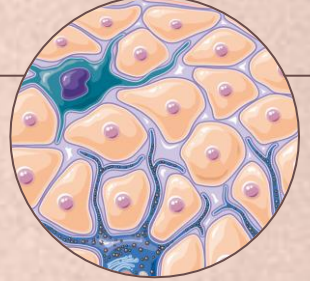
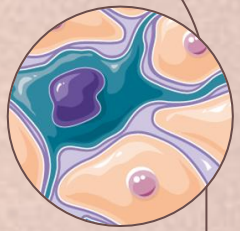




TERM PAPER PRESENTATAION

Many-to-Many-Alignment-Learning-for-Histopathology-Image-Classification-using-Vision-Language-Models

SPECIAL TOPICS IN DEEP LEARNING

ICS 590
Dr. Muzammil Behzad

PRESENTED BY

Quratulain Arshad
g202523250@kfupm.edu.sa

Terminologies

Pathology	Branch of medical science that is focused on the study and diagnosis of disease.
Histopathology	Diagnosis and study of diseases of the tissues, and involves examining tissues and/or cells under a microscope.
Computational Pathology	A brand-new discipline that aims to enhance patient care by utilizing advances in artificial intelligence and data generated from anatomic and clinical pathology.
Vision Language Model	Fusion of vision and natural language models. It ingests images and their respective textual descriptions as inputs and learns to associate the knowledge from the two modalities.
Whole Slide Images	Also called “virtual” microscopy, involves digitally scanning a tissue slide containing thin sections of tissue specimens for microscopic examination and storing it as digital images. This process allows for remote collaboration.
Weakly Supervised Learning	Weak supervision covers a wide range of methods where models are trained using partial, inexact, or otherwise inaccurate information that is simpler to supply than hand-labeled data
Self-Supervised Learning	A paradigm in machine learning where a model is trained on a task using the data itself to generate supervisory signals, rather than relying on externally-provided labels.
Contrastive Learning	A self-supervised learning technique where the model learns by pulling similar samples closer and pushing dissimilar samples apart in the embedding space
Zero Shot Learning	A technique where a model predicts classes it has never been trained on, using semantic descriptions (text prompts or attributes)

Table of contents

01 Challenges in Histopathology

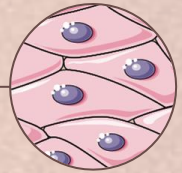
02 Limitations of Learning Models

03 Problem Statement

04 Proposed Methodology

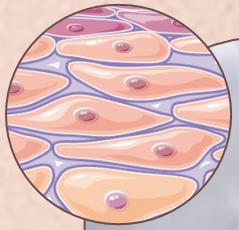
05 Results

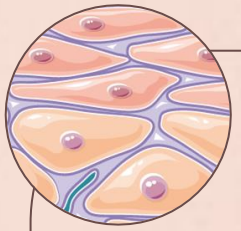
06 Conclusion



01

Challenges in Histopathology





Challenges in Histopathology

Problems with CNNs & ViTs

Deep learning models CNNs and vision transformers (ViTs) have demonstrated strong performance in tasks such as tumor detection, subtype classification, and cellular composition analysis.

But these approaches remain **heavily dependent on large, expertly annotated datasets**, which are **costly, time-consuming to curate, and often institution-specific**.

Histopathology-specific Challenges

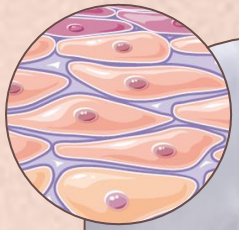
Histopathology introduces substantial domain-specific challenges:

- Different hospitals use different **staining protocols**
- Different scanners produce different **resolutions and color profiles**
- Tissue preparation varies between labs
- Datasets differ in **morphology, artifacts, and disease distribution**
- inter-observer differences i.e. **two or more clinicians/pathologists do not agree** on the diagnosis, label, or annotation of the same histopathology image

Leads to domain shift, causing trained models to perform poorly when applied across different clinical centers

02

Limitations of Learning Models



Categories of Models for Histology Image Calssification & Segmentation

Weakly Supervised

Use data with labels at a broad level, without needing detailed annotations for every instance

Self Supervised

Learn from the data itself without using labels, using pretext tasks to boost downstream task performance

Vision Language Supervised

Integrate textual descriptions with visual data to pre-train deep models.

