PathFLIP: Fine-grained Language-Image Pretraining for Versatile Computational Pathology

Fengchun Liu[*1], **Songhan Jiang**[*1], **Linghan Cai**[1], **Ziyue Wang**[2], **Yongbing Zhang**[1]

[1]Harbin Institute of Technology, Shenzhen, School of Computer Science and Technology
[2]National University of Singapore, Department of Electronic and Computer Engineering
{24s051033, 23s151019, cailh}@stu.hit.edu.cn, e1374378@u.nus.edu, ybzhang08@hit.edu.cn

Abstract

While Vision-Language Models (VLMs) have achieved notable progress in computational pathology (CPath), the gigapixel scale and spatial heterogeneity of Whole Slide Images (WSIs) continue to pose challenges for multimodal understanding. Existing alignment methods struggle to capture fine-grained correspondences between textual descriptions and visual cues across thousands of patches from a slide, compromising their performance on downstream tasks. In this paper, we propose PathFLIP (**Path**ology **F**ine-grained **L**anguage-**I**mage **P**retraining), a novel framework for holistic WSI interpretation. PathFLIP decomposes slide-level captions into region-level subcaptions and generates text-conditioned region embeddings to facilitate precise visual-language grounding. By harnessing Large Language Models (LLMs), PathFLIP can seamlessly follow diverse clinical instructions and adapt to varied diagnostic contexts. Furthermore, it exhibits versatile capabilities across multiple paradigms, efficiently handling slide-level classification and retrieval, fine-grained lesion localization, and instruction following. Extensive experiments demonstrate that PathFLIP outperforms existing large-scale pathological VLMs on four representative benchmarks while requiring significantly less training data, paving the way for fine-grained, instruction-aware WSI interpretation in clinical practice.

Code — https://github.com/cyclexfy/PathFLIP

Introduction

Whole Slide Images (WSIs) are fundamental to pathology diagnosis(Cai et al. 2025), yet their gigapixel resolution and complex spatial structures pose significant challenges for multimodal learning(Jiang et al. 2024, 2025). Since each slide contains thousands of heterogeneous patches, achieving fine-grained visual-textual alignment is challenging. To address this, the emergence of large-scale multimodal datasets (Chen et al. 2025), pairing WSIs with expert-authored slide captions and driving progress in the field. These captions are extended and detail-rich, reflecting clinical workflows and supporting diagnostic reasoning.

This expanding multimodal opportunity aligns with advances in Vision-Language Models (VLMs), particularly

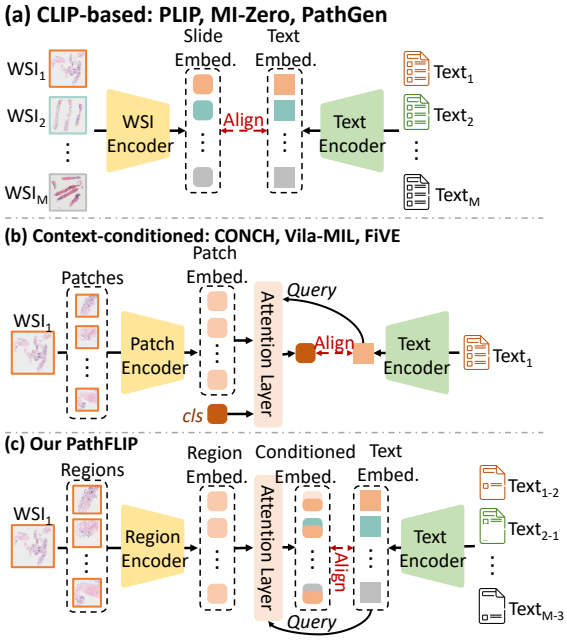

Figure 1: Comparison of PathFLIP with previous methods in vision-language pathology modeling. (a) CLIP-based methods perform coarse global feature alignment between images and text. (b) Context-conditioned approaches use textual cues to guide attention but focus on global alignment. (c) Our PathFLIP enables fine-grained pathology analysis by aligning localized image embeddings with semantically matched text segments.

those leveraging contrastive pretraining paradigms exemplified by CLIP (Radford et al. 2021; Zheng et al. 2024; Xiao et al. 2025), which have shown impressive capabilities in learning joint image-text representations across general domains. However, when applied to computational pathology (CPath), these models (Lu et al. 2023; Huang et al. 2023) face unique challenges arising from the gigapixel scale and spatial heterogeneity of WSIs (Figure 1(a)). Most existing methods extract global feature of a slide from thousands of patches, followed by global alignment with slide-level captions. Nevertheless, this coarse-grained strategy fails to cap-

[*]These authors contributed equally.

ture fine-grained correspondences between visual regions and textual descriptions (Sun et al. 2025b). To overcome these limitations, some approaches adapt VLMs to the multiple instance learning (MIL) framework, aggregating patch-level features into slide-level representations through attention mechanisms guided by textual cues (Shi et al. 2024; Ahmed et al. 2024; Tang et al. 2025; Dong et al. 2025), as shown in Figure 1(b). These methods leverage contextual signals from captions to highlight semantically relevant regions before performing slide-text alignment between visual and textual embeddings (Guo et al. 2025; Li et al. 2024). While these strategies enhance localization, they typically depend on weak or implicit region-text supervision and fall short of fully emulating the diagnostic reasoning of pathologists, which requires grounding from region-level observations to slide-level conclusions (Ghezloo et al. 2025).

Targeting the above challenges, we propose PathFLIP (**Path**ology **F**ine-grained **L**anguage-Image **P**retraining), a novel framework for fine-grained WSI analysis. As illustrated in Figure 1(c), PathFLIP establishes region-text correspondences by aligning localized visual features with semantically matched text segments. Concretely, we design a Region Q-Former to extract region-level embeddings from the WSI. Meanwhile, the slide-level captions are decomposed into region-level subcaptions that reflect localized pathological semantics and serve as anchors to guide region-level grounding, promoting the disentanglement of interpretable clinical concepts. At the region level, we introduce a self-supervised contrastive objective that treats region-subcaption pairs from the same slide as positives, while considering all other combinations in the batch as negatives. This encourages precise and discriminative region-text alignment. To retain a global understanding, we align WSI representations with their corresponding full captions using a global contrastive loss. This pretraining strategy enhances alignment accuracy and interpretability by enabling fine-grained supervision without manual regional annotations.

