Libra-MIL: Multimodal Prototypes Stereoscopic Infused with Task-specific Language Priors for Few-shot Whole Slide Image Classification

Zhenfeng Zhuang[1*], **Fangyu Zhou**[1*], **Liansheng Wang**[1,2†]

[1] Department of Computer Science at School of Informatics, Xiamen University, Xiamen, China.
[2] National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen, China.
{zhuangzhenfeng, zhoufangyu}@stu.xmu.edu.cn, lswang@xmu.edu.cn

Abstract

While Large Language Models (LLMs) are emerging as a promising direction in computational pathology, the substantial computational cost of giga-pixel Whole Slide Images (WSIs) necessitates the use of Multi-Instance Learning (MIL) to enable effective modeling. A key challenge is that pathological tasks typically provide only bag-level labels, while instance-level descriptions generated by LLMs often suffer from bias due to a lack of fine-grained medical knowledge. To address this, we propose that constructing task-specific pathological entity prototypes is crucial for learning generalizable features and enhancing model interpretability. Furthermore, existing vision-language MIL methods often employ unidirectional guidance, limiting cross-modal synergy. In this paper, we introduce a novel approach, Multimodal Prototype-based Multi-Instance Learning, that promotes bidirectional interaction through a balanced information compression scheme. Specifically, we leverage a frozen LLM to generate task-specific pathological entity descriptions, which are learned as text prototypes. Concurrently, the vision branch learns instance-level prototypes to mitigate the model's reliance on redundant data. For the fusion stage, we employ the Stereoscopic Optimal Transport (SOT) algorithm, which is based on a similarity metric, thereby facilitating broader semantic alignment in a higher-dimensional space. We conduct few-shot classification and explainability experiments on three distinct cancer datasets, and the results demonstrate the superior generalization capabilities of our proposed method.

Code, Appendix — https://github.com/zfy07/Libra-MIL

Introduction

AI-based histopathology is a promising direction for assisting diagnostics, especially in cancer diagnosis and grading (Omar et al. 2024). H&E-stained WSIs, despite compression, have extremely high resolution (e.g., 40× magnification, 0.25μm/pixel, approximately 40,000 × 40,000 pixels) and large storage needs, challenging end-to-end training on fine-grained pixel data. Multiple instance learning (MIL), a weakly-supervised framework, addresses sparse instance-level labels (typically only WSI-level) and resource consumption in this area (Gadermayr and Tschuchnig 2024). In MIL paradigm, WSIs are treated as bags and their constituent patches as instances. Typically, pretrained feature extractors (e.g., UNI (Chen et al. 2024a), CONCH (Lu et al. 2024)) are employed to embed these patches, and the resulting features are then aggregated to obtain bag-level embeddings for WSI-level prediction tasks. Although MIL-based approaches have achieved strong performance, their diagnostic capabilities have shown signs of saturation. To overcome this limitation, recent studies have explored the integration of multimodal information—such as large language models (LLMs) or domain-specific textual knowledge—to guide more precise vision-language multiple instance learning (VLMIL) (Shi et al. 2024). In the pathology domain, such methods have been increasingly proposed and have demonstrated promising results across various diagnostic tasks by leveraging LLMs.

Despite recent progress in integrating multimodal guidance through LLMs, the scarcity of annotated pathological data remains a fundamental challenge. In this context, few-shot learning has emerged as a promising approach to enhance model generalization under limited supervision. By incorporating strategies such as prototype learning, meta-learning, or transfer learning, few-shot learning can uncover latent feature patterns and improve the model's ability to recognize novel categories, thereby alleviating the performance bottleneck caused by annotation constraints (Qu et al. 2024).

As illustrated in Fig. 1(a-c), previous approaches have not integrated multimodal prototypes and typically rely on text-guided mechanisms to assist the learning of visual features. While these models have demonstrated strong classification performance across various tasks, several limitations remain. First, instance-level descriptions generated by LLMs often suffer from inaccuracies due to the lack of fine-grained medical knowledge. Manually crafted prompts that use only class names lack pathological prior knowledge and offer limited discriminative guidance, particularly for ambiguous categories. Second, existing VLMIL frameworks predominantly adopt unidirectional guidance (e.g., text-to-image cross-modal attention or similarity learning), which restricts the potential for synergistic, bidirectional interactions (Huang et al. 2023; Li et al. 2023). This limits effective information exchange, weakens collaborative reasoning, and results in shallow modality alignment. Consequently, these models fail to fully exploit the mutual enhancement

