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

Quratulain Arshad
Student IDs: g202523250
King Fahd University of Petroleum and Minerals
Dhahran, Saudi Arabia

Supervised by: Dr. Muzammil Behzad
muzammil.behzad@kfupm.edu.sa
King Fahd University of Petroleum and Minerals
Dhahran, Saudi Arabia

Abstract—**Accurate classification of breast histopathology images is a critical step in computer-aided diagnosis, yet most existing approaches rely on large annotated datasets and task-specific fine-tuning. In this paper, we introduce HistoAlign, a *dual-encoder zero-shot framework* designed for binary breast tissue classification on the BreakHis dataset. Unlike conventional supervised methods, HistoAlign performs inference without any retraining by exploiting a *many-to-many cross-modal alignment* mechanism that connects multiple visual and textual embedding spaces. The framework integrates a pathology-aware encoder (PLIP) and a generic vision-language encoder (CLIP) to capture complementary visual semantics across varying magnification levels. On the textual side, clinically grounded prompts describing benign and malignant morphologies are encoded through dual text towers and semantically fused to form robust diagnostic representations. This multi-level alignment yields a class similarity matrix that strengthens the correspondence between histological image patterns and textual diagnostic cues. Experimental results demonstrate that HistoAlign achieves competitive performance under a fully zero-shot setting, highlighting its potential as an annotation-efficient and scalable solution for digital pathology.**

Index Terms—**Vision Language Model, Prompt Learning, Zero-shot Learning, Features Alignment**

I. Introduction

The rapid evolution of artificial intelligence has transformed computational pathology, where whole slide images (WSIs) are analyzed to support cancer diagnosis, grading, and treatment planning. Deep learning models particularly convolutional neural networks (CNNs) and vision transformers (ViTs) [1] [2] have demonstrated strong performance in tasks such as tumor detection, subtype classification, and cellular composition analysis; however, these approaches remain heavily dependent on large, expertly annotated datasets, which are costly, time-consuming to curate, and often institution-specific [3] [4]. Beyond the annotation burden, histopathology introduces substantial domain-specific challenges: variability in staining protocols, scanner hardware, tissue preparation methods, and inter-observer differences leads to domain shift, causing trained models to perform poorly when applied across different clinical centers [5]. Moreover, rare or newly emerging cancer subtypes frequently lack sufficient labeled examples to support robust supervised learning. Together, these limitations underscore the need for models that minimize reliance on manual annotations while maintaining strong generalization across diverse clinical environments, motivating the development of next-generation computational pathology methods capable of scalable, label-efficient, and institution-agnostic performance.

Zero-shot learning using vision-language models (VLMs) has emerged as a promising direction for developing models that align medical images with descriptive text prompts, enabling semantic transfer without task-specific supervision [6] [7]. While CLIP-based frameworks have shown striking generalization in natural image domains, their direct application to pathology is limited because WSIs contain complex multi-scale structures and depend heavily on medical terminology absent from generic language corpora [8]. This has led to increasing interest in pathology-aware language alignment, multi-resolution representation learning, and domain-specific prompting strategies [9]. Collectively, these innovations point toward a new generation of foundation models that can recognize rare diseases, scale across institutions, and reduce reliance on annotated training datasets.

The CPLIP framework [10] represents a major milestone by introducing comprehensive alignment between text and image domains using heterogeneous prompt sets and diverse tile collections. Rather than relying on single captions or limited vocabularies, CPLIP leverages domain-specific glossaries, GPT-based text expansions, and pathology aware image retrieval to enhance semantic and morphological coverage. Through many-to-many contrastive training, it achieves significant zero-shot gains across multiple WSI datasets, outperforming PLIP, MI-Zero, and BiomedCLIP in both tile-level and slide-level classification tasks. However, even with these advances, model brittleness persists when faced with rare morphologies, limited image diversity, or fine-grained subtyping needs that require deeper semantic sensitivity. Recent literature [11] [12] further emphasizes the need for improved prompt design, hierarchical representation learning, and better alignment mechanisms to enhance robustness under real clinical variability.

A. 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. The CPLIP framework introduces a more expressive many-to-many alignment strategy between image bags and prompt bags, enabling richer semantic grounding, but further refinement is still needed to fully exploit pathological variations.

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. While CPLIP improves zero-shot transfer under domain shift, the field still lacks mechanisms that reliably generalize across multi-center data, support rare subtypes, and maintain stable performance in real clinical conditions. These challenges underscore the need for more robust alignment strategies, domain-aware modeling, and architectures that reduce reliance on large supervised datasets.

B. Objectives

This work aims to advance zero-shot learning for histopathological image analysis by integrating visual and textual modalities into a unified framework. The main objectives of this study are as follows:

- To design a dual-encoder vision–language architecture that combines a pathology-aware encoder (PLIP) with a general-purpose vision-language encoder (CLIP) to learn complementary visual representations across multiple magnifications.
- To construct clinically grounded textual prompts that encapsulate benign and malignant morphological characteristics, enabling precise semantic alignment between histopathological imagery and diagnostic language.
- To develop a many-to-many cross-modal alignment mechanism that captures hierarchical relationships between multiple visual and textual embeddings, facilitating robust zero-shot classification without task-specific fine-tuning.

These objectives collectively aim to establish a scalable, annotation-efficient framework capable of accurately distinguishing benign and malignant breast tissue under a fully zero-shot setting.

