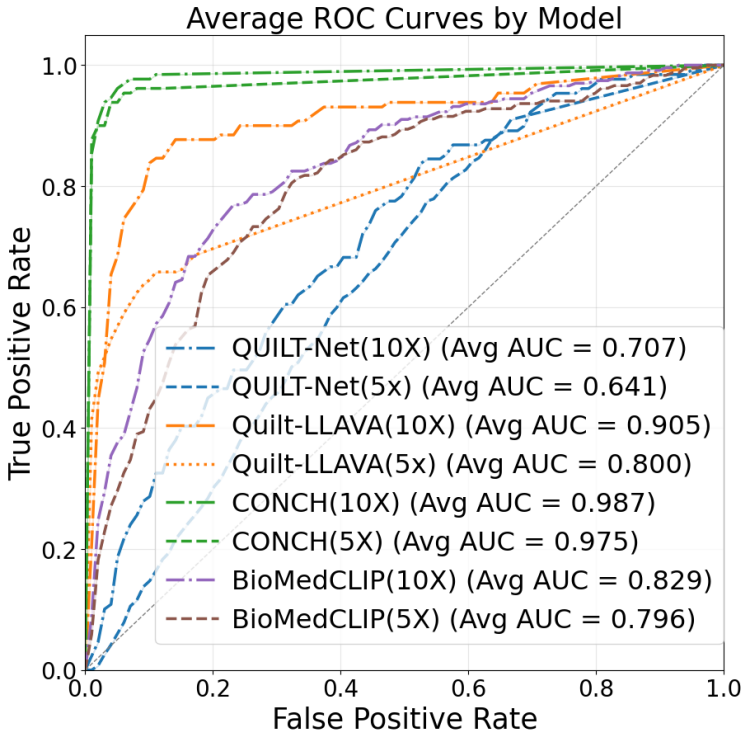


Figure 3.9: Average AUC curves comparing model performance at different magnification levels.

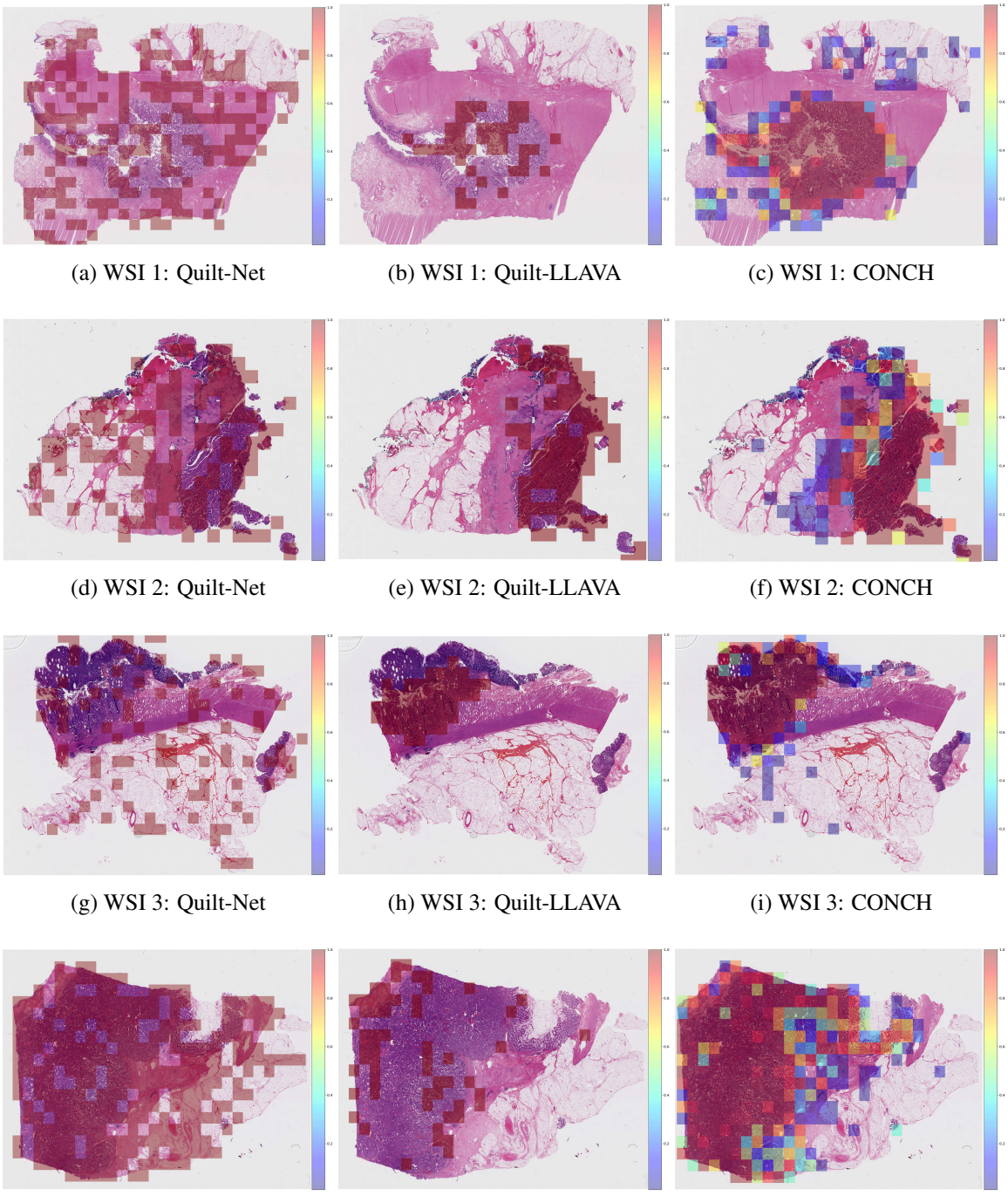
that the performance in CONCH is nearly similar regardless of the magnification level, indicating that CONCH is more robust to changes in resolution. The experiments on different magnification levels was conducted by randomly sampling 1000 WSIs from our in-house dataset.