[*] These authors contributed equally.
[†] Corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

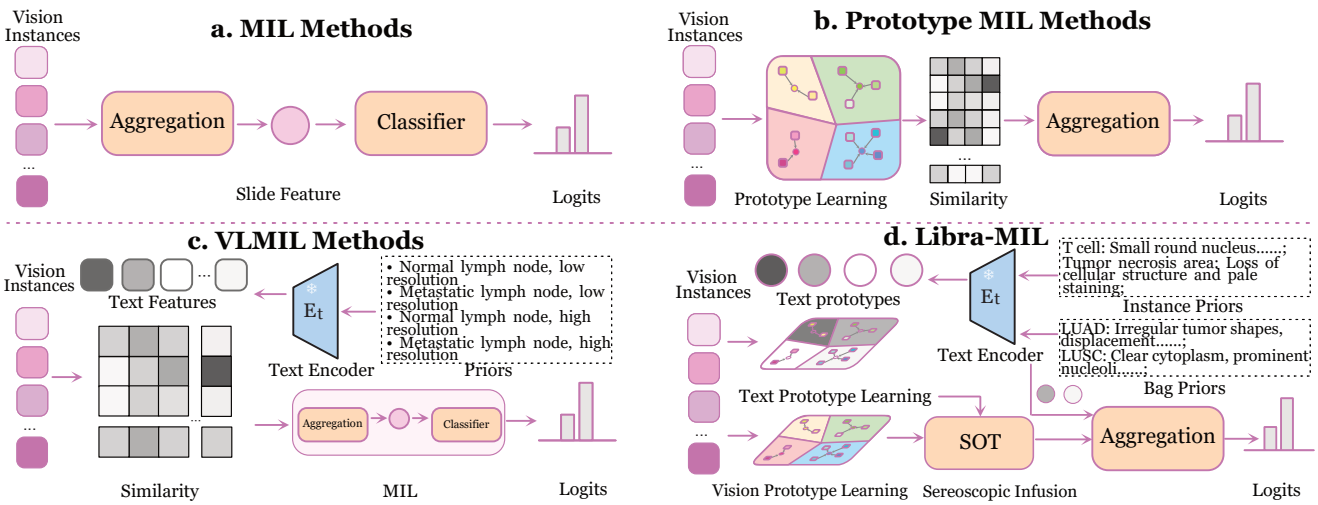

Figure 1: Brief comparison with related MIL methods. a) Basic MIL simply performs aggregation and sorting operations. b) Prototype MIL learns class-aware logits via instance similarity directly. c) VLMIL fused LLM Priors with the text-guided similarity. d) Libra-MIL uses Multimodal prototype learning with Stereoscopic Optimal Transport (SOT) on similarities.

potential of visual and textual representations. Third, current prototype-based learning methods typically rely solely on image-text similarity, neglecting the construction and integration of a unified, scale-invariant multimodal similarity space. Using query similarity as the final decision, without accounting for the deficiencies in the knowledge and embedding accuracy of the textual branch, results in suboptimal alignment and fusion.

To address the mentioned limitations, we propose Libra-MIL, with the following contributions:

- We first introduce a Task Specific approach for Textual Priors Generation. Leveraging LLMs, we produce priors at both the bag level (e.g., tumors often exhibit expansive growth) and the instance level (e.g., nuclear division, inflammatory cell infiltration), tailored to specific tasks. This design is inspired by the diagnostic reasoning of pathologists, who integrate the global and local context. By incorporating these textual prompts with visual information, the model is guided to focus on morphological features that are more aligned with the task-specific diagnostic objectives.
- Dual-Prototype Multimodal Learner enhances generalization in few-shot learning tasks through prototype-based learning. The modality-specific similarities are then stereoscopically infused within a unified embedding space under an optimal transport framework. This fusion minimizes the matching cost between similarities, enabling structure-aware alignment and integration, thereby improving cross-modal semantic consistency and interoperability.
- Libra-MIL consistently outperforms baselines across various few-shot learning settings on three pathological datasets. It surpasses state-of-the-art (SOTA) approaches by an average margin of 2.43% across accuracy, F1-score, and AUC. Extensive ablation studies further demonstrate the effectiveness of each proposed module, and the framework also offers prototype-based interpretability to enhance model transparency.