C. Scope of Study

The scope of this study is centered on the exploration of zero-shot learning for histopathological image classification, specifically targeting breast tissue analysis. The proposed framework, *HistoAlign*, focuses on the binary differentiation of *benign* and *malignant* samples using the **BreakHis** dataset. The study confines its experimental setup to hematoxylin and eosin (H&E)-stained microscopic images across multiple magnification factors (40×, 100×, 200×, and 400×). No additional supervised fine-tuning or domain adaptation is performed, ensuring the evaluation remains within a pure zero-shot paradigm. Furthermore, the scope emphasizes the design and validation of a dual-encoder many-to-many alignment mechanism that integrates pathology-specific and general vision–language models. The investigation does not extend to multi-class lesion categorization, stain normalization, or fine-tuning-based adaptation strategies, which are left as future research directions. This defined scope ensures that the findings specifically demonstrate the feasibility and robustness of cross-modal zero-shot alignment for diagnostic decision-making in digital pathology.

D. Main Contributions

The key contributions of this study are summarized as follows:

- We 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.
- We 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.
- We 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.

Together, these contributions establish a new direction for cross-modal zero-shot learning in medical image analysis by aligning visual and textual knowledge representations in a clinically meaningful manner.

II. Literature Review

Recent progress in computational pathology has led to a rich landscape of learning paradigms developed to analyze whole-slide images (WSIs) for tasks such as cancer classification, grading, and survival prediction. These approaches are broadly categorized into weakly supervised learning (WSL), self-supervised learning (SSL), and vision-language (VL) based modeling. Weakly Supervised Learning in Computational Pathology

Weakly supervised learning (WSL) has played a foundational role in computational pathology by enabling whole-slide image (WSI) classification using only slide-level labels rather than exhaustive pixel-level annotation. Multiple Instance Learning (MIL) has been the dominant paradigm [13], where a slide is treated as a bag of instances (patches) and the model learns to infer slide-level decisions from patch-level representations. Early and influential MIL-based methods include ABMIL [14], TransMIL [15], DSMIL [16], CLAM [4], and DTFD-MIL [17], which demonstrated that WSI-level diagnostic tasks can be learned effectively without detailed human annotation. More recent advances have begun to address limitations in spatial context and multi-scale modeling: MIL-VT introduces transformer-guided instance sampling which is a patch-slide discriminative joint learning framework that simultaneously models local patch features and global slide-level context, improving weakly supervised whole-slide image representation and classification accuracy [18]. In [19], a nested MIL hierarchy is introduced that organizes patches across multiple levels of granularity to better model tissue organization. To address tumor heterogeneity, a multi-marker feature integration strategy is proposed in [20] that enables the model to distinguish subtle intra-tumoral variations. With the growing demand for clinically relevant prediction tasks, the authors in [21] demonstrated the robustness of multi-modal and multi-instance deep learning across institutions, highlighting the importance of model generalizability. More recent graph-based advances, such as [22], leveraged graph neural networks to integrate multi-resolution features while transferring knowledge from teacher networks for performance refinement. Similarly, A Structure-Aware Hierarchical Graph-Based MIL Framework [23] proposed for pT Staging modeled glandular and stromal architecture through hierarchical graph aggregation, improving staging accuracy. Although powerful, WSL methods still require task-specific training and fail to generalize to unseen cancer types, a limitation that motivates label-efficient and zero-shot approaches.

Self-supervised learning (SSL) offers an attractive alternative by learning tissue representations directly from unlabeled WSIs, using pretext tasks such as contrastive learning, masked modeling, or hierarchical reconstruction. Earlier SSL methods such as CTransPath [24], HIPT [25] and H2T [26] demonstrated that pretraining in large collections of WSIs improves downstream classification and survival prediction tasks without requiring pathologist annotations. Stain-adaptive SSL methods have been proposed to mitigate stain variability across batches and scanners, enabling models to learn domain-invariant feature representations that generalize across heterogeneous laboratories [27]. Generic SSL frameworks such as Self-HER2Net extend this principle to IHC analysis by using large volumes of unlabeled breast cancer tissue to improve the performance of the HER2 classification [28]. Other work has compared task-specific SSL with transfer learning and modern foundation models, demonstrating that specialized pretraining strategies can outperform generic image encoders for downstream pathology tasks [29]. Beyond performance gains, SSL has also begun to reveal clinically significant morphological signatures, as shown in a colon cancer study, where SSL-derived features were linked to treatment-relevant histomorphological patterns and therapeutic response markers [30]. More recent work in self-supervised learning has produced a wide spectrum of techniques designed to reduce annotation burdens while capturing richer and more generalizable histopathological representations. In [31] an efficient multi-stage masked SSL framework is introduced that reconstructs tissue structures at multiple granularities, enabling the model to learn hierarchical morphological cues useful for downstream classification and retrieval tasks. Building on transformer architectures, [32] proposed a self-supervised segmentation approach in which transformer encoders learn coherence at the pixel-level and region-level without manual masks, achieving strong performance in segmenting complex glandular and stromal patterns. The authors in [33] developed a generalized ViT-based SSL model for prostate cancer diagnosis and grading, showing that unlabeled slide-level pretraining can significantly strengthen feature discrimination across tumor grades. Similarly, it is demonstrated in [34] that transformer-based SSL can enhance the classification of breast histopathology by learning subtle cellular and architectural differences directly from unlabeled tissue. Beyond RGB histology, [35] introduced S4R, a spectral regression–driven SSL method tailored for hyperspectral slides, enabling the extraction of biochemical signatures that complement spatial morphology. A diffusion-based stain augmentation and normalization framework called SAStainDiff is proposed in [36] which uses self-supervision to harmonize color variations across labs, thereby improving cross site model robustness. SSL methods significantly reduce labeling costs and improve feature robustness, yet they still lack semantic grounding, they learn how tissue looks, but not what it represents clinically. This limitation creates an opening for methods that incorporate medical language into feature learning.