Building upon these foundations, PathFLIP demonstrates exceptional capabilities, seamlessly adapting across diverse multimodal pathology tasks. Through efficient integration with Large Language Models (LLMs), PathFLIP acquires advanced instruction-following abilities such as caption generation and visual question answering (VQA) within a unified framework. At the region level, PathFLIP excels at semantic understanding, accurately identifying lesion-relevant areas within WSIs. This enables detailed region-level visual grounding of pathological descriptions. For slide-level tasks, PathFLIP delivers robust performance in zero-shot classification and retrieval, achieving competitive results and significantly surpassing previous large-scale VLMs across four benchmarks. Collectively, these strengths position PathFLIP as a versatile, unified model for computational pathology, capable of effectively addressing various multimodal understanding challenges. Key contributions are as follows:

- We propose PathFLIP, a fine-grained language-image pretraining framework specifically for WSIs. By leveraging subcaption–guided embedding alignment, PathFLIP enables accurate region-level grounding without expert annotations. These localized semantics are further aggregated to support holistic slide-level diagnosis and comprehensive multimodal interpretation, closely aligning with the typical workflow of pathologists.
- By integrating LLMs, PathFLIP gains the flexibility to address a wide range of multimodal pathology tasks at both the region and slide levels, including lesion localization, instruction-following, and pathological diagnosis.
- Extensive experiments demonstrate that PathFLIP consistently outperforms existing large-scale VLMs on diverse clinical benchmarks, requiring significantly less training data. Its interpretable and multimodal capabilities enable practical deployment in diagnostic support and clinical AI applications.

Related Work

Pretraining in Computational Pathology

Pretraining in computational pathology leverages large-scale datasets and self-supervised learning to produce general-purpose visual representation models (Li et al. 2025). Unimodal models focus solely on visual features, utilizing hierarchical structures in WSIs through pyramid Vision Transformers (Chen et al. 2022) and masked autoencoding with self-distillation (Chen et al. 2024; Xu et al. 2024). Recent works bridge the gap between visual and textual modalities by employing cross-modal contrastive learning (Wang et al. 2024; Sun et al. 2025a). PathGen (Sun et al. 2025b) retrieves semantically relevant patches using cross-modal signals to train a high-quality patch encoder. TITAN (Ding et al. 2024) aligns regional embeddings with region captions and slide-level reports, while CONCH (Lu et al. 2024) enables fine-grained alignment via context-aware patch embeddings. In contrast, our PathFLIP unifies slide-level and fine-grained alignment within a single framework, achieving superior performance with significantly fewer training samples.

Pathological Vision-Language Models

While CLIP-based methods in computational pathology (Lu et al. 2023; Huang et al. 2023; Sun et al. 2025b; Liu et al. 2025; Yuan et al. 2018) apply contrastive learning to align slide representations with their global captions, they lack supervision signals for fine-grained region-text alignment. Motivated by the success of LLMs, several studies have explored more effective ways to bridge pathology images and textual semantics. For example, PathAlign (Ahmed et al. 2024) extends the BLIP framework with LLMs to support diverse WSI applications. SlideChat (Chen et al. 2025) utilizes a slide encoder to obtain global slide embeddings, which are sent into an LLM for downstream slide-level multimodal tasks. PRISM2 (Shaikovski et al. 2025) introduces a two-stage training framework that aligns slide embeddings with diagnostic summaries, and then incorporates LLM-guided supervision using clinical prompts. Nevertheless, existing methods fail to achieve fine-grained reasoning and ground visual evidence based on textual cues. In contrast, our approach unifies multimodal understanding across regions and slide levels, while supporting diverse pathology tasks in real-world clinical settings.

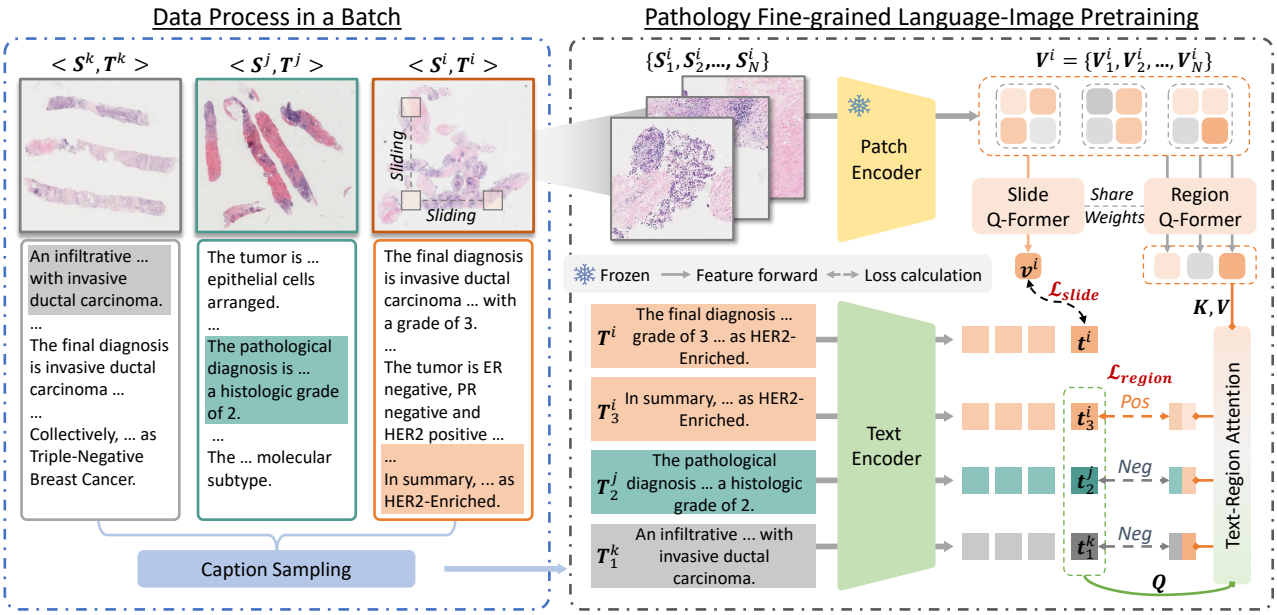

Figure 2: Overview of PathFLIP. Given a slide-caption pair $<S^i, T^i>$, the slide S^i is divided into N regions $\{S_1^i, \ldots, S_N^i\}$. We use Slide Q-Former and Region Q-Former to extract slide-level and region-level features. Captions $\{T^k, T^j, T^i\}$ are decomposed and sampled to obtain region-level subcaptions $\{T_1^k, T_2^j, T_3^i\}$. The slide-level contrastive loss $\mathcal{L}_{slide}$ aligns the global image feature v^i with its corresponding text feature t^i. The region-level contrastive loss $\mathcal{L}_{region}$ encourages alignment between region-image and subcaption pairs from the same slide as positive pairs, while treating all others in the batch as negatives.