Related Work

Vision-Language Multiple Instance Learning

Vision-Language Multiple Instance Learning (VLMIL) builds upon traditional computational pathology frameworks. Early MIL approaches include ABMIL (Ilse, Tomczak, and Welling 2018), which integrated attention mechanisms; CLAM (Lu et al. 2021), enhancing data efficiency through instance clustering and attention pooling; TransMIL (Shao et al. 2021), modeling inter-patch dependencies via mutual attention; DSMIL (Li, Li, and Eliceiri 2021), fusing multi-scale WSI features with pyramid fusion; H2MIL (Hou et al. 2022), incorporating multi-resolution information through heterogeneous graph learning; and DTFD-MIL (Zhang et al. 2022), employing a pseudo-bag strategy for representation enhancement. Recently, VL frameworks have advanced beyond traditional MIL by integrating vision-language models like CLIP (Radford et al. 2021) and LLMs. TOP-MIL (Qu et al. 2023) leverages CLIP features and GPT-4 linguistic priors for few-shot WSI classification. ViLa-MIL (Shi et al. 2024) fuses vision-language information via prototype-guided decoding and dual-scale text prompts. FOCUS (Guo et al. 2025) combines pathological features with language priors through progressive three-stage compression, while Concept-MIL (Sun et al. 2025) predicts pathological concepts to represent key instances through concept-activated patches.

Prototypical Networks for Few-shot Learning

Few-shot learning addresses overfitting in data-scarce scenarios by "learning to learn," enabling rapid adaptation to novel classes with limited supervision. Prototypical Networks (Snell, Swersky, and Zemel 2017) are a representa-

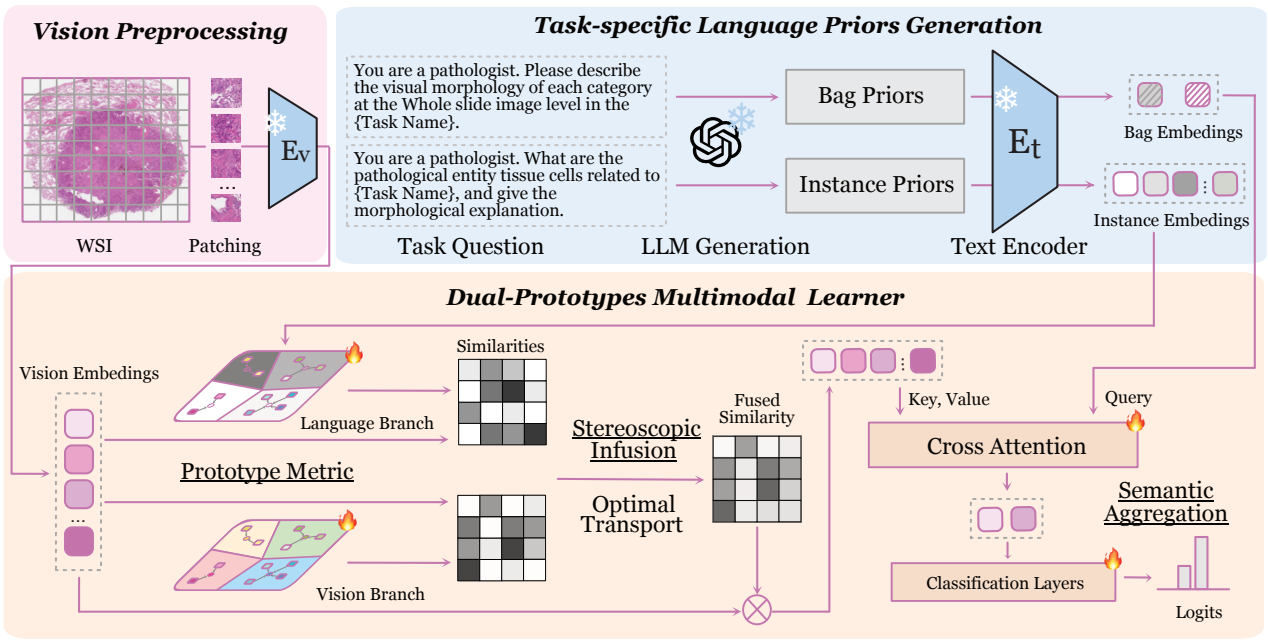

Figure 2: Overview of Libra-MIL. The framework first employs Vision Preprocessing and Task-specific Language Priors Generation modules to perform modality-specific preprocessing and prior embedding. The dual-prototype multimodal learner then integrates instance-level representations from both modalities into a unified similarity space through prototype-based modeling and Sinkhorn optimal transport.

tive approach that learns a metric space where classification is performed based on distances to class prototypes, offering an effective inductive bias for low-data regimes.