Vision Language Supervised Models

Adhering to the conventional VL model training approach, these methodologies leverage paired visual-textual data within a contrastive learning framework to ensure that representations of similar Visual-Textual concepts are drawn **closer together**, while divergent ones are **distanced**

What they are

Models that are **trained using paired image and text data**.

During training, each image is linked with one or more text descriptions.

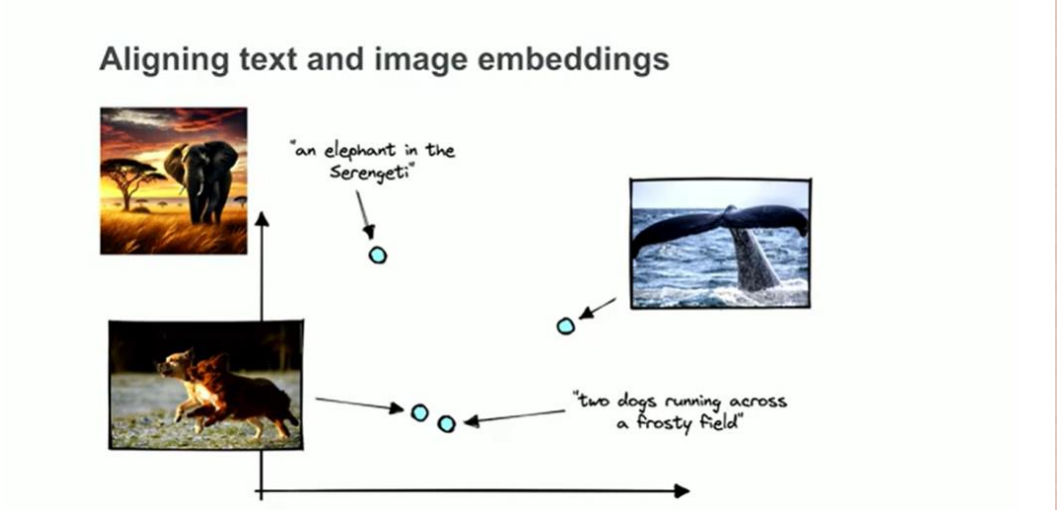
The supervision signal is the similarity between the image and the corresponding text caption.

How they work

They learn to align **visual features** with **language features**.

They require **supervision in the form of image–text pairs**.