Methodology

Overall Structure of PathFLIP

The overall framework of PathFLIP is illustrated in Figure 2. We describe the visual feature extraction process at both slide and region levels, then construct region-level subcaptions and their textual representations. We then introduce a contrastive pretraining objective that jointly aligns visual and textual features, incorporating both text-conditioned region-level alignment and global slide-level alignment. This framework enables to learn semantically aligned visual representations, facilitating effective transfer to various downstream tasks in Figure 3.

Visual Representation

Region-level Visual Feature Extraction Given the gigapixel resolution of WSIs, we adopt a two-stage partitioning strategy to extract region-level visual features from an input slide S^i. First, the slide is divided into N non-overlapping regions $\{S_n^i\}_{n=1}^N$ using a sliding window. Each region S_n^i is further partitioned into M non-overlapping patches, and each patch is encoded into a d-dimensional feature vector using a pretrained pathology-specific visual encoder. This results in a region-level feature matrix $V_n^i \in \mathbb{R}^{M\times d}$ for each region n of slide S^i. The slide-level feature matrix is obtained by concatenating all patch-level features across all regions, denoted as $V^i \in \mathbb{R}^{MN\times d}$.

Region and Slide Q-Former After feature extraction, we obtain N non-overlapping regions per slide, each consisting of M independently encoded patch features. Due to the lack of spatial dependency modeling and redundancy among patch representations, compact and informative semantic aggregation becomes essential. Inspired by BLIP-2 (Li et al. 2023b), we adopt a Q-Former module to enhance intra-region semantic fusion. It consists of self-attention, cross-attention, and feed-forward layers, and uses a learnable set of semantic query tokens $Q \in \mathbb{R}^{N_q\times d}$. Since $N_q \ll M$, the Q-Former efficiently distills the salient semantics.

We apply the Region Q-Former $f_{rv}(\cdot,\cdot)$ to each region feature V_n^i to obtain fine-grained embeddings:

$$v_n^i = f_{rv}(Q, V_n^i)\,, \tag{1}$$

where $v_n^i \in \mathbb{R}^{N_q\times d}$ represents the semantic embedding of region n in slide S^i. The resulting region embeddings $\{v_n^i\}_{n=1}^N$ are concatenated along the first dimension to form $\mathbf{v}^{\text{reg}} \in \mathbb{R}^{NN_q\times d}$. Similarly, the Slide Q-Former $f_{sv}(\cdot,\cdot)$ is applied to the full patch set V^i of slide S^i to derive a global representation:

$$v^i = f_{sv}(Q, V^i)\,, \tag{2}$$

where $v^i \in \mathbb{R}^{N_q\times d}$ denotes the holistic embedding of the slide. The Q-Former shares weights at the region and slide levels to ensure consistency and efficiency in feature aggregation across scales. The resulting global embedding v^i and regional features $\mathbf{v}^{\text{reg}}$ collaboratively support holistic slide understanding and fine-grained vision-language alignment.

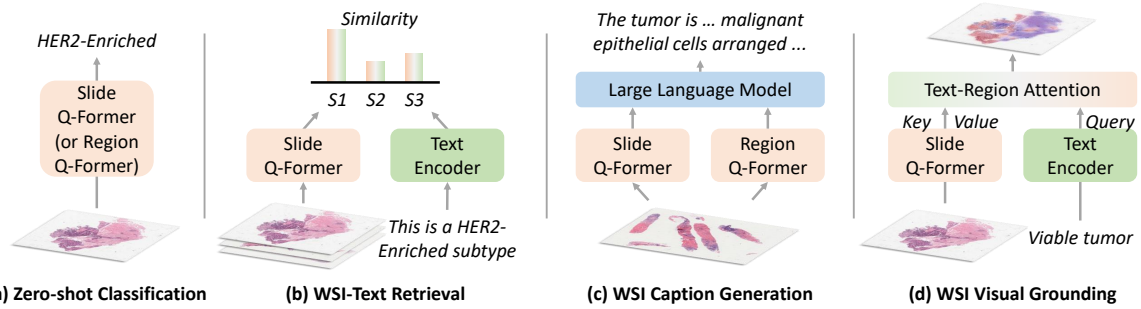

Figure 3: PathFLIP serves as a versatile tool in computational pathology. It accommodates a diverse range of multimodal pathology tasks at both slide and region levels. In (b), "Si" refers to the similarity between the i-th slide and the input text.

Text Representation

The pretraining data used for our PathFLIP follows SlideBench (Chen et al. 2025), which provides slide-level captions containing long and detailed descriptions. These captions reflect the standard clinical workflow of identifying ROIs, conducting diagnostic evaluations, and integrating findings into a slide-level conclusion. To align the fine-grained textual descriptions with the regional visual features set $\mathbf{v}^{\text{reg}}$ extracted, we construct a corresponding set of fine-grained textual representations. Specifically, each slide-level caption is first segmented into individual sentences. Then, we construct K subtext segments $\{T_k^i\}_{k=1}^K$ by randomly sampling subsets of sentences from the original caption T^i using a probabilistic strategy. Each subcaption is subsequently encoded into a fixed-dimensional representation $t_k^i \in \mathbb{R}^d$ using a learnable text encoder $f_t(\cdot,\cdot)$:

$$t_k^i = f_t([\text{CLS}], T_k^i)\ . \tag{3}$$

The complete set of region-level textual features $\{t_k^i\}_{k=1}^K$ is defined as $\mathbf{t}^{\text{sub}} \in \mathbb{R}^{K\times d}$. We preserve the full-text caption and encode it into a global textual representation $t^i \in \mathbb{R}^d$ as:

$$t^i = f_t([\text{CLS}], T^i)\ . \tag{4}$$

During training, subcaptions are randomly sampled for region-level alignment, while the full caption supports global representation learning.