Though not explicitly designed for few-shot tasks, several MIL methods have adopted prototype-based mechanisms to improve performance and interpretability. ProtoMIL (Rymarczyk et al. 2022) incorporates prototype features into instance representations. PAMIL (Liu et al. 2024) combines prototype learning with attention for bag-level prediction. TP-MIL (Yang et al. 2023) optimizes patch-level spaces via instance-prototype distances, while QP-MIL (Gou et al. 2025) introduces a vision-language queryable prototype framework for incremental WSI classification. These methods highlight prototype learning's broader relevance beyond the few-shot setting.

Multimodal Fusion in Computational Pathology

While WSIs are central to computational pathology, the complexity of disease demands multimodal fusion with clinical or genomic data for more comprehensive analysis. MCAT (Chen et al. 2021) employs early fusion via co-attention and Set Transformers between image patches and gene embeddings. PathOmics (Ding et al. 2023) aligns image and omics features in a shared latent space via MSE loss. MOTCat (Xu and Chen 2023) uses optimal transport-based co-attention for global alignment. SURVPATH (Jaume et al. 2024) encodes transcriptomic data as pathway-informed tokens and combines them with image patches in a Transformer for survival prediction.

Methodology

We propose Libra-MIL, a few-shot learning framework for pathological images comprising four key components (Fig. 2): Vision Preprocessing (VP) for WSI feature extraction; Task-specific Language Priors Generation (T-LPG) for generating global and entity-level textual cues; a Dual-Prototype Multimodal Learner (DPM) that aligns visual and textual prototypes via Stereoscopic Optimal Transport (SOT); and Semantic Aggregation (SA) for description-guided querying and information fusion.

Preliminary

MIL is a weakly supervised learning paradigm where data are organized into bags $B^{(i)}$ for i^{th} sample, each containing instances $\{\mathbf{x}_j^{(i)}\}_{j=1}^{n^{(i)}}$, while only bag-level labels $y^{(i)}$ (discrete or continuous) are available. The core assumption is that $y^{(i)}$ depends on the collective properties or the presence of key instances within the bag. This paradigm is suitable for scenarios where instance-level annotations are unavailable.

$$f(B^{(i)}) = f(\{x_1^{(i)}, \cdots, x_{n^{(i)}}^{(i)}\}) \approx y^{(i)} \tag{1}$$

$f(\cdot)$ enabling bag-level prediction and implicit instance importance inference. For classification tasks, $y^{(i)} \in \{1, ..c\}$.

In **vision preprocessing**, let a bag $B^{(i)} \in \mathbb{R}^{H\times W\times 3}$ denote a WSI, which is divided into a set of patches $\{a_j^{(i)}\} \in \mathbb{R}^{h\times w\times 3}$ inspired by CLAM (Lu et al. 2021). Due to high computational cost, end-to-end training on pixel-level patches is impractical. A common alternative uses a

pretrained vision encoder E_v to extract patch-level features: $\mathbf{x}_{ij} = E_v(a_j^{(i)})$.

Task-specific Language Priors Generation

To incorporate high-level semantic cues and pathological domain knowledge, we introduce a T-LPG module that constructs instance-level and bag-level textual descriptions through task-specific prompts.