Examples: **CLIP, PLIP, BiomedCLIP, CONCH**



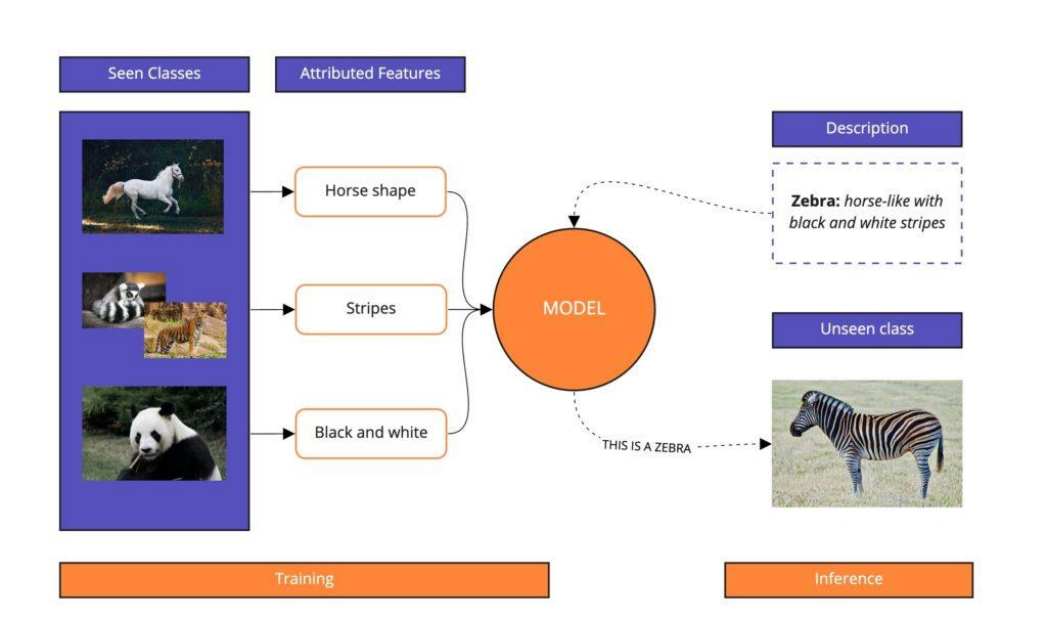
Limitations of Learning Models

Paradigm	Traditional Histopathology	Supervised deep learning in Histopathology	Weakly-supervised learning in Histopathology	Self-supervised learning in Histopathology	Vision-language models for Computational Pathology
Key Strengths	Clinical gold standard Leverages broad human expertise	High accuracy with large curated Labels Models can learn spatial/graph context	Uses slide-level labels; attention maps offer interpretability	Learns from vast unlabelled corpora Strong downstream features	Open-vocabulary recognition; versatile zero/few-shot transfer
Key Limitations	Inter-observer variability Labour-intensive Overlooks tissue heterogeneity	Requires extensive annotations Performance drops under domain shift	Still label-dependent May miss subtle local cues Bag design sensitive	Computationally demanding Heavy augmentation No inherent language grounding	Require millions of image text pairs and high computational cost Often single-prompt, Single-scale alignment

Zero Shot Learning

What is it

A learning setup where the model is asked to classify or recognize classes **that it never saw during training**.



How it Works

At inference time you provide text prompts for new, unseen classes.

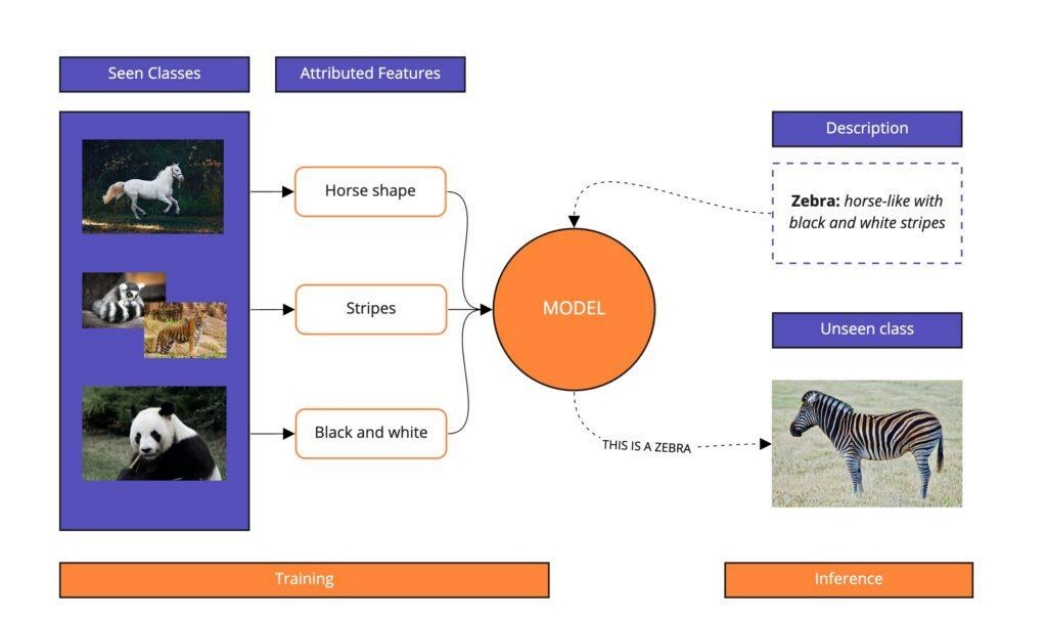
The model computes similarity between the image embedding and the text embedding.

It predicts the class with the highest similarity score

Zero Shot Learning

What is it

A learning setup where the model is asked to classify or recognize classes **that it never saw during training**.



How it Works

At inference time you provide text prompts for new, unseen classes.

The model computes similarity between the image embedding and the text embedding.