3.5.2 Attention Maps Analysis

Figure 3.10 presents attention maps generated for histopathological analysis of randomly selected invasive cancer tissue samples. The heatmaps for BioMedCLIP, shown in Appendix Fig. A.1 are generated from a different set of whole slide images to supplement the visual analysis with additional examples. While these slides differ from those used for the other models, the heatmaps provide valuable qualitative insights into BioMedCLIP's attention mechanisms. These visualizations represent probability scores assigned at the patch level by the different models to various regions of WSI. The process begins with segmenting each WSI into patches and classifying them as either tissue or background using an in-house CNN. Tissue patches are then processed through the models to obtain probability scores. By default, Quilt-Net and CONCH generate continuous confidence scores, which are used to represent each patch in the WSI and construct a heatmap. In contrast, Quilt-LLAVA does not inherently produce such scores; therefore, binary labels are used to highlight patches or regions identified as invasive. After scoring, the patches are reconstructed with their corresponding probability values to create comprehensive attention maps that highlight regions of interest across the entire WSI, potentially indicating areas of malignancy or specific tissue characteristics. We selected 4 WSIs that cover a diversity of forms of colorectal cancer invading into different levels of depth into the colonic wall.

All four models displayed different attention behavior in the underlying images. Quilt-Net randomly identified high-attention areas throughout the image, with no significant shift towards areas containing invasive cancer versus areas with cancer. Quilt-LLAVA displayed high-attention patches found within the invasive cancer, but was rather inconsistent in its approach as some areas of the invasive cancer was not highlighted. However, most high attention maps were accurately identified within cancer. CONCH showed the most accurate attention maps of invasive cancer and consistently highlighted its presence throughout the patches. CONCH was more precise in all images but highlighted low-attention areas that were distant from the cancer. CONCH could also highlight at medium-level attention areas of a precursor lesion that is on the verge of becoming cancer and altered tissue areas adjacent to the invasive cancer. Overall, per the review of a board-certified pathologist, CONCH most accurately mimicked the general approach by pathologists in

addressing these tissues. Most attention is drawn towards the invasive cancer area, and second-order areas are revised to detect relevant findings, such as precursor lesions and mild changes in the peritumoral area that can be relevant for invasive cancer.



(a) WSI 1: Quilt-Net (b) WSI 1: Quilt-LLAVA (c) WSI 1: CONCH

(d) WSI 2: Quilt-Net (e) WSI 2: Quilt-LLAVA (f) WSI 2: CONCH

(g) WSI 3: Quilt-Net (h) WSI 3: Quilt-LLAVA (i) WSI 3: CONCH

(j) WSI 4: Quilt-Net (k) WSI 4: Quilt-LLAVA (l) WSI 4: CONCH

Figure 3.10: Comparison of different models across multiple WSI samples

Below are detailed analysis on the four WSIs from the board-certified pathologist:

- WSI 1 represents a diverticular disease which has progressed into invasive cancer that breaches into the muscularis propria. Quilt-Net targets the whole colon wall with no preference for the invasive cancer versus the non-invaded areas. Quilt-LLAVA targets the invasive cancer and peri-invasive cancer area accurately. CONCH gives high attention at the invasive cancer consistently and highlights at medium attention the precursor area in the epithelium and the affected peri-cancer areas. It notes at low attention the unaffected normal tissue further away.

- WSI 2 represents a classical invasive cancer that reached the resection margin and invades into the subserosal connective tissue. Quilt-Net produces randomized high-attention area throughout the image, Quilt-LLAVA accurately targets the invasive cancer, and CONCH shows high-attention for invasive cancer, medium-attention for the affected pericancer areas, and low-attention to areas without cancer.

- WSI 3 represents a classical invasive cancer that is restricted to the muscularis propria, arising from a precursor adenoma. Quilt-Net gives randomized high-attention area throughout the image, Quilt-LLAVA targets the cancer area while ignoring the precursor lesion, and CONCH targets the cancer area accurately, and at medium- attention the precursor lesion. It further gives low attention to the non-invasive area.

- WSI 4 represents a very large cluster of cancer with reactive epithelium at the surface. It invades deeply into the wall, into the subserosal connective tissue. Quilt-Net targets the invasive cancer a bit more, but large areas of rather non-invaded tissues. Quilt-LLAVA seems to highlight the cancer, but only in areas that are adjacent to the non-tumoral tissues. CONCH accurately targets the cancer but appears to give low attention to an area of the cancer that is less aggressive while overcalling the reactive epithelium that overlies the cancer.