Vision-Langugae Alignment

Text-conditioned Region Feature Alignment To capture fine-grained pathological semantics, it is necessary to establish semantic associations between the regional feature set $\mathbf{v}^{\text{reg}}$ and the subcaption feature set $\mathbf{t}^{\text{sub}}$, thereby enabling region-level alignment between WSI features and text representations. To this end, we employ a cross-attention mechanism, where subcaption features are used as queries to attend to regional visual features. This process is shown as:

$$\mathbf{v}^{\text{tc}} = \text{softmax}\left(\frac{W_q\mathbf{t}^{\text{sub}}(W_k\mathbf{v}^{\text{reg}})^\top}{\sqrt{d}}\right)W_v\mathbf{v}^{\text{reg}}\ , \tag{5}$$

where $\mathbf{v}^{\text{tc}}$ denotes the text-conditioned query results of attended features and W_q, W_k, and W_v are learnable projection matrices.

To enhance region-text alignment, we incorporate subcaption features from other image-text pairs within the same batch as negative queries. For a given image-text pair $\langle S^i, T^i\rangle$, the subcaptions of T^i serve as positive conditions, while those from other texts T^j $(j \neq i)$ act as negatives for contrastive learning. To reduce memory and computational cost, we randomly sample one subcaption from each of the other $B-1$ pairs, rather than using all. Consequently, the total number of text-conditioned queries becomes $L = K + B - 1$, where K is the number of subcaptions for T^i and B is the batch size. We adopt the LogSigmoid loss for regional contrastive learning:

$$\mathcal{L}_{region} = \frac{1}{L}\sum_{l=1}^{L} -\log\left(\frac{1}{1+\exp(y_l\cdot -\eta < \mathbf{v}_l^{\text{tc}}, \mathbf{t}_l^{\text{sub}} >)}\right)\ , \tag{6}$$

where $y_l = +1$ for positive pairs and -1 for negative ones, and η is a learnable temperature parameter. The similarity $< \mathbf{v}_l^{\text{tc}}, \mathbf{t}_l^{\text{sub}} >$ between the text-conditioned visual query result $\mathbf{v}_l^{\text{tc}}$ and the textual query $\mathbf{t}_l^{\text{sub}}$ is computed via dot product. Final region-level alignment is obtained by mean pooling over the L query results. This self-supervised formulation encourages alignment of semantically related regions and subcaptions, which we find essential for fine-grained visual-text grounding.

Global Feature Alignment We adopt a contrastive learning strategy to align global-level semantics of vision and language modalities. Within each training batch, matched image-text pairs are treated as positive samples, while mismatched pairs are treated as negatives. We compute the cosine similarity between global image embeddings v^i and text representations t^i. We average the N_q query features in v^i to obtain $\bar{v}^i$ for image-text similarity computation. The contrastive loss is formulated as:

$$\mathcal{L}_{i2t} = -\log\left(\frac{\exp(<\bar{v}^i, t^i>/\tau)}{\Sigma_{b=1}^{B}\exp(<\bar{v}^i, t^b>/\tau)}\right)\ , \tag{7}$$

$$\mathcal{L}_{t2i} = -\log\left(\frac{\exp(<\bar{v}^i, t^i>/\tau)}{\Sigma_{b=1}^{B}\exp(<\bar{v}^b, t^i>/\tau)}\right)\ , \tag{8}$$

where B is the batch size, $<\cdot,\cdot>$ represents the cosine similarity calculation and τ is the temperature hyperparameter.

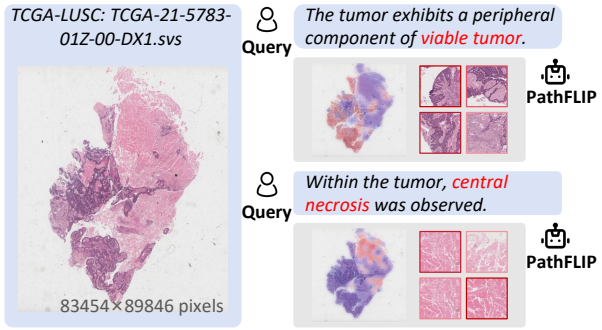

Figure 4: Visual grounding results. High-attention areas are highlighted in red in the heatmap, with red boxes marking the corresponding regions.

The total global contrastive loss is:

$$\mathcal{L}_{slide} = \frac{1}{2}(\mathcal{L}_{t2i} + \mathcal{L}_{i2t}) \,. \tag{9}$$

Our final loss of vision-language alignment is the sum of global loss $\mathcal{L}_{slide}$ and regional loss $\mathcal{L}_{region}$:

$$\mathcal{L} = \mathcal{L}_{region} + \mathcal{L}_{slide}. \tag{10}$$

Transfer for Visual Understanding

Supervised Instruction Fine-tuning Building upon the pretrained model, we fine-tune it by integrating with an LLM to enable instruction-following for tasks such as captioning and question answering. Specifically, we concatenate the slide-level embedding v^i with the set of region-level embeddings $\{v_n^i\}_{n=1}^N$ and project them into the LLM's embedding space using a linear transformation $\sigma(\cdot)$. A task-specific prompt embedding t_p (e.g., for caption generation) is prepended to the projected visual features, which are then fed into the LLM for autoregressive text generation. The entire model is optimised using a language modeling objective:

$$\mathcal{L}_{lm} = \sum_{j=1}^{L_t} -\log P_{lm}(\mathcal{T}_j^i \mid \mathcal{T}_{<j}^i, \sigma([v^i, v_1^i, \ldots, v_N^i]), t_p), \tag{11}$$

where L_t denotes the length of the generated sequence, P_{lm} is the token-level conditional distribution predicted by the LLM, $\mathcal{T}_j^i$ indicates the j-th ground truth token, and $\mathcal{T}_{<j}^i$ represents the preceding tokens in the sequence. The model acquires enhanced multimodal comprehension and effective instruction-following abilities by integrating visual features with the LLM in this unified framework.

Generalizable Visual-Language Understanding We project pathology images and textual descriptions into a shared embedding space, where semantically aligned image-text pairs are pulled closer via contrastive learning. For an input slide-caption pair $\langle S^i, T^j \rangle$, we extract their embeddings v^i and t^j, and compute the similarity score. Moreover, our Region Q-Former module outputs region-level embeddings v_n^i, enabling fine-grained visual-language understanding.

Visual Grounding Our method aligns with the diagnostic workflow of pathologists by integrating regional observations into coherent narratives to support downstream interpretation and visual grounding. For each slide, region-level representations are extracted using the Region Q-Former. A text-region attention layer is then applied to compute attention scores between each region and its corresponding subcaption, enabling precise region-level alignment and grounding. As shown in Figure 4, we visualize the attention maps corresponding to two subcaptions from a given slide. PathFLIP successfully localizes regions of interest such as "viable tumor" and "central necrosis." These results demonstrate that our framework enables fine-grained visual grounding guided by natural language prompts, facilitating transferable reasoning and clinical localization. Notably, the grounding ability is unique to PathFLIP and is not available in most previous approaches. The detailed analyses of the key feature are in supplementary materials.