It predicts the class with the highest similarity score

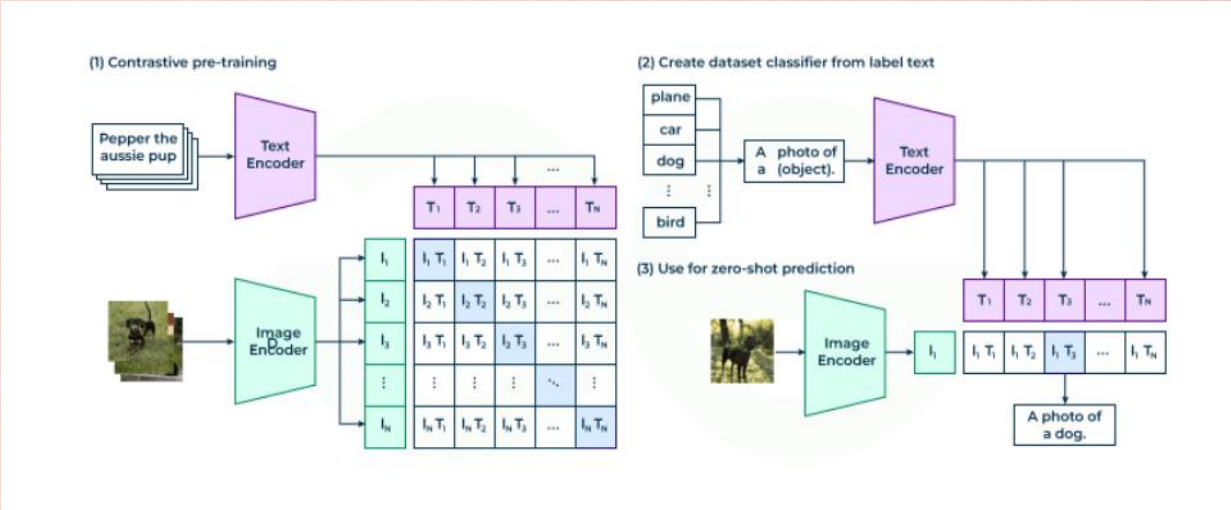
Contrastive Language Image Pretraining (CLIP)

An advanced AI multimodal model developed by OpenAI and UC Berkeley that combines knowledge of English Language concepts with semantic knowledge of images.

It has the unique ability to understand and relate both textual descriptions and images

It is first trained on many image–text pairs to compare correct vs incorrect matches, so that it learns a shared understanding of language and images.

CLIP has:
an image encoder (sees pictures)
a text encoder (reads sentences)



CLIP Versions

Model	Year	Backbone	Training Data	Goal / Improvement	Key Strengths
CLIP (Original)	2021	ResNet / ViT	400M image–text pairs	Foundational contrastive vision–language model	Excellent zero-shot, broad generalization
OpenCLIP	2022	ViT (various sizes)	LAION-400M/LAION-2B	Fully open-source re-implementation	Scalable, wide model zoo, reproducible
BioCLIP	2023	ViT-B/16	PMC papers + medical images	Adapts CLIP to biomedical domain	Better medical text understanding
MedCLIP	2022	ViT-B	MIMIC-CXR + clinical notes	Medical VL alignment for radiology	Strong on CXR zero-shot tasks
PLIP	2023	ViT-L	1M histopathology image–text pairs	Pathology-specific CLIP	Best zero-shot WSI patch classification
BiomedCLIP	2023	ViT-L/14	15M biomedical captions	Large-scale biomedical VL pretraining	Strong in cross-modality medical tasks
CONCH	2024	ViT-B	Multi-magnification histopathology	Multi-scale VL alignment	Much better for multi-resolution pathology tasks
CPLIP	2024	ViT-L	5M histopathology prompt bags	**Many-to-many image–text alignment**	State-of-the-art WSI zero-shot classification

CPLIP: Zero-Shot Learning for Histopathology with Comprehensive Vision-Language Alignment

Computer Vision and Pattern Recognition Conference (CVPR) 2024

Comprehensive Pathology Language Image Pre-training (CPLIP)

Designed specifically for **histopathology images**, like whole slide images and patches from tissue biopsies

Uses a **pathology specific dictionary**, generates pathology style text descriptions with language models, then aligns those descriptions with pathology images.