Chapter 4

Conclusion and Future direction

4.1 Conclusion

This study has systematically investigated the impact of prompt engineering on the zero-shot diagnostic capabilities of VLMs in computational pathology. Through our comprehensive analysis of BioMedCLIP, Quilt-Net, Quilt-LLAVA, and CONCH on a large-scale dataset of 3,507 digestive system whole slide images, we have established several key findings that advance the field's understanding of how VLMs can be effectively deployed for pathological diagnosis. Our results demonstrate that prompt design significantly influences model performance, with anatomical precision emerging as a critical factor in diagnostic accuracy. The consistent degradation in performance observed when reducing anatomical specificity highlights the importance of domain-specific contextual cues in guiding model attention toward diagnostically relevant features. This finding parallels the diagnostic process of human pathologists, who rely on precise anatomical context to interpret histological patterns correctly. The comparative performance analysis across models revealed that CONCH consistently outperformed BioMedCLIP, Quilt-Net and the significantly larger Quilt-LLAVA model, achieving an impressive AUC with optimally formulated prompts. This suggests that domain-specific architectural design and training approaches are more crucial than raw parameter count for computational pathology applications. CONCH's superior performance can be attributed to its effective contrastive learning strategy specifically tuned on histopathology image-text pairs, which appears to enable better generalization across diverse tissue types. BioMedCLIP

demonstrated strong generalization capabilities across tissue types, highlighting the effectiveness of large-scale contrastive pretraining in the biomedical domain. Its consistent performance reinforces the value of domain-aligned pretraining even in zero-shot diagnostic settings. Our attention map analysis, validated by expert pathologists, provided visual confirmation that models with superior quantitative metrics also demonstrated more clinically relevant attention patterns. CONCH consistently highlighted invasive cancer regions with high attention, while also appropriately identifying precursor lesions and affected peri-cancer areas with medium attention, a pattern that closely resembles the diagnostic approach of human pathologists. This alignment between quantitative performance and qualitative attention distribution strengthens confidence in the clinical relevance of our findings. The observation that information density in prompts affects model performance supports an information-theoretic perspective of prompt engineering, where intermediate levels of detail yield optimal results. Similarly, the finding that expert role framing did not enhance (and sometimes degraded) model performance challenges assumptions about effective prompt strategies in specialized domains. These insights establish foundational guidelines for implementing VLMs in computational pathology workflows. By systematically optimizing prompts with appropriate anatomical precision and information density, while leveraging models with domain-appropriate architectures, researchers and developers can significantly enhance diagnostic accuracy in zero-shot settings. This approach holds particular promise for rare pathologies where annotated training data is limited. Future work should extend these findings across broader tissue types and more complex diagnostic tasks, incorporate multimodal data sources, and develop interactive systems that enable pathologists to iteratively refine prompts during diagnostic sessions. Additionally, exploring approaches to enhance model robustness to staining variations and further improving explainability will be crucial for clinical adoption. In conclusion, this research demonstrates that VLMs, when guided by carefully engineered prompts, can achieve impressive diagnostic accuracy in computational pathology. The established relationship between prompt design, model architecture, and diagnostic performance provides a solid foundation for developing more robust, interpretable, and clinically viable AI systems that can augment pathologist capabilities and ultimately improve patient care. In conclusion, this research demonstrates that VLMs, when guided by carefully engineered prompts, can achieve impressive diagnostic accuracy in computational pathology. The established

relationship between prompt design, model architecture, and diagnostic performance provides a solid foundation for developing more robust, interpretable, and clinically viable AI systems that can augment pathologist capabilities and ultimately improve patient care.

4.2 Future Direction

Our systematic investigation of VLMs for zero-shot diagnostic pathology has revealed several promising directions for future research that can rely upon our key findings. The significant performance variations observed across different prompt structures and model architectures point to critical areas for advancement in computational pathology. The marked impact of anatomical precision on model performance suggests that further refinement of prompt engineering approaches could yield substantial improvements in diagnostic accuracy. Future research should develop comprehensive anatomical reference frameworks that can be systematically incorporated into prompts. This approach would extend beyond simple organ identification to include detailed tissue layers, cell types, and histological structures relevant to specific diagnostic tasks. Additionally, researchers should investigate algorithmic approaches to automatically generate and optimize prompts based on specific tissue types and diagnostic contexts. Such systems could employ meta-learning techniques to identify prompt patterns that maximize diagnostic accuracy across diverse pathological conditions, potentially discovering prompt structures that outperform manually crafted ones.

The creation of standardized prompt libraries optimized for different pathological tasks would also benefit the broader research community. These libraries could serve as benchmarks for evaluating new models and establishing consistent reporting standards across studies. By sharing optimized prompts for tasks such as cancer detection, subtyping, and grading, researchers could accelerate progress while maintaining comparability across different institutional implementations. Our finding that CONCH outperformed the significantly larger Quilt-LLAVA model highlights the importance of domain-specific architectural design over raw parameter count. Future architectural research should focus on specialized pre-training strategies exclusively tailored to histopathological data. This approach might include developing novel self-supervised learning objectives that specifically target the identification of diagnostically relevant tissue patterns and cellular arrangements.

Pre-training on expanded histopathology-specific datasets would likely yield further improvements, particularly if these datasets encompass diverse organ systems and pathological conditions.