Experiments

Experimental Setup

Datasets We use the SlideInstruction dataset (Chen et al. 2025) for pretraining, which includes WSIs from the public TCGA database (Tomczak, Czerwińska, and Wiznerowicz 2015), comprising 4,915 WSI-caption pairs from 4,028 patients. Each WSI is associated with a long and detailed, clinically meaningful caption. To validate the PathFLIP performance for various tasks, we use the SlideBench dataset, which includes the remaining test set from TCGA and external datasets from BCNB (Xu et al. 2021). For zero-shot classification, we additionally use CPTAC (Edwards et al. 2015), which contains genetic mutation information. We also employ the large-scale Quilt-1M (Ikezogwo et al. 2023) as an external benchmark for the retrieval task.

Comparison Methods We compare PathFLIP with state-of-the-art methods for classification and retrieval tasks. These methods can be categorized into two groups. The first group includes models without LLMs, which are typically used as foundation models (FMs) to extract task-agnostic features. Representative examples include CONCH (Lu et al. 2024), PLIP (Huang et al. 2023), and Prov-GigaPath (Xu et al. 2024). The second group consists of VLMs designed for pathology and enhanced by LLMs. These models are often capable of following natural language instructions, enabling more flexible downstream applications. Representative methods include Quilt-LLaVA (Seyfioglu et al. 2024), LLaVA-Med (Li et al. 2023a), and PathGen-LLava (Sun et al. 2025b). Notably, these LLM-based methods are primarily developed for patch-level pathology images. In particular, Cpath-Omni (Sun et al. 2025a) and SlideChat (Chen et al. 2025) incorporate dedicated slide encoders to generate a global slide-level representation as inputs to the LLMs.

Implementation Details We implement PathFLIP in PyTorch 2.7.0 and conduct experiments on 8 NVIDIA RTX 4090 GPUs. Each WSI is partitioned into non-overlapping 4096×4096 regions at $20\times$ magnification and further cropped into 256×256 patches for feature extraction us-

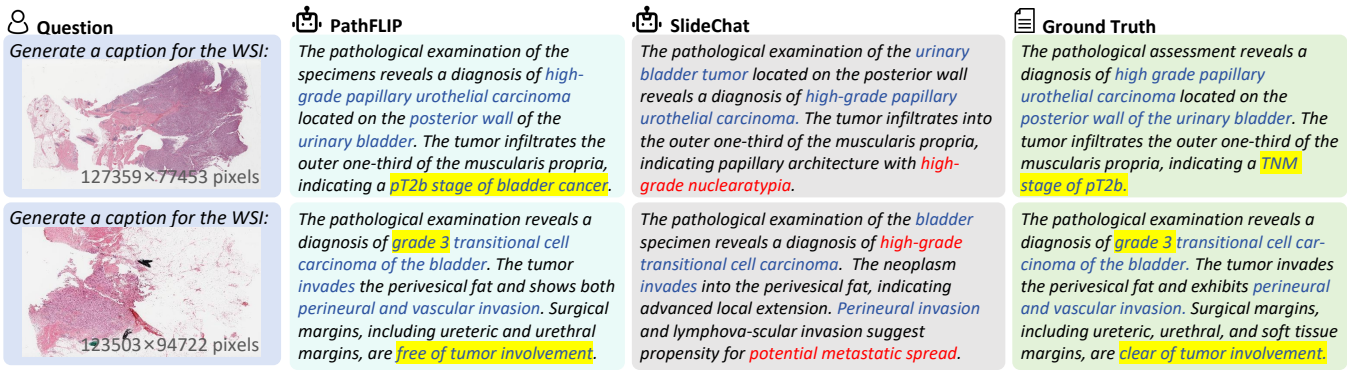

Figure 5: Caption generation comparison. Blue indicates correct matches, red indicates incorrect or imprecise matches, and yellow backgrounds emphasize important information matches.

AUC	BRCA			COAD			LUAD			LSCC			Average
Model	*PIK3CA*	*MAP3K1*	*GATA3*	*KRAS*	*PIK3CA*	*TP53*	*TP53*	*STK11*	*KRAS*	*TP53*	*PIK3R1*	*KEAP1*	
MI-Zero (Lu et al. 2023)	0.5321	0.5542	0.5468	0.5022	0.5438	0.5370	0.5470	0.5847	0.4979	0.6171	0.5910	0.5252	0.5483
PLIP (Huang et al. 2023)	0.4637	0.5818	0.4454	0.4809	0.5455	0.5078	0.5384	0.6427	0.5252	**0.7298**	0.5820	0.5676	0.5509
CONCH (Lu et al. 2024)	0.5349	0.6380	0.5638	0.5104	0.6384	0.5893	0.6647	0.4935	0.5862	0.5608	0.6101	0.5680	0.5798
Prov-GigaPath (Xu et al. 2024)	0.5633	0.6235	0.5544	0.5615	0.6184	0.6249	0.6403	0.6282	0.5092	0.7072	0.5943	0.6252	0.6042
PathGen (Sun et al. 2025b)	0.5184	0.4904	0.4750	0.4949	0.5807	0.5695	0.5708	0.4825	0.5197	0.4731	0.5099	0.4898	0.5146
PathAlign (Ahmed et al. 2024)	0.5875	0.6073	**0.6865**	0.6096	0.6276	0.6534	**0.6693**	0.6143	0.5871	0.7208	0.6301	0.6427	0.6364
LLaVA-Med (Li et al. 2023a)	0.5755	0.5770	0.5598	0.5690	0.5802	0.5748	0.5657	0.5711	0.5563	0.5875	0.5678	0.5694	0.5712
MedDr (He et al. 2024)	0.5433	0.5947	0.5245	0.5371	0.5924	0.6075	0.6096	0.6349	0.5817	0.7131	0.6240	0.5927	0.5963
Quilt-LLaVA (Seyfioglu et al. 2024)	0.6122	0.6107	0.5981	0.5945	0.6077	0.6018	0.5962	0.6079	0.5896	0.6143	0.6002	0.6019	0.6029
PathGen-LLava (Sun et al. 2025b)	0.5860	**0.6484**	0.6252	0.6351	0.6427	0.6380	0.5482	0.6332	0.6247	0.6466	0.5699	0.6334	0.6193
CPath-Omni (Sun et al. 2025a)	0.6423	0.6301	0.6355	0.6430	**0.6623**	0.6671	0.6520	0.6418	0.6502	0.6389	0.6309	0.6151	0.6424
PathFLIP (w/o Region Q-Former)	0.5152	0.6043	0.5432	0.5492	0.5651	0.5443	0.5034	0.5641	0.5034	0.5736	0.5066	0.5976	0.5475
PathFLIP (w/o Slide Q-Former)	0.5863	0.5450	0.5660	0.5467	0.5643	0.5486	0.5357	0.5698	0.5691	0.5721	0.5927	0.5279	0.5604
PathFLIP (w/o T-R Atten.)	0.5367	0.5921	0.5541	0.5025	0.5132	0.5226	0.5585	0.5229	0.5595	0.5898	0.5770	0.5036	0.5444
PathFLIP (Ours)	**0.6575**	0.6339	0.6796	**0.6720**	0.6585	**0.6713**	0.6452	**0.6867**	**0.6646**	0.6209	**0.6423**	**0.7287**	**0.6634**