In the context of a classification task T, we design two task-specific prompts based on the **CHAT** framework (Sahoo et al. 2024). One capturing entity-level pathological concepts and the other describing whole-slide morphological patterns. These prompts, denoted by $q \in Q$, are provided as input to a large language model $\text{LLM}(\cdot)$, which simulates a pathologist's diagnostic reasoning to produce language-based priors. Specifically, the instance-level priors encode fine-grained pathological entities, while the bag-level priors reflect the global morphological context of the slide. The process for generating these two types of priors is:

$$P_{\text{ins}} = \text{LLM}(q_{\text{ins}}), P_{\text{bag}} = \text{LLM}(q_{\text{bag}}) \tag{2}$$

where P_{ins} and P_{bag} are further embedded into shared semantic space using a frozen text encoder E_t:

$$z_{\text{ins}} = E_t(P_{\text{ins}}), z_{\text{bag}} = E_t(P_{\text{bag}}) \tag{3}$$

where $z_{\text{ins}} \in \mathbb{R}^{k_t \times d}$, $z_{\text{bag}} \in \mathbb{R}^{c \times d}$ are the resulting instance-level and bag-level language embeddings. k_t is the number of instance priors. d is the hidden size of E_t.

Dual-Prototype Multimodal Learner

To facilitate effective alignment between vision and language modalities under limited supervision in few-shot learning, we propose a two branches in prototype learner that integrates instance-level representations from both branches into a unified similarity space via prototype-based modeling and Stereoscopic Optimal Transport (SOT).

Prototypes metrics To construct a shared semantic space for vision-language alignment, we employ two complementary sets of prototypes: learnable visual prototypes and text prototypes derived from language priors.

Let the set of visual instances bag $B^{(i)}$ be $\mathcal{X}^{(i)} = \{\mathbf{x}_{ij}\}_{j=1}^{n^{(i)}}$, where $\mathbf{x}_{ij} \in \mathbb{R}^d$ elements. We introduce a set of learnable vision prototypes $\mathcal{P}_v$ with K_v elements. For multimodal domain-specific prior knowledge, we also introduce text prototypes derived from instance-level textual descriptions. These K_t language priors describe relevant pathological entities' embeddings:

$$\begin{cases} \mathcal{P}v = \{\mathbf{p}_v^{(k_v)}\}_{k_v=1}^{K_v}, & \mathbf{p}_v^{(k_v)} \in \mathbb{R}^d \\ \mathcal{P}t = \{\mathbf{p}_t^{(k_t)} = z_{\text{ins}}^{(k_t)}\}_{k_t=1}^{K_t}, & \mathbf{p}_t^{(k_t)} \in \mathbb{R}^d \end{cases} \tag{4}$$

Dual modality prototypes are initialized with $\mathcal{P}_v, \mathcal{P}_t$ and jointly optimized with model parameters. They act as semantic anchors to capture task-relevant instance patterns.

The insance-to-prototype similarity $S^{(i)}(j,k)$ is computed modality-wise as follows:

$$\begin{cases} \mathbf{S}_v^{(i)}(j,k) = \cos(\mathbf{x}_{ij}, \mathbf{p}_v^{(k)}) = \dfrac{\mathbf{x}_{ij}^\top \mathbf{p}_v^{(k)}}{\|\mathbf{x}_{ij}\| \cdot \|\mathbf{p}_v^{(k)}\|}, & \mathbf{p}_v^{(k)} \in \mathcal{P}_v \\ \mathbf{S}_t^{(i)}(j,k) = \cos(\mathbf{x}_{ij}, \mathbf{p}_t^{(k)}) = \dfrac{\mathbf{x}_{ij}^\top \mathbf{p}_t^{(k)}}{\|\mathbf{x}_{ij}\| \cdot \|\mathbf{p}_t^{(k)}\|}, & \mathbf{p}_t^{(k)} \in \mathcal{P}_t \end{cases} \tag{5}$$

The bimodal similarity quantifies the semantic relevance of each visual instance across different modalities, reflecting its alignment with task-specific concepts. The model is able to enhance discriminative capability and semantic understanding of individual instances by similarities infusion as follow.

Stereoscopic Infusion Since similarity values are dimensionless and lie in a shared semantic space, we interpret the two similarity matrices—visual-to-visual $\mathbf{S}_v^{(i)} \in \mathbb{R}^{n^{(i)} \times K_v}$ and visual-to-textual $\mathbf{S}_t^{(i)} \in \mathbb{R}^{n^{(i)} \times K_t}$ as two slices in a conceptual three-dimensional latent space. Each visual instance $\mathbf{x}_{ij}$ can be semantically aligned with prototypes from both visual and textual modalities, forming a multimodal correspondence field.