The observed variations in performance across magnification levels underscore the need for architectures that explicitly incorporate multi-resolution analysis capabilities. Future models should implement parallel processing streams that simultaneously leverage information from different magnification levels, integrating these diverse perspectives through cross-scale attention mechanisms. This approach would more closely mirror the diagnostic process of human pathologists, who routinely navigate between low and high magnification views during assessment. Computational efficiency remains a significant challenge when processing gigapixel whole slide images. Future research should focus on developing architectures that can efficiently analyze such large-scale images without sacrificing diagnostic accuracy. Promising approaches include hierarchical patch processing strategies, region-of-interest identification systems, and attention mechanisms that prioritize diagnostically relevant regions based on preliminary low-resolution scans.

While our study focused on digestive pathology, future work should validate and extend these approaches across diverse organ systems. This cross-organ investigation would reveal whether the prompt engineering strategies identified in our work generalize to different tissue contexts or whether organ-specific adaptations are necessary. The systematic examination of prompt transferability across tissue types could yield valuable insights about the fundamental visual-semantic relationships in histopathology that transcend specific anatomical contexts. Current pathology datasets typically overrepresent common conditions while underrepresenting rare pathologies. Future research should address this imbalance by incorporating more examples of rare cancers and uncommon histological variants. When direct data collection is limited by the inherent rarity of these conditions, synthetic data generation approaches could be employed to augment available samples.

Future studies should also move beyond binary invasive and non-invasive classification to address more complex diagnostic tasks. These include multi-class grading schemes, prediction of molecular subtypes from morphological features, assessment of treatment response indicators, and identification of prognostic markers. Each of these more nuanced diagnostic tasks will likely require specialized prompt engineering approaches that carefully structure the model's attention and reasoning process. The promising performance demonstrated in our study warrants prospective clinical

validation of optimally-prompted VLMs in real-world settings. Future clinical studies should focus particularly on challenging cases where inter-observer variability among pathologists is high, as these represent scenarios where computational assistance could provide the greatest clinical value.

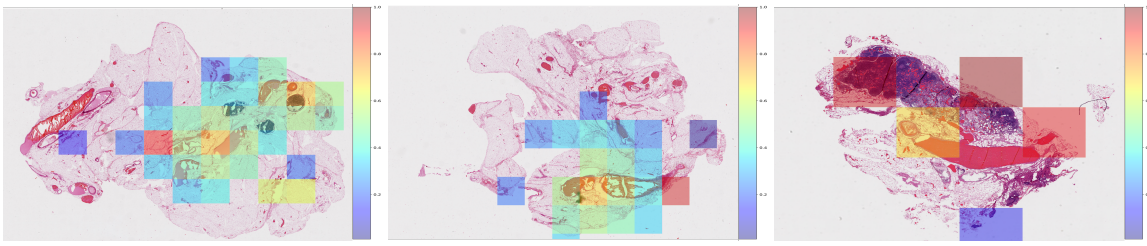
The sensitivity of VLMs to prompt design could be leveraged as a feature rather than a limitation through interactive systems where pathologists iteratively refine model prompts during diagnostic sessions. Such human-AI collaborative workflows would combine the pathologist's domain expertise with the model's computational capabilities, potentially yielding diagnostic accuracy superior to either alone. For clinical adoption, improved explainability remains essential. Future research should extend attention map approaches to provide more granular and interpretable explanations of diagnostic reasoning. These might include natural language explanations that link model decisions to specific histological features, comparison visualizations with reference cases, and uncertainty quantification for different aspects of the diagnosis.

Variations in tissue preparation and staining protocols represent a significant challenge for computational pathology systems deployed across different laboratories and institutions. Future work should investigate how prompt engineering can address these variations, potentially through specific prompt components that instruct models to account for staining intensity differences or preparation artifacts. The ability to rapidly adapt models to new diagnostic contexts with minimal examples would significantly enhance clinical utility. Future research should explore how few-shot learning can be incorporated into prompts to quickly adapt pre-trained models to novel pathologies or rare variants. Finally, the integration of additional data modalities beyond H&E images represents an important frontier. Future systems should incorporate immunohistochemistry results, molecular profiles, and relevant clinical metadata through multimodal prompt designs.

The promising results of our current study, particularly the strong performance of CONCH with anatomically precise prompts, demonstrate that VLMs hold significant potential for computational pathology. By addressing these future research directions, the field can move toward developing robust, interpretable diagnostic systems that enhance pathologists' efficiency and accuracy in clinical practice while potentially revealing new insights into disease morphology and progression.

Appendix A

My Appendix



(a) WSI 1 (b) WSI 2 (c) WSI 3

Figure A.1: Visualization of heatmaps across three WSIs using the BioMedCLIP model.