Table 1: Comparison of models on gene mutation prediction tasks across multiple cancer types in the CPTAC dataset. Bold indicates the best performance and underline indicates the second best performance. "T-R Atten." refers to the Text-Region Attention layer.

ing the pretrained CONCH (Lu et al. 2024). We set the Q-Former query token count to $N_q = 8$ and sample $K = 8$ subcaptions for regional alignment. The temperature hyper-parameter is set to $\tau = 0.1$, while the learnable parameter is initialized as $\eta = \log(1/0.07)$. The language module uses the lightweight Qwen3-0.6B (Yang et al. 2025), fine-tuned via LoRA (Hu et al. 2022).

Comparisons with Previous Studies

Zero-shot Classification We evaluate the generalisation capability of PathFLIP through zero-shot biomarker classification tasks on the CPTAC dataset, formulated as prompt-based tasks. Each task is formulated as a prompt-based classification problem, such as: *"a histopathology image with a PIK3CA gene mutation."* The first group of baselines excludes LLMs, using averaged patch embeddings for slide-level representations and classifying based on similarity with text prompts. The second group integrates LLMs but cannot directly handle gigapixel WSIs, so 30 patches are randomly sampled per slide and used with prompts for biomarker prediction. As shown in Table 1, PathFLIP achieves the highest average score 0.6634, outperforming CPath-Omni and PathAlign by 2.1% and 2.7%, respectively, while using significantly fewer model parameters and less training data. Further, when incorporating the Slide Q-Former, Region Q-Former, and text-region attention modules, the model achieves performance improvements of 11.59%, 10.3%, and 11.9%, respectively, demonstrating the effectiveness of fine-grained alignment. These results highlight the discriminative power of the slide-level representations at fine-grained and global levels.

Zero-shot Retrieval We evaluate the multimodal alignment capability through the slide-text retrieval tasks on the SlideBench dataset and a region-level Quilt dataset, as summarised in Table 2. On SlideBench, our slide-level embeddings significantly outperform the mean patch embeddings of FMs, achieving an average improvement of 15.50% on the ITR task and 13.04% on the TIR task over CONCH. We conduct region-level image-text retrieval tasks on the Quilt dataset to further validate the model's generative retrieval capability. PathFLIP achieves SOTA performance with 0.6611 and 0.6407 R@10 scores, respectively. Compared to PathGen, which achieves R@1 scores of 0.3246 and 0.2688, our method ranks second on other evaluation met-

Model	SlideBench						Quilt					
	Image-Text Retrieval (ITR)			Text-Image Retrieval (TIR)			Image-Text Retrieval (ITR)			Text-Image Retrieval (TIR)		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
MI-Zero (Lu et al. 2023)	0.0468	0.2120	0.2723	0.0491	0.2235	0.3725	0.1583	0.3316	0.4711	0.1516	0.3543	0.5036
PLIP (Huang et al. 2023)	0.0468	0.2163	0.3725	0.0336	0.2053	0.3341	0.1296	0.2813	0.4208	0.1152	0.2751	0.4092
CONCH (Lu et al. 2024)	0.0360	0.2276	0.3798	0.0576	0.2788	0.4399	0.2411	0.4612	0.6295	0.2377	0.4174	0.6351
Prov-GigaPath (Xu et al. 2024)	0.0468	0.2259	0.3822	0.0528	0.2235	0.3725	0.0123	0.1516	0.3070	0.0253	0.1482	0.3184
PathGen (Sun et al. 2025b)	0.0513	0.2165	0.3629	0.0535	0.2139	0.3581	**0.3246**	**0.4955**	0.6206	**0.2688**	**0.4645**	0.6043
PathAlign (Ahmed et al. 2024)	0.0336	0.2299	0.3341	0.0432	0.2139	0.3629	0.1813	0.3906	0.4169	0.1533	0.3910	0.5047
PathFLIP (w/o Region Q-Former)	0.0581	0.2103	0.3613	0.0481	0.2099	0.3760	0.1795	0.3627	0.5654	0.1829	0.3567	0.5611
PathFLIP (w/o Slide Q-Former)	0.1127	0.3639	0.5461	0.0811	0.3357	0.5504	0.1757	0.3333	0.6165	0.1719	0.3253	0.6293
PathFLIP (w/o T-R Atten.)	0.0781	0.2810	0.4823	0.0861	0.3249	0.4979	0.1774	0.3278	0.6301	0.1668	0.3257	0.6284
PathFLIP (Ours)	**0.1513**	**0.3869**	**0.5704**	**0.1392**	**0.4380**	**0.5903**	0.2710	0.4708	**0.6611**	0.2646	0.4291	**0.6407**

Table 2: Retrieval performance on the SlideBench and Quilt datasets for Image-Text Retrieval (ITR) and Text-Image Retrieval (TIR) tasks. The evaluation metric is Recall@K. Bold indicates the best performance in each column, and underline denotes the second best. All evaluations are conducted with a batch size of 64 from the dataset.