Inspired by Optimal Transport (OT) (Brenier 1991), we propose to iteratively fuse $\mathbf{S}_v^{(i)}$ and $\mathbf{S}_t^{(i)}$ into a unified alignment map by modeling the coupling between instance-level distributions and prototype-level distributions. Specifically, we first define the instance-wise similarity cost matrix $\mathbf{C} \in \mathbb{R}^{K_v \times K_t}$ as the cross-modal discrepancy between vision and text prototypes, computed by:

$$\mathbf{C}(k_v, k_t) = 1 - \cos\left(\mathbf{p}_v^{(k_v)}, \mathbf{p}_t^{(k_t)}\right) \tag{6}$$

Then, we define marginal distributions over the two prototype spaces, Δ^K denoates the K-dimensional probability simplex:

$$\boldsymbol{\mu}^{(i)} \in \Delta^{K_v}, \quad \boldsymbol{\nu}^{(i)} \in \Delta^{K_t} \tag{7}$$

where $\boldsymbol{\mu}^{(i)}$ and $\boldsymbol{\nu}^{(i)}$ are estimated via average attention distributions from $\mathbf{S}_v^{(i)}$ and $\mathbf{S}_t^{(i)}$, respectively:

$$\begin{cases} \boldsymbol{\mu}^{(i)} = \text{softmax}\left(\frac{1}{n^{(i)}} \sum_{j=1}^{n^{(i)}} \mathbf{S}_v^{(i)}(j,:)\right) \\ \boldsymbol{\nu}^{(i)} = \text{softmax}\left(\frac{1}{n^{(i)}} \sum_{j=1}^{n^{(i)}} \mathbf{S}_t^{(i)}(j,:)\right) \end{cases} \tag{8}$$

We then solve the entropy-regularized optimal transport problem to find the transport plan $\mathbf{T}^{(i)} \in \mathbb{R}^{K_v \times K_t}$:

$$\mathbf{T}^{(i)} = \underset{\mathbf{T} \in \Pi(\boldsymbol{\mu}^{(i)}, \boldsymbol{\nu}^{(i)})}{\arg\min} \langle \mathbf{T}, \mathbf{C} \rangle - \varepsilon \mathcal{H}(\mathbf{T}) \tag{9}$$

where $\Pi(\boldsymbol{\mu}^{(i)}, \boldsymbol{\nu}^{(i)})$ denotes the set of transport plans with marginals $\boldsymbol{\mu}^{(i)}$ and $\boldsymbol{\nu}^{(i)}$, $\mathcal{H}(\mathbf{T}) = -\sum_{k_v,k_t} \mathbf{T}_{k_v,k_t} \log \mathbf{T}_{k_v,k_t}$ is the entropy regularization term, and $\varepsilon > 0$ controls the smoothness. To efficiently solve this $\mathbf{T}$, we apply the Sinkhorn algorithm (Cuturi 2013) (Appendix B.1).

Finally, we fused the similarity between instance $\mathbf{x}_{ij}$ and prototype pair (k_v, k_t) is derived by transporting the original similarity maps via the learned plan:

$$\tilde{\mathbf{S}}_j^{(i)} = \mathbf{S}_v^{(i)}(j,:) \mathbf{T}^{(i)} \mathbf{S}_t^{(i)}(j,:)^\top \tag{10}$$

This scalar fusion score $\tilde{\mathbf{S}}_j^{(i)} \in \mathbb{R}$ acts as a unified multimodal relevance indicator for each instance.

Semantic Aggregation This scalar fusion score is used to reweight the visual instances, and a query-based attention aggregation mechanism, which dynamically summarizes instance information guided by a task-specific query derived from the bag-level textual prior. The fused attention map is obtained by marginalizing over the prompt dimension:

$$\begin{cases} \alpha_{\text{fused}}^{(i)} = \sum_j \tilde{\mathbf{S}}_{j,:}^{(i)} \\ X_{\text{fused}}^{(i)} = \sum_k \alpha_{\text{fused}}^{(k)} \cdot \mathcal{X}^{(i)} \end{cases} \tag{11}$$

Next, bag priors as query representation z_{bag} are used to perform cross-attention over the fused visual tokens:

$$H^{(i)} = \text{CrossAttn}(z_{\text{bag}}, X_{\text{fused}}^{(i)}, X_{\text{fused}}^{(i)}) \tag{12}$$

Finally, the aggregated bag representation $H \in \mathbb{R}^{1\times d}$ is forwarded through a classification head $g(\cdot)$ followed by an activation function $\sigma(\cdot)$ to produce the bag-level logits:

$$\tilde{p}^{(i)} = \sigma(g(H^{(i)})) \tag{13}$$

This querying-based aggregation allows the model to focus on task-relevant fine-grained by semantic prompts, enhancing both interpretability and discriminative power.