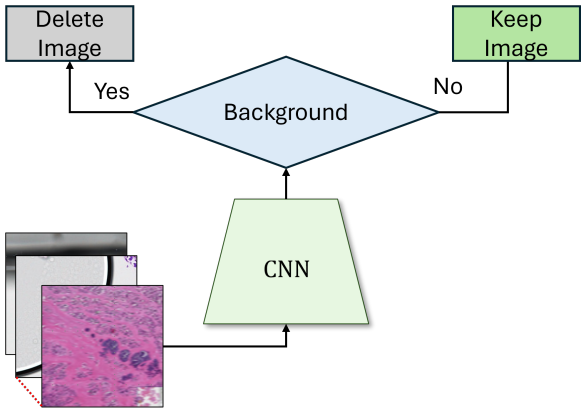


Figure A.2: Cleaning pipeline for patches extracted from WSI using custom CNN

References

Abbet, C., Zlobec, I., Bozorgtabar, B., & Thiran, J.-P. (2020). *Divide-and-rule: Self-supervised learning for survival analysis in colorectal cancer.* Retrieved from https://arxiv.org/abs/2007.03292

Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C. L., Batra, D., & Parikh, D. (2016). *Vqa: Visual question answering.* Retrieved from https://arxiv.org/abs/1505.00468

Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jindi, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical bert embeddings. *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 72–78.

Azizi, S., Culp, L., Freyberg, J., Mustafa, B., Baur, S., Kornblith, S., ... Natarajan, V. (2022). *Robust and efficient medical imaging with self-supervision.* Retrieved from https://arxiv.org/abs/2205.09723

Bardes, A., Ponce, J., & LeCun, Y. (2022). *Vicreg: Variance-invariance-covariance regularization for self-supervised learning.* Retrieved from https://arxiv.org/abs/2105.04906

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... Liang, P. (2022). *On the opportunities and risks of foundation models.* Retrieved from https://arxiv.org/abs/2108.07258

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). *Language models are few-shot learners.* Retrieved from https://arxiv.org/abs/2005.14165

Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam,

K. J., ... Fuchs, T. J. (2019, August). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, *25*(8), 1301–1309.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). *Emerging properties in self-supervised vision transformers.* Retrieved from https://arxiv.org/abs/2104.14294

Chanda, D., Aryal, M., Soltani, N. Y., & Ganji, M. (2024). *A new era in computational pathology: A survey on foundation and vision-language models.* Retrieved from https://arxiv.org/abs/2408.14496

Chen, L., Bentley, P., Mori, K., Misawa, K., Fujiwara, M., & Rueckert, D. (2022). Self-supervised learning for medical image analysis using image context restoration. *Medical image analysis*, *75*, 102304.

Chen, R. J., Chen, C., Li, Y., Chen, T. Y., Trister, A. D., Krishnan, R. G., & Mahmood, F. (2022). *Scaling vision transformers to gigapixel images via hierarchical self-supervised learning.* Retrieved from https://arxiv.org/abs/2206.02647

Chen, R. J., Ding, T., Lu, M. Y., Williamson, D. F. K., Jaume, G., Song, A. H., ... Mahmood, F. (2024, March). Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, *30*(3), 850–862.

Chen, R. J., & Krishnan, R. G. (2022). *Self-supervised vision transformers learn visual concepts in histopathology.* Retrieved from https://arxiv.org/abs/2203.00585

Chen, R. J., Lu, M. Y., Williamson, D. F. K., Chen, T. Y., Lipkova, J., Shaban, M., ... Mahmood, F. (2021). *Pan-cancer integrative histology-genomic analysis via interpretable multimodal deep learning.* Retrieved from https://arxiv.org/abs/2108.02278

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *International conference on machine learning*, 1597–1607.

Child, R., Gray, S., Radford, A., & Sutskever, I. (2019). *Generating long sequences with sparse transformers.* Retrieved from https://arxiv.org/abs/1904.10509

Ciga, O., Xu, T., & Martel, A. L. (2021). *Self supervised contrastive learning for digital histopathology.* Retrieved from https://arxiv.org/abs/2011.13971

Coudray, N., Ocampo, P. S., Sakellaropoulos, T., Narula, N., Snuderl, M., Fenyö, D., ... Tsirigos, A. (2018, September). Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat Med*, *24*(10), 1559–1567.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 ieee conference on computer vision and pattern recognition* (p. 248-255). doi: 10.1109/CVPR.2009.5206848

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *Bert: Pre-training of deep bidirectional transformers for language understanding.* Retrieved from https://arxiv.org/abs/1810.04805

Dietterich, T. G., & Bakiri, G. (1995). *Solving multiclass learning problems via error-correcting output codes.* Retrieved from https://arxiv.org/abs/cs/9501101

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... Houlsby, N. (2021). *An image is worth 16x16 words: Transformers for image recognition at scale.* Retrieved from https://arxiv.org/abs/2010.11929

Ehteshami Bejnordi, B., Veta, M., Johannes van Diest, P., van Ginneken, B., Karssemeijer, N., Litjens, G., ... Venâncio, R. (2017, December). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, *318*(22), 2199–2210.

Filiot, A., Huszar, F., Dăugavietis, G., Huang, J., Zhou, Z., Gray, M., & Graham, S. (2023). Foundation models for computational pathology: Do transformer models with self-supervised learning actually live up to the hype? *Medical Image Analysis*, *87*, 102805.