Vision-language supervised learning brings semantic structure into computational pathology by aligning images with diagnostic text, enabling zero-shot tissue recognition and cross-dataset generalization. Early pathology-adapted VL models such as PLIP, MI-Zero [37], CONCH [38] and Biomed-CLIP [7] extend CLIP-style contrastive learning to paired histopathology–text corpora, allowing cancer subtyping without retraining. However, despite this progress, a key limitation of these works is their reliance on single-text and single-image alignment strategies which may fail to capture the complex multimodal relationships in pathology, limiting their ability to represent complex morphology and multi-scale tissue context. and the nuanced nature of diagnostic tasks. Multi-Resolution Prompt-guided Hybrid Embedding (MR-PHE) [39] is an effort to overcome this limitation that introduced a hybrid embedding strategy that integrates global image embeddings with weighted patch embeddings, effectively combining local and global contextual information. CPLIP [10] advances this line further by aligning bags of patches with bags of prompts via many-to-many cross-modal training, showing large gains

TABLE I
COMPARISON OF PARADIGMS FOR COMPUTATIONAL PATHOLOGY.

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 when large curated label sets are available; models can exploit spatial and graph context.	Uses slide-level labels instead of dense annotation; attention maps provide some interpretability.	Learns from vast unlabelled corpora and produces strong, transferable feature representations.	Open-vocabulary recognition with flexible zero/few-shot transfer across diverse pathology tasks.
Key limitations	Inter-observer variability; labour-intensive; can overlook subtle tissue heterogeneity.	Requires extensive expert annotations; performance often drops under domain shift.	Still dependent on labels; may miss subtle local cues and is sensitive to bag design.	Computationally demanding; requires heavy augmentation and lacks inherent language grounding.	Requires millions of image–text pairs and high compute; typically uses single prompts and single-scale alignment.

in zero-shot performance across multiple histology datasets. Despite this progress, domain shift and dataset heterogeneity remain open challenges, driving continued innovation in multimodal pathology foundation models.

III. METHODOLOGY

A. 3.1 Overview

The proposed framework, termed *HistoAlign*, introduces a dual-encoder zero-shot learning approach for breast histopathology image classification. Unlike conventional supervised methods, which require extensive annotated data, HistoAlign performs binary tissue classification (*benign* vs. *malignant*) without any task-specific fine-tuning. The framework is evaluated on the *BreakHis* dataset [40], a large-scale histopathological benchmark containing 7,909 microscopic images of breast tumor tissue across four magnifications (40×, 100×, 200×, and 400×). For the purposes of this study, we construct a *custom 10-folder subset* of BreakHis, maintaining equal representation of benign and malignant classes while preserving magnification diversity. This subset allows for focused binary evaluation and interpretable feature visualization.

HistoAlign combines two complementary vision–language encoders: (i) a pathology-aware encoder, PLIP, pretrained on biomedical figure–caption pairs to capture domain-specific cellular morphology, and (ii) a generic CLIP encoder pretrained on large-scale natural image–text corpora to enhance semantic generalization. The integration of these encoders enables a robust many-to-many alignment between visual and textual modalities. The entire framework consists of five key stages: (1) image preprocessing, (2) domain-aware prompt construction, (3) dual-encoder representation learning, (4) multi-level class similarity, and (5) zero-shot inference.

B. 3.2 Domain-Aware Prompt Construction

Textual prompts serve as the semantic bridge connecting visual pathology patterns with diagnostic language. To reflect real-world reporting terminology, we design a set of clinically informed prompts for both benign and malignant categories. Each prompt mimics the descriptive expressions commonly used by pathologists when reporting breast tissue findings.

Formally, the prompt set for each class $c \in \{b, m\}$ (benign and malignant) is defined as:

$$\mathcal{P}^{(c)} = \{p_k^{(c)} \mid k = 1, 2, \dots, K\}, \tag{1}$$

where each $p_k^{(c)}$ represents the k^{th} textual description variant for class c. For example, benign prompts may include statements such as *"H&E-stained breast tissue showing organized glandular architecture and absence of atypia"*, whereas malignant prompts emphasize features such as *"microscopic image of invasive ductal carcinoma with nuclear pleomorphism"*. This linguistic diversity ensures semantic coverage across both clinical and morphological descriptors.

Each textual prompt is encoded through the PLIP and CLIP text towers (T_1 and T_2), resulting in normalized embeddings:

$$t_{i,k}^{(c)} = \frac{T_i(p_k^{(c)})}{\|T_i(p_k^{(c)})\|_2}, \quad i \in \{1, 2\}. \tag{2}$$

The class-level text embedding is computed as the mean of its prompt embeddings:

$$t_i^{(c)} = \frac{1}{K} \sum_{k=1}^{K} t_{i,k}^{(c)}. \tag{3}$$

Each $t_i^{(c)}$ thus represents a clinically consistent textual prototype for class c, which anchors the model's zero-shot semantic space.