Uses a **many to many contrastive loss** instead of simple one image one caption

Better **zero shot classification and segmentation** in histopathology, more interpretable for pathologists

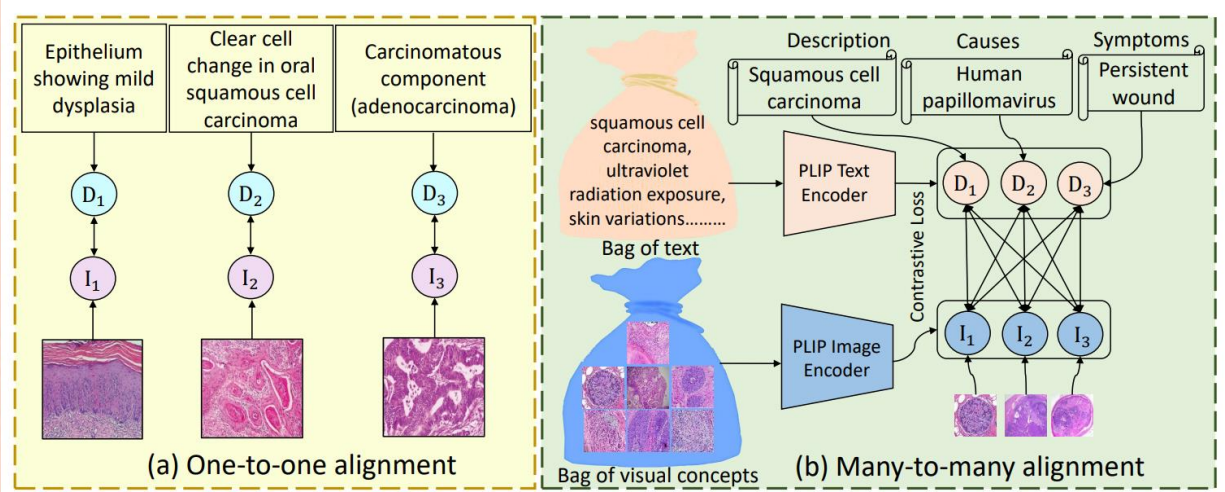
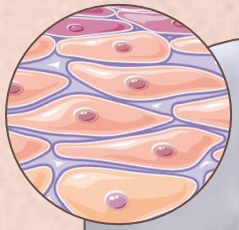


Figure 2. **(a)** Displays the traditional one-to-one alignment in computational pathology VL models like PLIP [14], BiomedCLIP [37], and MI-Zero [23], where each histology image is aligned with a single textual description during fine-tuning. **(b)** Our proposed approach of many-to-many alignment, where bags of correlated texts are aligned with bags of correlated histology images during fine-tuning, offers a richer, interconnected data set for model training.

03

Problem Statement



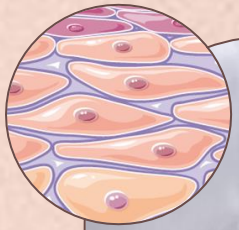
Problem Statement

- Despite the rapid progress of vision-language models in computational pathology, significant challenges remain before zero-shot histopathology becomes a reliable clinical tool. One major limitation concerns modality alignment, as most current methods rely on one-to-one alignment between an image patch and a single text prompt. This does not adequately reflect the inherently multi-scale and multi-semantic nature of pathology images, which can encode morphology, disease etiology, grading indicators, and contextual tissue organization.

- A second challenge is dataset heterogeneity and domain shift. Whole slide images vary widely due to differences in staining protocols, scanners, magnification, and institution-specific workflows. As a result, vision-language models trained on one dataset frequently underperform on external data, especially for unseen disease categories or rare cancer presentations

03

Proposed Methodology



Proposed Methodology

Contributions

- Propose a novel architecture that integrates a pathology-specific vision–language encoder (PLIP) with a general-purpose CLIP model to capture both domain-dependent and generic visual–textual semantics for histopathology image classification.

- Introduce a new alignment strategy that establishes interactions between multiple visual and textual embeddings, enabling richer semantic correspondence between histological patterns and diagnostic textual concepts. This alignment forms a multi-level similarity matrix that enhances discriminative learning under zero-shot conditions.

- Design clinically relevant textual prompts that reflect benign and malignant morphological cues, bridging the linguistic gap between medical reporting terminology and visual pathology features.

- Extensive experiments on the **BreakHis dataset** demonstrate the effectiveness of the proposed approach in classifying breast histopathology images without task-specific fine-tuning, highlighting its scalability and annotation efficiency for digital pathology applications.