Fuchs, T. J., & Buhmann, J. M. (2011, October). Computational pathology: Challenges and promises for tissue analysis. *Computerized Medical Imaging and Graphics*, *35*(7–8), 515–530. Retrieved from http://dx.doi.org/10.1016/j.compmedimag.2011.02.006 doi: 10.1016/j.compmedimag.2011.02.006

Gadiya, S., Anand, D., & Sethi, A. (2019). *Histographs: Graphs in histopathology.* Retrieved from https://arxiv.org/abs/1908.05020

Gou, J., Yu, B., Maybank, S. J., & Tao, D. (2021, March). Knowledge distillation: A survey. *International Journal of Computer Vision*, *129*(6), 1789–1819. Retrieved from http://

dx.doi.org/10.1007/s11263-021-01453-z doi: 10.1007/s11263-021-01453-z

Gu, A., & Dao, T. (2024). *Mamba: Linear-time sequence modeling with selective state spaces.* Retrieved from https://arxiv.org/abs/2312.00752

Gu, A., Goel, K., & Ré, C. (2022). *Efficiently modeling long sequences with structured state spaces.* Retrieved from https://arxiv.org/abs/2111.00396

Gu, J., Han, Z., Chen, S., Beirami, A., He, B., Zhang, G., ... Torr, P. (2023). *A systematic survey of prompt engineering on vision-language foundation models.* Retrieved from https://arxiv.org/abs/2307.12980

Gurcan, M. N., Boucheron, L. E., Can, A., Madabhushi, A., Rajpoot, N. M., & Yener, B. (2009). Histopathological image analysis: A review. *IEEE Reviews in Biomedical Engineering*, *2*, 147-171. doi: 10.1109/RBME.2009.2034865

He, B., Bergenstråhle, L., Stenbeck, L., Abid, A., Andersson, A., Borg, Å., ... Zou, J. (2020, August). Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature Biomedical Engineering*, *4*(8), 827–834.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2021). *Masked autoencoders are scalable vision learners.* Retrieved from https://arxiv.org/abs/2111.06377

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). *Momentum contrast for unsupervised visual representation learning.* Retrieved from https://arxiv.org/abs/1911.05722

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep residual learning for image recognition.* Retrieved from https://arxiv.org/abs/1512.03385

Hinton, G., Vinyals, O., & Dean, J. (2015). *Distilling the knowledge in a neural network.* Retrieved from https://arxiv.org/abs/1503.02531

Hossain, M. S., Karuniawati, H., Jairoun, A. A., Urbi, Z., Ooi, D. J., John, A., ... Hadi, M. A. (2022, March). Colorectal cancer: A review of carcinogenesis, global epidemiology, current challenges, risk factors, preventive and treatment strategies. *Cancers (Basel)*, *14*(7).

Hosseini, M. S., Bejnordi, B. E., Trinh, V. Q.-H., Hasan, D., Li, X., Kim, T., ... Plataniotis, K. N. (2024). *Computational pathology: A survey review and the way forward.* Retrieved from https://arxiv.org/abs/2304.05482

Hou, L., Agarwal, A., Samaras, D., Kurc, T. M., Gupta, R. R., & Saltz, J. H. (2019). Robust

histopathology image analysis: To label or to synthesize? In *2019 ieee/cvf conference on computer vision and pattern recognition (cvpr)* (p. 8525-8534). doi: 10.1109/CVPR.2019.00873

Ignatov, A., & Malivenko, G. (2024). *Nct-crc-he: Not all histopathological datasets are equally useful.* Retrieved from https://arxiv.org/abs/2409.11546

Ikezogwo, W. O., Seyfioglu, M. S., Ghezloo, F., Geva, D. S. C., Mohammed, F. S., Anand, P. K., ... Shapiro, L. (2025). *Quilt-1m: One million image-text pairs for histopathology.* Retrieved from https://arxiv.org/abs/2306.11207

Ilse, M., Tomczak, J. M., & Welling, M. (2018). *Attention-based deep multiple instance learning.* Retrieved from https://arxiv.org/abs/1802.04712

Jaume, G., Pati, P., Bozorgtabar, B., Foncubierta-Rodríguez, A., Feroce, F., Anniciello, A. M., ... Goksel, O. (2021). *Quantifying explainers of graph neural networks in computational pathology.* Retrieved from https://arxiv.org/abs/2011.12646

Kang, M., Song, H., Park, S., Yoo, D., & Pereira, S. (2023). *Benchmarking self-supervised learning on diverse pathology datasets.* Retrieved from https://arxiv.org/abs/2212.04690

Katharopoulos, A., Vyas, A., Pappas, N., & Fleuret, F. (2020). *Transformers are rnns: Fast autoregressive transformers with linear attention.* Retrieved from https://arxiv.org/abs/2006.16236

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., ... Krishnan, D. (2021). *Supervised contrastive learning.* Retrieved from https://arxiv.org/abs/2004.11362

Kirilenko, D., Andreychuk, A., Panov, A., & Yakovlev, K. (2022). *Transpath: Learning heuristics for grid-based pathfinding via transformers.* Retrieved from https://arxiv.org/abs/2212.11730

Kitaev, N., Łukasz Kaiser, & Levskaya, A. (2020). *Reformer: The efficient transformer.* Retrieved from https://arxiv.org/abs/2001.04451

Koohbanani, N. A., Unnikrishnan, B., Khurram, S. A., Krishnaswamy, P., & Khalifa, N. (2021). Self-path: Self-supervision for classification of pathology images with limited annotations. *IEEE Transactions on Medical Imaging*, *40*(10), 2845–2856.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, & K. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 25). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234–1240.