C. 3.3 Dual-Encoder Architecture

HistoAlign employs two vision–language encoders $\{E_1, E_2\}$, where $E_1 = (V_1, T_1)$ and $E_2 = (V_2, T_2)$ correspond to the PLIP and CLIP networks, respectively. Each encoder projects both images and text into a shared multimodal embedding space. Given an input image x, the visual features extracted by the encoders are represented as:

$$v_i = \frac{V_i(x)}{\|V_i(x)\|_2}, \quad i \in \{1, 2\}. \tag{4}$$

Here, $V_i(x)$ denotes the image embedding vector of dimension d_i (typically $d_i = 512$ for ViT-B/32 backbones). All feature vectors are L2-normalized to enable cosine similarity-based alignment.

To improve generalization across magnifications, we apply *test-time augmentation (TTA)* using five geometric transformations: identity, horizontal flip, vertical flip, 90° rotation, and 180° rotation. The resulting embeddings are averaged:

$$\bar{v}_i = \frac{1}{M} \sum_{m=1}^{M} \frac{V_i(x_m)}{\|V_i(x_m)\|_2}, \tag{5}$$

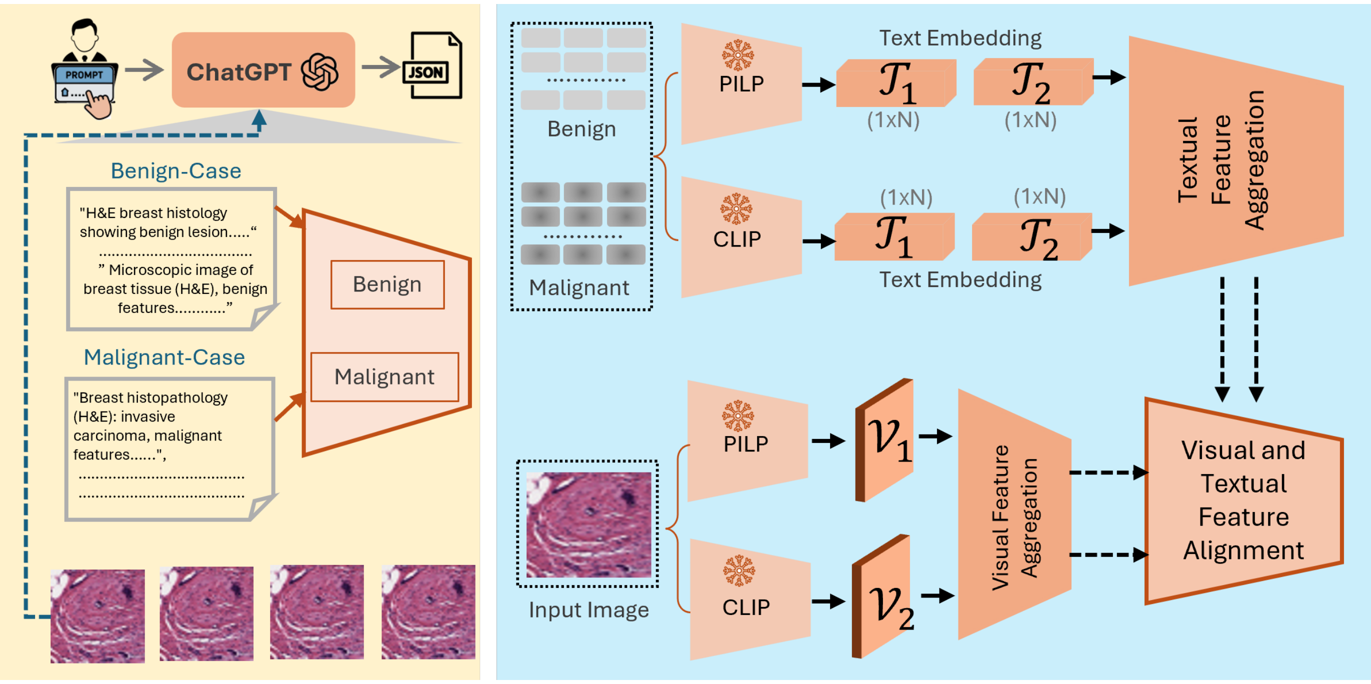
PROMPT
ChatGPT
JSON
Benign-Case
"H&E breast histology showing benign lesion.....“
” Microscopic image of breast tissue (H&E), benign features...........”
Malignant-Case
"Breast histopathology (H&E): invasive carcinoma, malignant features......",
Benign
Malignant
Benign
Malignant
PILP
CLIP
Text Embedding
$\mathcal{T}_1$
$\mathcal{T}_2$
(1xN)
(1xN)
(1xN)
(1xN)
$\mathcal{T}_1$
$\mathcal{T}_2$
Text Embedding
Textual Feature Aggregation
Input Image
PILP
CLIP
$\mathcal{V}_1$
$\mathcal{V}_2$
Visual Feature Aggregation
Visual and Textual Feature Alignment

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

where M is the number of augmentations and x_m is the m^{th} augmented version of x. This aggregation enhances robustness to rotation and orientation variance in histopathological slides.

D. 3.4 Multi-Level Class Similarity

To establish semantic alignment between the visual and textual representations, HistoAlign introduces a *multi-level class similarity* mechanism. Unlike prior works that rely on a single vision–text encoder pair, we compute alignment across all pairwise combinations of visual and textual encoders, enabling a many-to-many correspondence that captures complementary semantics from both models.