Proposed Methodology

HistoAlign: A Dual-Encoder Many-to-Many Alignment Framework for Zero-ShotHistopathology Classification

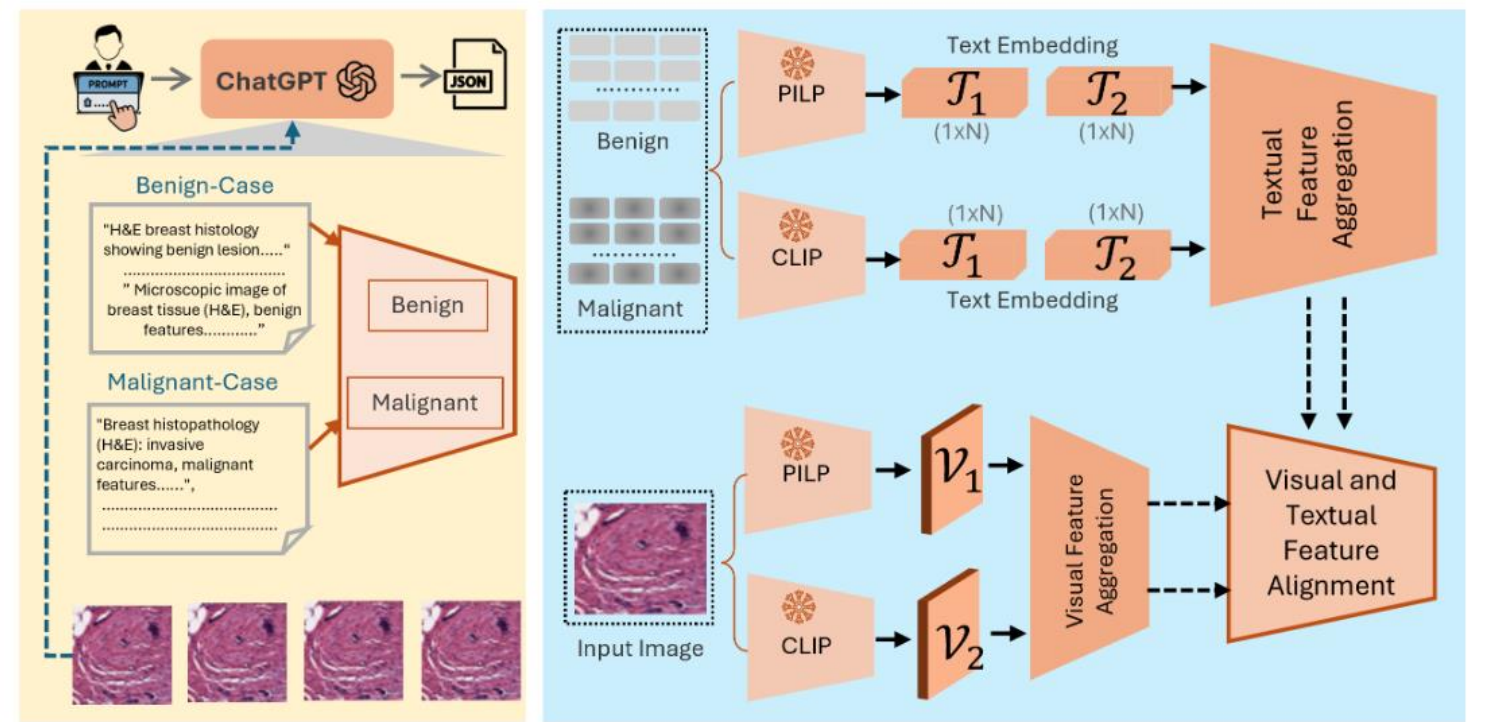


Fig. 1. Overview of the proposed training-free dual vision–language framework for zero-/few-shot classification.

Results

Experimnetal Setup

Dataset Description.

All experiments are conducted on the **BreakHis dataset**, a publicly available benchmark for breast histopathology image classification.

BreakHis contains 7,909 microscopic images of benign and malignant breast tumors collected from 82 patients, acquired under four optical magnifications (40×, 100×, 200×, and 400×).
Each image is an H&E-stained RGB patch of size 700×460 pixels.