Li, B., Li, Y., & Eliceiri, K. W. (2021). *Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning.* Retrieved from https://arxiv.org/abs/2011.08939

Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., ... Gao, J. (2023). *Llava-med: Training a large language-and-vision assistant for biomedicine in one day.* Retrieved from https://arxiv.org/abs/2306.00890

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., ... Sánchez, C. I. (2017, December). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*, 60–88. Retrieved from http://dx.doi.org/10.1016/j.media.2017.07.005 doi: 10.1016/j.media.2017.07.005

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). *Visual instruction tuning.* Retrieved from https://arxiv.org/abs/2304.08485

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Lu, M. Y., Chen, B., Williamson, D. F. K., Chen, R. J., Liang, I., Ding, T., ... Mahmood, F. (2023). *Towards a visual-language foundation model for computational pathology.* Retrieved from https://arxiv.org/abs/2307.12914

Lu, M. Y., Chen, B., Williamson, D. F. K., Chen, R. J., Zhao, M., Chow, A. K., ... Mahmood, F. (2024, October). A multimodal generative AI copilot for human pathology. *Nature*, *634*(8033), 466–473.

Lu, M. Y., Chen, B., Zhang, A., Williamson, D. F. K., Chen, R. J., Ding, T., ... Mahmood, F. (2023). *Visual language pretrained multiple instance zero-shot transfer for histopathology images.* Retrieved from https://arxiv.org/abs/2306.07831

Lu, M. Y., Kong, D., Lipkova, J., Chen, R. J., Singh, R., Williamson, D. F. K., ... Mahmood, F. (2020). *Federated learning for computational pathology on gigapixel whole slide images.* Retrieved from https://arxiv.org/abs/2009.10190

Lu, M. Y., Williamson, D. F. K., Chen, T. Y., Chen, R. J., Barbieri, M., & Mahmood, F. (2020). *Data efficient and weakly supervised computational pathology on whole slide images.* Retrieved from https://arxiv.org/abs/2004.09666

M., G., V., A., S., M.-M., & H., M. (2020). Concept attribution: Explaining cnn decisions to physicians. *Computers in Biology and Medicine*, *123*, 103865. Retrieved from https://www.sciencedirect.com/science/article/pii/S0010482520302225 doi: https://doi.org/10.1016/j.compbiomed.2020.103865

Nguyen, E., Goel, K., Gu, A., Downs, G. W., Shah, P., Dao, T., ... Ré, C. (2022). *S4nd: Modeling images and videos as multidimensional signals using state spaces.* Retrieved from https://arxiv.org/abs/2210.06583

Park, W., Kim, D., Lu, Y., & Cho, M. (2019). *Relational knowledge distillation.* Retrieved from https://arxiv.org/abs/1904.05068

Qu, L., Yang, D., Huang, D., Guo, Q., Luo, R., Zhang, S., & Wang, X. (2024). *Pathology-knowledge enhanced multi-instance prompt learning for few-shot whole slide image classification.* Retrieved from https://arxiv.org/abs/2407.10814

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... others (2021a). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... Sutskever, I. (2021b). *Learning transferable visual models from natural language supervision.* Retrieved from https://arxiv.org/abs/2103.00020

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training.* Retrieved from https://openai.com/blog/

language-unsupervised/ (OpenAI Blog)

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... Liu, P. J. (2023). *Exploring the limits of transfer learning with a unified text-to-text transformer.* Retrieved from https://arxiv.org/abs/1910.10683

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., ... Sutskever, I. (2021). *Zero-shot text-to-image generation.* Retrieved from https://arxiv.org/abs/2102.12092

Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., & Bengio, Y. (2015). *Fitnets: Hints for thin deep nets.* Retrieved from https://arxiv.org/abs/1412.6550

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2025). *A systematic survey of prompt engineering in large language models: Techniques and applications.* Retrieved from https://arxiv.org/abs/2402.07927

Sellergren, A., Wang, X., Lee, N., Zhang, Z., Huang, H., Seong, M., ... Lungren, M. (2023). Foundation models for general medical image classification. *Nature Medicine*, *29*(9), 2289–2300.

Seyfioglu, M. S., Ikezogwo, W. O., Ghezloo, F., Krishna, R., & Shapiro, L. (2025). *Quilt-llava: Visual instruction tuning by extracting localized narratives from open-source histopathology videos.* Retrieved from https://arxiv.org/abs/2312.04746

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... Natarajan, V. (2023, August). Large language models encode clinical knowledge. *Nature*, *620*(7972), 172–180.