Methods	BLEU-1	BLEU-2	BLEU-3	Rouge-L	Qwen.Em. Similarity
LLava-Med† (Li et al. 2023a)	0.18	0.12	0.06	0.12	0.58
MedDr† (He et al. 2024)	0.31	0.16	0.06	0.07	0.68
Quilt-LLava† (Seyfioglu et al. 2024)	0.23	0.09	0.04	0.16	0.56
PathGen-LLava† (Sun et al. 2025b)	0.35	0.17	0.09	0.16	0.65
CPath-Omni (Sun et al. 2025a)	0.33	0.19	0.10	0.17	0.70
SlideChat (Chen et al. 2025)	0.37	0.21	0.12	0.24	0.71
PathFLIP (w/o Region Q-Former)	0.31	0.19	0.12	0.33	0.64
PathFLIP (w/o Slide Q-Former)	0.21	0.13	0.08	0.32	0.67
PathFLIP (w/o T-R Atten.)	0.32	0.19	0.12	**0.34**	0.64
PathFLIP (w/o finetune Qwen3)	0.25	0.14	0.09	0.33	0.64
PathFLIP (w/ finetune Qwen3)	**0.38**	**0.23**	**0.13**	**0.34**	**0.76**

Table 3: WSI caption performance across different methods on SlideBench. Due to the inability to process the entire WSI simultaneously, methods marked with † use 30 randomly sampled patches from each slide as input. Bold indicates the best and underline denotes the second best.

rics while being trained on a significantly smaller dataset, demonstrating the strong potential of our proposed PathFLIP. The Region Q-Former module is crucial in retrieval performance, boosting average recall by 12.90% for ITR task and 12.79% for TIR task. Among all comparison methods, PathFLIP shows superior capability in capturing fine-grained visual-language correlations, highlighting its effectiveness and efficiency, even with limited data resources.

Methods	SlideBench			BCNB VQA	Average
	Microscopy	Diagnosis	Clinical		
LLava-Med† (Li et al. 2023a)	0.4734	0.3278	0.4796	0.3010	0.3955
MedDr† (He et al. 2024)	0.7330	0.5778	0.7425	0.3367	0.5975
Quilt-LLava† (Seyfioglu et al. 2024)	0.5776	0.3596	0.5307	0.3219	0.4475
PathGen-LLava† (Sun et al. 2025b)	0.6867	0.4969	0.6336	0.4557	0.5682
CPath-Omni (Sun et al. 2025a)	0.6370	0.5241	0.5926	0.4567	0.5526
SlideChat (Chen et al. 2025)	**0.8764**	0.7327	0.8426	0.5414	0.7483
Ours (w/ finetune Qwen3)	0.8611	**0.7628**	**0.8941**	**0.5865**	**0.7761**

Table 4: VQA performance of different methods on the SlideBench and BCNB. Methods marked with † use 30 randomly sampled patches from each slide. Bold indicates the best performance and underline denotes the second best.

WSI Caption and Visual Question Answering As shown in Table 3 and Table 4, we evaluate caption generation performance using BLEU, ROUGE-L, and Qwen3-Embedding-0.6B (Zhang et al. 2025) for similarity scoring. We observe that Quilt-LLava and PathGen-LLava, which rely on patch-level inputs, perform worse than models capable of holistically processing the entire slide. Notably, our PathFLIP significantly outperforms SlideChat, achieving a 10% improvement in Rouge-L and a 5% gain in Qwen embedding similarity in the captioning task. For the VQA task, PathFLIP also outperforms SlideChat with average improvements of 2.78%. Specifically, on the external BCNB dataset, PathFLIP achieves a strong score of 0.58. These results demonstrate that our model can efficiently generate descriptive captions for WSIs and effectively interpret complex instructions. Furthermore, when applying instruction fine-tuning on a lightweight model like Qwen3-0.6B, we observe a 12% improvement in similarity scores for captioning. These findings underscore the model's strong multimodal understanding and instruction-following capabilities for gigapixel-scale WSIs. They also highlight the practical potential of our pretraining approach when integrated with lightweight LLMs in real-world clinical applications. Additional experimental results are in supplementary materials.

Conclusion

We propose PathFLIP, a fine-grained language-image pretraining framework for general-purpose pathological image analysis. By aligning region-level visual features with subcaption semantics via contrastive learning, PathFLIP achieves accurate and interpretable grounding without manual annotations. Integrated with large language models, PathFLIP supports diverse multimodal tasks such as WSI captioning and VQA, and consistently outperforms existing pathology VLMs on multiple benchmarks. Beyond technical advancements, PathFLIP facilitates interpretable AI-assisted diagnosis and clinician-AI communication, paving the way for autonomous AI agents in digital pathology to enhance clinical decision-making and healthcare.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under 62031023 & 62331011; and in part by the Shenzhen Science and Technology Project under GXWD20220818170353009.

References

Ahmed, F.; Sellergren, A.; Yang, L.; Xu, S.; Babenko, B.; Ward, A.; Olson, N.; Mohtashamian, A.; Matias, Y.; Corrado, G. S.; et al. 2024. Pathalign: A vision-language model for whole slide images in histopathology. *arXiv preprint arXiv:2406.19578*.

Cai, L.; Huang, S.; Zhang, Y.; Lu, J.; and Zhang, Y. 2025. AttriMIL: Revisiting attention-based multiple instance learning for whole-slide pathological image classification from a perspective of instance attributes. *Medical Image Analysis*, 103631.

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

Chen, R. J.; Ding, T.; Lu, M. Y.; Williamson, D. F.; Jaume, G.; Song, A. H.; Chen, B.; Zhang, A.; Shao, D.; Shaban, M.; et al. 2024. Towards a general-purpose foundation model for computational pathology. *Nature medicine*, 30(3): 850–862.

Chen, Y.; Wang, G.; Ji, Y.; Li, Y.; Ye, J.; Li, T.; Hu, M.; Yu, R.; Qiao, Y.; and He, J. 2025. Slidechat: A large vision-language assistant for whole-slide pathology image understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 5134–5143.

Ding, T.; Wagner, S. J.; Song, A. H.; Chen, R. J.; Lu, M. Y.; Zhang, A.; Vaidya, A. J.; Jaume, G.; Shaban, M.; Kim, A.; et al. 2024. Multimodal whole slide foundation model for pathology. *arXiv preprint arXiv:2411.19666*.

Dong, J.; Jiang, J.; Jiang, K.; Li, J.; Cai, L.; and Zhang, Y. 2025. Disentangled Pseudo-bag Augmentation for Whole Slide Image Multiple Instance Learning. *IEEE Transactions on Medical Imaging*.

Edwards, N. J.; Oberti, M.; Thangudu, R. R.; Cai, S.; McGarvey, P. B.; Jacob, S.; Madhavan, S.; and Ketchum, K. A. 2015. The CPTAC data portal: a resource for cancer proteomics research. *Journal of proteome research*, 14(6): 2707–2713.