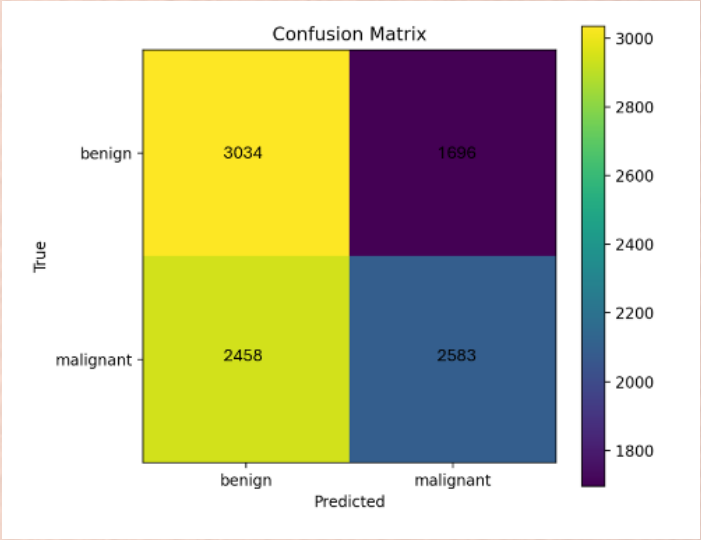
For this study, we curate a custom 10-folder subset comprising an equal number of benign and malignant cases (five folder search) to create a balanced binary evaluation set. This subset preserves magnification diversity and reflects real diagnostic variance in tissue morphology.

Results

TABLE II

ZERO-SHOT CLASSIFICATION RESULTS OF **HISTOALIGN** ON THE BREAKHIS BINARY SUBSET. METRICS ARE COMPUTED PER CLASS AND AVERAGED ACROSS MAGNIFICATIONS.

Class	Precision	Recall	F1-Score	Accuracy	AUC
Benign	0.5075	0.6416	0.5668	54.0	0.5599
Malignant	0.5529	0.4158	0.4746	51.0	0.5599
Macro Avg.	0.5302	0.5287	0.5207	52.51	0.5599
Weighted Avg.	0.5309	0.5251	0.5192	52.51	0.5599



Results

TABLE III

ABLATION STUDY ON DIFFERENT ENCODER COMBINATIONS AND ALIGNMENT STRATEGIES FOR ZERO-SHOT CLASSIFICATION ON THE BREAKHIS BINARY SUBSET. AT, ACC, PRE, REC, F1-S REPRESENTS ALIGNMENT TYPE, ACCURACY, PRECISION, RECALL, F1-SCORE, WHILE M AND F REPRESENTS MANY AND FUSION. METRICS ARE AVERAGED ACROSS MAGNIFICATIONS.

Configuration	AT	Acc	Pre	Rec	F1-S
PLIP (V_1–T_1) Only	One-to-One	49.3	0.491	0.503	0.497
CLIP (V_2–T_2) Only	One-to-One	50.7	0.510	0.495	0.502
Cross Pair (V_1–T_2)	One-to-One	51.6	0.519	0.503	0.511
Cross Pair (V_2–T_1)	One-to-One	51.9	0.522	0.508	0.515
Dual Encoders (T_2-to-T_2)	Joint Fusion	52.1	0.526	0.513	0.520
HistoAlign (Proposed)	**M-to-M-F**	**52.51**	**0.530**	**0.529**	**0.521**

Conclusion

In this work, **HistoAlign**, is presented which is a dual-encoder zero-shot learning framework designed for breast histopathologyvimage classification.

By leveraging a many-to-many cross-modal alignment between pathology-specific and generalvvision–language encoders, the proposed model effectively bridges visual and textual representations without requiringany supervised fine-tuning.

Comprehensive experiments on the **BreakHis dataset**, including a curated 10-folder binary subset, demonstrated that **HistoAlign** can transfer semantic knowledge from large-scale natural image–text pairs to highly domain-specific medical imagery.

Limitations

Despite its promise, several limitations persist.

The reliance on handcrafted clinical prompts may constrain linguistic coverage and hinder generalization to unseen tissue morphologies.

Furthermore, the visual encoders, pretrained primarily on natural images, are suboptimal for microscopic texture representation, leading to occasional misclassification of low-grade malignancies

Future Directions

Future work will focus on addressing these challenges through adaptive prompt optimization, stain-invariant visual pretraining, and the integration of few-shot or weakly supervised fine-tuning mechanisms to reduce domain bias.

Extending the framework to multi-class classification and cross-dataset generalization tasks will also be explored to enhance the clinical applicability of vision–language models in digital pathology.

Thanks!