Smith, J. T. H., Warrington, A., & Linderman, S. W. (2023). *Simplified state space layers for sequence modeling.* Retrieved from https://arxiv.org/abs/2208.04933

Song, A. H., Jaume, G., Williamson, D. F. K., Lu, M. Y., Vaidya, A., Miller, T. R., & Mahmood, F. (2023, October). Artificial intelligence for digital and computational pathology. *Nature Reviews Bioengineering*, *1*(12), 930–949. Retrieved from http://dx.doi.org/10.1038/s44222-023-00096-8 doi: 10.1038/s44222-023-00096-8

Sun, Y., Zhu, C., Zheng, S., Zhang, K., Sun, L., Shui, Z., ... Yang, L. (2024). *Pathasst: A generative foundation ai assistant towards artificial general intelligence of pathology.* Retrieved from https://arxiv.org/abs/2305.15072

Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J.-M., Ciompi, F., & van der Laak, J. (2019, December). Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical Image Analysis*, *58*, 101544. Retrieved from http://dx.doi.org/10.1016/j.media.2019.101544 doi: 10.1016/j.media.2019.101544

Tellez, D., Litjens, G., van der Laak, J., & Ciompi, F. (2021, February). Neural image compression for gigapixel histopathology image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *43*(2), 567–578. Retrieved from http://dx.doi.org/10.1109/TPAMI.2019.2936841 doi: 10.1109/tpami.2019.2936841

Tizhoosh, H. R., & Pantanowitz, L. (2018, November). Artificial intelligence and digital pathology: Challenges and opportunities. *J Pathol Inform*, *9*, 38.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., ... others (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... Scialom, T. (2023). *Llama 2: Open foundation and fine-tuned chat models.* Retrieved from https://arxiv.org/abs/2307.09288

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2023). *Attention is all you need.* Retrieved from https://arxiv.org/abs/1706.03762

Veeling, B. S., Linmans, J., Winkens, J., Cohen, T., & Welling, M. (2018, June). Rotation equivariant CNNs for digital pathology.

Veta, M., Heng, Y. J., Stathonikos, N., Bejnordi, B. E., Beca, F., Wollmann, T., ... Pluim, J. P. (2019). Predicting breast tumor proliferation from whole-slide images: The tupac16 challenge. *Medical Image Analysis*, *54*, 111-121. Retrieved from https://www.sciencedirect.com/science/article/pii/S1361841518305231 doi: https://doi.org/10.1016/j.media.2019.02.012

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). *Show and tell: A neural image caption generator.* Retrieved from https://arxiv.org/abs/1411.4555

Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Liu, S., ... Fuchs, T. J. (2024). *Virchow: A million-slide digital pathology foundation model.* Retrieved from https://arxiv.org/abs/2309.07778

Vorontsov, E., Bozkurt, A., Casson, A., Shaikovski, G., Zelechowski, M., Severson, K., ... Fuchs, T. J. (2024, October). A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature Medicine*, *30*(10), 2924–2935.

Wang, B., Liu, J., Karimnazarov, J., & Thompson, N. (2024, March). Task supportive and personalized human-large language model interaction: A user study. In *Proceedings of the 2024 acm sigir conference on human information interaction and retrieval* (p. 370–375). ACM. Retrieved from http://dx.doi.org/10.1145/3627508.3638344 doi: 10.1145/3627508.3638344

Wang, J., Shi, E., Yu, S., Wu, Z., Ma, C., Dai, H., ... Zhang, S. (2024). *Prompt engineering for healthcare: Methodologies and applications.* Retrieved from https://arxiv.org/abs/2304.14670

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., ... Shao, L. (2021). Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 568–578.

Wang, X., Yan, Y., Tang, P., Bai, X., & Liu, W. (2018, February). Revisiting multiple instance neural networks. *Pattern Recognition*, *74*, 15–24. Retrieved from http://dx.doi.org/10.1016/j.patcog.2017.08.026 doi: 10.1016/j.patcog.2017.08.026

Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., ... Han, X. (2022). Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis*, *81*, 102559. Retrieved from https://www.sciencedirect.com/science/article/pii/S1361841522002043 doi: https://doi.org/10.1016/j.media.2022.102559

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... others (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., ... Zhou, D. (2023). *Chain-of-thought prompting elicits reasoning in large language models.* Retrieved from https://arxiv.org/abs/2201.11903

Wu, C., Yin, S., Qi, W., Wang, X., Tang, Z., & Duan, N. (2023). *Visual chatgpt: Talking, drawing and editing with visual foundation models.* Retrieved from https://arxiv.org/abs/2303.04671

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., ... Hu, H. (2022). *Simmim: A simple framework for masked image modeling.* Retrieved from https://arxiv.org/abs/2111.09886

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., & Wu, Y. (2022). *Coca: Contrastive captioners are image-text foundation models.* Retrieved from https://arxiv.org/abs/2205.01917

Yuan, L., Chen, D., Chen, Y.-L., Codella, N., Dai, X., Gao, J., ... Zhang, P. (2021). *Florence: A new foundation model for computer vision.* Retrieved from https://arxiv.org/abs/2111.11432

Zbontar, J., Jing, L., Misra, I., LeCun, Y., & Deny, S. (2021). *Barlow twins: Self-supervised learning via redundancy reduction.* Retrieved from https://arxiv.org/abs/2103.03230

Zhan, C., Zhang, Y., Lin, Y., Wang, G., & Wang, H. (2023). *Unidcp: Unifying multiple medical vision-language tasks via dynamic cross-modal learnable prompts.* Retrieved from https://arxiv.org/abs/2312.11171

Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., ... Poon, H. (2025). *Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs.* Retrieved from https://arxiv.org/abs/2303.00915

Zhou, Y., Graham, S., Koohbanani, N. A., Shaban, M., Heng, P.-A., & Rajpoot, N. (2019). *Cgc-net: Cell graph convolutional network for grading of colorectal cancer histology images.* Retrieved from https://arxiv.org/abs/1909.01068