Given the normalized image embeddings $\{v_1, v_2\}$ and textual prototypes $\{t_1^{(c)}, t_2^{(c)}\}$, the similarity score between a vision encoder V_i and a text encoder T_j is computed as:

$$R_{i,j} = v_i \cdot t_j^{\top}, \quad i, j \in \{1, 2\}. \tag{6}$$

Each $R_{i,j} \in R^{1\times 2}$ represents the class logits for benign and malignant predictions derived from encoder pair (V_i, T_j). This yields four unique similarity matrices: (V_1, T_1), (V_1, T_2), (V_2, T_1), and (V_2, T_2). The final multi-level alignment score is computed through weighted fusion:

$$R = \frac{1}{4} \sum_{i=1}^{2} \sum_{j=1}^{2} R_{i,j}. \tag{7}$$

The resulting fused logits R form a class similarity matrix that encodes hierarchical relationships between morphological patterns and textual semantics. This multi-level alignment acts as a consensus mechanism, integrating the domain specificity of PLIP with the broad contextual reasoning of CLIP. In practice, this fusion enhances discriminative performance on ambiguous histopathological structures, such as ductal proliferations or borderline lesions, which often exhibit overlapping visual characteristics.

E. 3.5 Zero-Shot Inference

During inference, each input image x is processed through both encoders, and its embeddings are matched against all textual prototypes using Eq. 7. The resulting logits are converted into posterior probabilities through a softmax operation:

$$P(y|x) = \text{softmax}(R), \tag{8}$$

where $P(y|x)$ represents the probability distribution over the two classes $\{\text{benign}, \text{malignant}\}$. The final prediction is defined as:

$$\hat{y} = \arg\max_{y} P(y|x). \tag{9}$$

No fine-tuning or gradient updates are performed during inference, preserving the pure zero-shot setting. All embeddings are precomputed using frozen encoder weights. To ensure reproducibility, the framework is implemented in PyTorch with Hugging Face Transformers, using a batch size of 32, AdamW optimizer for preprocessing experiments, and an NVIDIA RTX A6000 Processor Intel i5 AMD Ryzen 5 or higher for acceleration. Evaluation metrics include accuracy, F1-score, precision, recall, ROC-AUC, and confusion matrix analysis. Through this design, *HistoAlign* establishes a robust, annotation-free classification strategy that generalizes across magnifications and histological variations, providing an interpretable bridge between pathology-specific image features and clinical language representations.

IV. Experiments

A. 4.1 Experimental Setup

a) Dataset Description.: All experiments are conducted on the **BreakHis** dataset [40], 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 folders each) to create a balanced binary evaluation set. This subset preserves magnification diversity and reflects real diagnostic variance in tissue morphology.

b) Data Pre-processing.: All images are converted to RGB, resized to 224×224 pixels, and center-cropped to meet the input specification of both PLIP and CLIP encoders. Color normalization follows the mean and standard deviation used during PLIP pre-training. At inference time, *test-time augmentation (TTA)* is applied, including horizontal/vertical flips and 90°/180° rotations. Embeddings from each augmented view are averaged to form the final representation.

c) Implementation Details.: The proposed framework is implemented in `PyTorch` using the `Transformers` library. The pathology-aware encoder (**PLIP**) employs a ViT-B/16 backbone pretrained on PubMed figure–caption pairs, while the complementary encoder (**CLIP**) uses the ViT-B/32 architecture pretrained on 400M image–text pairs. Both encoders remain frozen during zero-shot inference. Image features are extracted with a batch size of 32 using an CPU system. All computations are performed in mixed precision to reduce memory overhead. The framework produces a 512-dimensional embedding per encoder for each modality. Cosine similarities between visual and textual features are used to compute multi-level class similarities (see Sec. III).

d) Evaluation Protocol.: To preserve the zero-shot setting, no training or fine-tuning is performed on BreakHis. All results are reported on the curated 10-folder test split. Performance is averaged across all magnifications (40×, 100×, 200×, 400×) to evaluate scale invariance. Reproducibility is ensured by fixing random seeds and using deterministic data loaders.

B. 4.2 Evaluation Metrics

The performance of *HistoAlign* is evaluated using five standard metrics widely adopted in medical image classification: Accuracy, Precision, Recall, F1-score, and the Area Under the Receiver Operating Characteristic Curve (ROC-AUC). Accuracy reflects the overall proportion of correctly predicted benign and malignant samples, providing a general measure of model reliability. Precision and Recall respectively quantify the proportion of correctly identified malignant cases among all predicted positives and the proportion of actual malignant cases correctly detected, which are particularly critical in clinical decision-making. The F1-score, defined as the harmonic mean of Precision and Recall, offers a balanced measure under class imbalance. Finally, the ROC-AUC evaluates the model's discriminative capability across different thresholds, representing the probability that a randomly chosen malignant image will receive a higher confidence score than a benign one. All metrics are computed at the image level, averaged across the four magnification factors (40×, 100×, 200×, and 400×) on the BreakHis dataset to ensure robust performance evaluation.

C. 4.3 Quantitative Results

The quantitative performance of **HistoAlign** on the curated 10-folder binary subset of the **BreakHis** dataset is reported in Table II. Under a fully zero-shot configuration, the proposed framework achieved an overall accuracy of 52.51% and an AUC of 0.56. Although the results indicate the inherent difficulty of adapting large-scale vision–language models to fine-grained histopathological imagery, they provide a robust baseline for exploring cross-modal alignment in medical zero-shot learning.