Ghezloo, F.; Seyfioglu, M. S.; Soraki, R.; Ikezogwo, W. O.; Li, B.; Vivekanandan, T.; Elmore, J. G.; Krishna, R.; and Shapiro, L. 2025. Pathfinder: A multi-modal multi-agent system for medical diagnostic decision-making applied to histopathology. *arXiv preprint arXiv:2502.08916*.

Guo, Z.; Xiong, C.; Ma, J.; Sun, Q.; Feng, L.; Wang, J.; and Chen, H. 2025. Focus: Knowledge-enhanced adaptive visual compression for few-shot whole slide image classification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 15590–15600.

He, S.; Nie, Y.; Chen, Z.; Cai, Z.; Wang, H.; Yang, S.; and Chen, H. 2024. Meddr: Diagnosis-guided bootstrapping for large-scale medical vision-language learning. *CoRR*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Huang, Z.; Bianchi, F.; Yuksekgonul, M.; Montine, T. J.; and Zou, J. 2023. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9): 2307–2316.

Ikezogwo, W.; Seyfioglu, S.; Ghezloo, F.; Geva, D.; Sheikh Mohammed, F.; Anand, P. K.; Krishna, R.; and Shapiro, L. 2023. Quilt-1m: One million image-text pairs for histopathology. *Advances in neural information processing systems*, 36: 37995–38017.

Jiang, S.; Cai, L.; Gan, Z.; Wang, Y.; Tang, G.; and Zhang, Y. 2025. Uncertainty-Aware Survival Analysis with Dirichlet Distribution for Multi-Scale Pathology and Genomics. *IEEE Transactions on Medical Imaging*.

Jiang, S.; Gan, Z.; Cai, L.; Wang, Y.; and Zhang, Y. 2024. Multimodal cross-task interaction for survival analysis in whole slide pathological images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 329–339. Springer.

Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; and Gao, J. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36: 28541–28564.

Li, D.; Wan, G.; Wu, X.; Wu, X.; Nirmal, A. J.; Lian, C. G.; Sorger, P. K.; Semenov, Y. R.; and Zhao, C. 2025. A Survey on Computational Pathology Foundation Models: Datasets, Adaptation Strategies, and Evaluation Tasks. *arXiv preprint arXiv:2501.15724*.

Li, H.; Chen, Y.; Chen, Y.; Yu, R.; Yang, W.; Wang, L.; Ding, B.; and Han, Y. 2024. Generalizable whole slide image classification with fine-grained visual-semantic interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11398–11407.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Liu, F.; Cai, L.; Wang, Z.; Fan, Z.; Yu, J.-g.; Chen, H.; and Zhang, Y. 2025. Prototype-Guided Cross-Modal Knowledge Enhancement for Adaptive Survival Prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 522–532. Springer.

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

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

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.;

et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Seyfioglu, M. S.; Ikezogwo, W. O.; Ghezloo, F.; Krishna, R.; and Shapiro, L. 2024. Quilt-llava: Visual instruction tuning by extracting localized narratives from open-source histopathology videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13183–13192.

Shaikovski, G.; Vorontsov, E.; Casson, A.; Viret, J.; Zimmermann, E.; Tenenholtz, N.; Wang, Y. K.; Bernhard, J. H.; Godrich, R. A.; Retamero, J. A.; et al. 2025. PRISM2: Unlocking Multi-Modal General Pathology AI with Clinical Dialogue. *arXiv preprint arXiv:2506.13063*.

Shi, J.; Li, C.; Gong, T.; Zheng, Y.; and Fu, H. 2024. Vilamil: Dual-scale vision-language multiple instance learning for whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11248–11258.

Sun, Y.; Si, Y.; Zhu, C.; Gong, X.; Zhang, K.; Chen, P.; Zhang, Y.; Shui, Z.; Lin, T.; and Yang, L. 2025a. Cpath-omni: A unified multimodal foundation model for patch and whole slide image analysis in computational pathology. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 10360–10371.

Sun, Y.; Zhang, Y.; Si, Y.; Zhu, C.; Zhang, K.; Shui, Z.; Li, J.; Gong, X.; LYU, X.; Lin, T.; et al. 2025b. PathGen-1.6 M: 1.6 Million Pathology Image-text Pairs Generation through Multi-agent Collaboration. In *The Thirteenth International Conference on Learning Representations*.

Tang, G.; Jiang, S.; Lu, J.; Cai, L.; and Zhang, Y. 2025. IPG-Phormer: Interpretable Pathology Graph-Transformer for Survival Analysis. *arXiv preprint arXiv:2508.12381*.

Tomczak, K.; Czerwińska, P.; and Wiznerowicz, M. 2015. Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia*, 2015(1): 68–77.

Wang, X.; Zhao, J.; Marostica, E.; Yuan, W.; Jin, J.; Zhang, J.; Li, R.; Tang, H.; Wang, K.; Li, Y.; et al. 2024. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature*, 634(8035): 970–978.

Xiao, R.; Kim, S.; Georgescu, M.-I.; Akata, Z.; and Alaniz, S. 2025. Flair: Vlm with fine-grained language-informed image representations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 24884–24894.

Xu, F.; Zhu, C.; Tang, W.; Wang, Y.; Zhang, Y.; Li, J.; Jiang, H.; Shi, Z.; Liu, J.; and Jin, M. 2021. Predicting axillary lymph node metastasis in early breast cancer using deep learning on primary tumor biopsy slides. *Frontiers in oncology*, 11: 759007.

Xu, H.; Usuyama, N.; Bagga, J.; Zhang, S.; Rao, R.; Naumann, T.; Wong, C.; Gero, Z.; González, J.; Gu, Y.; et al. 2024. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015): 181–188.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yuan, Y.; Liu, S.; Zhang, J.; Zhang, Y.; Dong, C.; and Lin, L. 2018. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 701–710.

Zhang, Y.; Li, M.; Long, D.; Zhang, X.; Lin, H.; Yang, B.; Xie, P.; Yang, A.; Liu, D.; Lin, J.; et al. 2025. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. *arXiv preprint arXiv:2506.05176*.

Zheng, K.; Zhang, Y.; Wu, W.; Lu, F.; Ma, S.; Jin, X.; Chen, W.; and Shen, Y. 2024. Dreamlip: Language-image pre-training with long captions. In *European Conference on Computer Vision*, 73–90. Springer.