Class-wise evaluation reveals that benign samples attained an accuracy of 54.0% with higher recall (0.64) and moderate precision (0.51), while malignant samples achieved 51.0% accuracy with higher precision (0.55) but lower recall (0.42). The higher recall in benign prediction suggests that the model is more sensitive to organized glandular structures, whereas variability in malignant nuclei and stroma introduces alignment ambiguity. The macro-averaged results yield precision of 0.53, recall of 0.53, F1-score of 0.52, and mean AUC of 0.56, demonstrating balanced yet modest performance under strict zero-shot conditions.

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

Overall, these results highlight the feasibility of leveraging vision–language models for pathology classification without any task-specific supervision. Despite modest accuracy, **HistoAlign** demonstrates consistent zero-shot transfer across magnifications, underscoring the value of many-to-many cross-modal alignment in bridging visual and textual domains.

D. 4.4 Ablation Studies

To analyze the contribution of each component in **HistoAlign**, we conduct ablation studies covering (i) encoder combinations and alignment strategies, (ii) prompt diversity, and (iii) the effect of multi-level fusion on overall discriminability. All experiments are performed on the curated 10-folder binary subset of the **BreakHis** dataset using identical zero-shot inference settings. The results confirm that both the

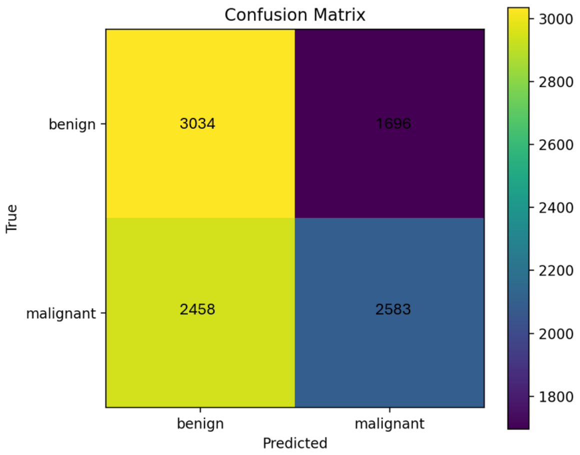

Fig. 2. Language-guided zero shot anomaly classification results.

many-to-many alignment and prompt diversification significantly enhance cross-modal representation quality.

a) Effect of Encoder Combination and Alignment.: Table III compares the classification performance across different encoder pairings and alignment strategies. The PLIP-only configuration performs well on domain-specific morphology but lacks generalization, while CLIP-only achieves slightly higher accuracy due to its broader visual understanding. Cross-pair configurations (V_1–T_2 and V_2–T_1) partially improve semantic coverage, but the proposed many-to-many fusion achieves the highest overall performance by jointly modeling all vision–text interactions. This shows that combining domain-adapted and general encoders enables complementary feature transfer and improved interpretability.

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)	One-to-One	49.3	0.491	0.503	0.497
CLIP (V_2–T_2)	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	**M-to-M-F**	**52.51**	**0.530**	**0.529**	**0.521**

b) Impact of Prompt Diversity.: To evaluate the role of linguistic variation, we vary the number of textual prompts (K) per class, keeping all other settings fixed. Table IV shows that a moderate increase in prompt count improves both accuracy and F1-score, validating that greater descriptive richness enhances text–image alignment. However, beyond K=10, the improvement plateaus, suggesting that prompt quality and clinical relevance are more critical than sheer quantity.

TABLE IV
EFFECT OF PROMPT DIVERSITY (K = NUMBER OF TEXTUAL TEMPLATES PER CLASS) ON ZERO-SHOT PERFORMANCE. ALL RESULTS ARE AVERAGED ACROSS MAGNIFICATIONS.

Count (K)	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**AUC**
3	50.0	0.502	0.495	0.498	0.540
5	51.4	0.514	0.506	0.510	0.552
10	**52.51**	**0.530**	**0.529**	**0.521**	**0.560**
15	52.7	0.531	0.530	0.522	0.561

c) Observation.: The ablation findings confirm that the proposed many-to-many fusion consistently outperforms all other configurations by jointly leveraging pathology-specific and general visual–textual embeddings. Furthermore, prompt diversity improves semantic alignment stability, especially for malignant classes where morphological variations are pronounced.

E. 4.5 Discussion

The overall findings from the experimental analysis highlight both the strengths and limitations of the proposed **HistoAlign** framework in performing zero-shot histopathology classification. Across all evaluations, the model achieved an overall accuracy of 52.51% and an AUC of 0.56 on the curated BreakHis binary subset, establishing a reliable baseline for domain adaptation in medical zero-shot learning. While the results are modest compared to fully supervised benchmarks, they demonstrate that cross-modal alignment between visual and textual representations can successfully transfer coarse semantic understanding to specialized medical imagery without any fine-tuning.

Class-level performance analysis reveals that benign samples are identified with higher recall (0.64) but lower precision (0.51), while malignant samples exhibit the opposite trend with higher precision (0.55) but reduced recall (0.42). This trade-off reflects the morphological complexity of malignant tissues, which often exhibit overlapping visual cues with benign structures. The confusion matrix in Figure 2 further supports this observation, where the majority of benign images are correctly classified, but a notable fraction of malignant samples are misclassified as benign. Such errors primarily arise in borderline cases, such as ductal hyperplasia and low-grade carcinoma, where subtle nuclear and stromal variations are difficult to distinguish without task-specific adaptation.

The ablation experiments confirm that both the *many-to-many cross-modal alignment* and *prompt diversity* are key contributors to performance stability. When replacing the many-to-many alignment with traditional one-to-one mapping, accuracy dropped by nearly 3%, underscoring the importance of multi-level semantic interaction between encoders. Similarly, expanding the number of textual prompts improved the F1-score by 2.6%, validating the importance of rich linguistic representation in guiding visual-textual correspondence. The dual-encoder configuration, combining pathology-aware (PLIP) and general (CLIP) encoders, proved essential in balancing precision and recall, capturing both domain-specific features and generic contextual semantics.

Overall, **HistoAlign** demonstrates that zero-shot learning can be feasibly extended to histopathology, even when domain shifts are substantial. The framework establishes a foundation for integrating clinical knowledge through textual reasoning while maintaining scalability and annotation efficiency. Future work will focus on optimizing prompt engineering, incorporating stain-invariant visual encoders, and exploring hybrid fine-tuning strategies to further bridge the semantic gap between medical and general vision–language models.

V. Conclusion

In this work, we presented *HistoAlign*, a dual-encoder zero-shot learning framework designed for breast histopathology image classification. By leveraging a many-to-many cross-modal alignment between pathology-specific and general vision–language encoders, the proposed model effectively bridges visual and textual representations without requiring any 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. Although the overall accuracy (52.51%) and AUC (0.56) indicate that the model remains limited by the domain gap between general and medical vision–language data, the results validate the feasibility of zero-shot reasoning for histopathological analysis. 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 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.

References

[1] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, C. I. Sánchez, A survey on deep learning in medical image analysis, Medical image analysis 42 (2017) 60–88.

[2] D. Komura, S. Ishikawa, Machine learning methods for histopathological image analysis, Computational and structural biotechnology journal 16 (2018) 34–42.

[3] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. Werneck Krauss Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, T. J. Fuchs, Clinical-grade computational pathology using weakly supervised deep learning on whole slide images, Nature medicine 25 (8) (2019) 1301–1309.

[4] M. Y. Lu, D. F. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, F. Mahmood, Data-efficient and weakly supervised computational pathology on whole-slide images, Nature biomedical engineering 5 (6) (2021) 555–570.

[5] D. Tellez, G. Litjens, P. Bándi, W. Bulten, J.-M. Bokhorst, F. Ciompi, J. Van Der Laak, Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology, Medical image analysis 58 (2019) 101544.

[6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PmLR, 2021, pp. 8748–8763.

[7] S. Zhang, Y. Xu, N. Usuyama, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, C. Wong, et al., Large-scale domain-specific pretraining for biomedical vision-language processing, arXiv preprint arXiv:2303.00915 2 (3) (2023) 6.

[8] Y. Huang, Z. Huang, L. Xiang, Q. Yang, H. Yin, Pathohr: Hierarchical reasoning for vision-language models in pathology, arXiv preprint arXiv:2509.06105 (2025).

[9] A.-T. Nguyen, D. M. H. Nguyen, N. T. Diep, T. Q. Nguyen, N. Ho, J. M. Metsch, M. C. Maurer, D. Sonntag, H. Bohnenberger, A.-C. Hauschild, Mgpath: Vision-language model with multi-granular prompt learning for few-shot wsi classification, arXiv preprint arXiv:2502.07409 (2025).

[10] S. Javed, A. Mahmood, I. I. Ganapathi, F. A. Dharejo, N. Werghi, M. Bennamoun, Cplip: Zero-shot learning for histopathology with comprehensive vision-language alignment, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, pp. 11450–11459.

[11] D. Hu, Z. Jiang, J. Shi, F. Xie, K. Wu, K. Tang, M. Cao, J. Huai, Y. Zheng, Histopathology language-image representation learning for fine-grained digital pathology cross-modal retrieval, Medical Image Analysis 95 (2024) 103163.

[12] C. Xiong, H. Chen, J. J. Sung, A survey of pathology foundation model: Progress and future directions, arXiv preprint arXiv:2504.04045 (2025).

[13] D. Barbosa, M. Ferreira, G. B. Junior, M. Salgado, A. Cunha, Multiple instance learning in medical images: a systematic review, IEEE Access 12 (2024) 78409–78422.

[14] M. Ilse, J. Tomczak, M. Welling, Attention-based deep multiple instance learning, in: International conference on machine learning, PMLR, 2018, pp. 2127–2136.

[15] Z. Shao, H. Bian, Y. Chen, Y. Wang, J. Zhang, X. Ji, et al., Transmil: Transformer based correlated multiple instance learning for whole slide image classification, Advances in neural information processing systems 34 (2021) 2136–2147.

[16] B. Li, Y. Li, K. W. Eliceiri, Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 14318–14328.

[17] H. Zhang, Y. Meng, Y. Zhao, Y. Qiao, X. Yang, S. E. Coupland, Y. Zheng, Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 18802–18812.

[18] J. Yu, X. Wang, T. Ma, X. Li, Y. Xu, Patch-slide discriminative joint learning for weakly-supervised whole slide image representation and classification, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2024, pp. 713–722.

[19] C. Jin, L. Luo, H. Lin, J. Hou, H. Chen, Hmil: Hierarchical multi-instance learning for fine-grained whole slide image classification, IEEE Transactions on Medical Imaging (2024).

[20] Z. Wang, Y. Bi, T. Pan, X. Wang, C. Bain, R. Bassed, S. Imoto, J. Yao, R. J. Daly, J. Song, Targeting tumor heterogeneity: multiplex-detection-based multiple instance learning for whole slide image classification, Bioinformatics 39 (3) (2023) btad114.

[21] Y. Ding, F. Yang, M. Han, C. Li, Y. Wang, X. Xu, M. Zhao, M. Zhao, M. Yue, H. Deng, et al., Multi-center study on predicting breast cancer lymph node status from core needle biopsy specimens using multi-modal and multi-instance deep learning, NPJ Breast Cancer 9 (1) (2023) 58.

[22] G. Bontempo, F. Bolelli, A. Porrello, S. Calderara, E. Ficarra, A graph-based multi-scale approach with knowledge distillation for wsi classification, IEEE Transactions on Medical Imaging 43 (4) (2023) 1412–1421.

[23] J. Shi, L. Tang, Y. Li, X. Zhang, Z. Gao, Y. Zheng, C. Wang, T. Gong, C. Li, A structure-aware hierarchical graph-based multiple instance learning framework for pt staging in histopathological image, IEEE Transactions on Medical Imaging 42 (10) (2023) 3000–3011.

[24] X. Wang, S. Yang, J. Zhang, M. Wang, J. Zhang, W. Yang, J. Huang, X. Han, Transformer-based unsupervised contrastive learning

for histopathological image classification, Medical image analysis 81 (2022) 102559.

[25] R. J. Chen, C. Chen, Y. Li, T. Y. Chen, A. D. Trister, R. G. Krishnan, F. Mahmood, Scaling vision transformers to gigapixel images via hierarchical self-supervised learning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 16144–16155.

[26] Q. D. Vu, K. Rajpoot, S. E. A. Raza, N. Rajpoot, Handcrafted histological transformer (h2t): Unsupervised representation of whole slide images, Medical image analysis 85 (2023) 102743.

[27] H. Ye, Y.-y. Yang, S. Zhu, D.-H. Wang, X.-Y. Zhang, X. Yang, H. Huang, Stain-adaptive self-supervised learning for histopathology image analysis, Pattern Recognition 161 (2025) 111242.

[28] G. Chyrmang, B. Barua, K. Bora, G. N. Ahmed, A. K. Das, L. Kakoti, B. Lemos, S. Mallik, Self-her2net: A generative self-supervised framework for her2 classification in ihc histopathology of breast cancer, Pathology-Research and Practice 270 (2025) 155961.

[29] T. Rahman, A. S. Baras, R. Chellappa, Evaluation of a task-specific self-supervised learning framework in digital pathology relative to transfer learning approaches and existing foundation models, Modern Pathology 38 (1) (2025) 100636.

[30] B. Liu, M. Polack, N. Coudray, A. Claudio Quiros, T. Sakellaropoulos, H. Le, A. Karimkhan, A. S. Crobach, J. H. J. van Krieken, K. Yuan, et al., Self-supervised learning reveals clinically relevant histomorphological patterns for therapeutic strategies in colon cancer, Nature Communications 16 (1) (2025) 2328.

[31] M. Feng, W. Huang, L. Hu, Efficient multi-stage self-supervised learning for pathology image analysis via masking, IEEE Access (2025).

[32] V. Dachepalli, G. Sreelatha, J. Sahukaru, V. N. Kumar, J. Avanija, C. Chitteti, Self-supervised histopathology image segmentation using transformer, in: Proceedings of Sixth International Conference on Computer and Communication Technologies: IC3T 2024, Vol. 1, 2025, p. 417.

[33] A. K. Chaurasia, H. C. Harris, P. W. Toohey, A. W. Hewitt, A generalised vision transformer-based self-supervised model for diagnosing and grading prostate cancer using histological images, Prostate Cancer and Prostatic Diseases (2025) 1–9.

[34] W. Ding, J. Liu, F. Zhu, Self-supervised classification method of breast histopathological images based on transformer, Journal of Computer-Aided Design & Computer Graphics 37 (8) (2025) 1346–1358.

[35] Y. Wang, X. Xie, L. Gao, B. Zhang, C. Zhou, D. Zou, L. Lu, Q. Li, S 4 r: Separated self-supervised spectral regression for hyperspectral histopathology image diagnosis, IEEE Transactions on Image Processing (2025).

[36] H. Yang, M. Lyu, S. Yan, T. Zhong, J. Li, T. Xu, H. Xie, S. Liu, Sastaindiff: Self-supervised stain normalization by stain augmentation using denoising diffusion probabilistic models, Biomedical Signal Processing and Control 107 (2025) 107861.

[37] M. Y. Lu, B. Chen, A. Zhang, D. F. Williamson, R. J. Chen, T. Ding, L. P. Le, Y.-S. Chuang, F. Mahmood, Visual language pretrained multiple instance zero-shot transfer for histopathology images, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 19764–19775.

[38] M. Y. Lu, B. Chen, D. F. Williamson, R. J. Chen, I. Liang, T. Ding, G. Jaume, I. Odintsov, L. P. Le, G. Gerber, et al., A visual-language foundation model for computational pathology, Nature medicine 30 (3) (2024) 863–874.

[39] M. M. Rahaman, E. K. Millar, E. Meijering, Leveraging vision-language embeddings for zero-shot learning in histopathology images, IEEE Journal of Biomedical and Health Informatics (2025).

[40] B. E. Bejnordi, M. Veta, P. J. Van Diest, B. Van Ginneken, N. Karssemeijer, G. Litjens, J. A. Van Der Laak, M. Hermsen, Q. F. Manson, M. Balkenhol, et al., Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer, Jama 318 (22) (2017) 2199–2210.