Training Strategy

Under few-shot weakly supervised learning (FSWL), the model is trained on limited WSIs per class to enhance robustness in low-data scenarios. Libra-MIL is optimized with cross-entropy loss to minimize the negative log-likelihood of ground truth $y^{(i)}$ in a batch N:

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N} y^{(i)} \log(\tilde{p}^{(i)}) \tag{14}$$

During Stereoscopic Infusion, the Sinkhorn algorithm performs iterative row- and column-wise normalization to satisfy marginal constraints, yielding a differentiable doubly stochastic matrix that minimizes the regularized transport cost for end-to-end training.

Experiments

Settings

Datasets We conduct experiments on three publicly available histopathology datasets: **TCGA-RCC(n=925)**, **TCGA-NSCLC(n=1052)** (Cancer Genome Atlas Research Network et al. 2013), and **CAMELYON16(n=399)** (Bejnordi et al. 2017) . TCGA-RCC contains WSIs of renal cell carcinoma, covering major subtypes of clear cell (KIRC), papillary (KIRP), and chromophobe (KICH). TCGA-NSCLC comprises lung cancer WSIs, including lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC), providing subtype annotations and high-resolution WSIs suitable for classification. Both datasets are from The Cancer Genome Atlas. CAMELYON16 is a benchmark dataset for detecting breast cancer, consisting of H&E-stained WSIs with detailed pixel-level annotations of tumor regions. Refer to C.4 for more detailed descriptions.

Metrics We evaluate all models under 1-, 4-, and 16-shot settings using accuracy (ACC), area under the curve (AUC), and F1-score, averaged over five-fold cross-validation with mean and standard deviation reported. To ensure fair comparison, all methods use identical dataset splits. Results in the table are reported as percentages. Implementation details, environments, and hyperparameters are provided in Appendix C.1, C.2, and C.5.

Comparison Results

We evaluated Libra-MIL against several competitive methods under K-shot (1, 4, 16) few-shot learning settings, as detailed in Table 1. The baseline methods included Max Pooling, ABMIL, TransMIL, CLAM, DSMIL, PAMIL, FOCUS, and TOP-MIL. Across these diverse settings, Libra-MIL demonstrably improved average ACC by 2.49%, AUC by 2.61%, and F1-score by 2.15% compared to the SOTA, and Libra-MIL's performance superiority persists with an increasing number of available shots. We also compare the computational efficiency of methods in Appendix C.7. While methods such as DSMIL, FOCUS, TOP-MIL, and Libra-MIL all incorporate multi-scale information, and the latter three all contain text priors. Our approach uniquely integrates multi-modal prototype-based Multiple Instance Learning (MIL). This addresses task-specific text and vision prior bias and often encountered in few-shot scenarios.

Ablation and Hyperparameters Studies

Modules To further explore the proposed components with the effectiveness of Libra-MIL, we performed the single module ablation experiments in Table 2. The ablation studies on TCGA-RCC in different few-shot settings investigate the influence of text-priors, vision prototypes, similarities stereoscopic infusion and bag query. The results demonstrate consistent performance improvements with each additional component. In the extremely one-shot setting (w/o Bag Querying), guidance from global information is crucial. The performance drop when replacing optimal transport with concatenation and cross-attention highlights the advantage of the proposed module. (Appendix C.5 for full).

Pathology Pre-training Models We further investigated the impact of cross-modal latent space misalignment by varying the visual encoder while keeping the text encoder constant. As shown in Table 3, different pathology foundation models (Phikon (Filiot et al. 2023), GigaPath (Xu et al. 2024), UNI (Chen et al. 2024b)) exhibited significant performance disparities under various few-shot settings compared to CONCH, despite utilizing the same language prior. This highlights the sensitivity of multimodal prototype learning to the alignment of the underlying feature spaces and suggests that cross-modal consistency plays a crucial role in MIL performance. The results underscore the importance of designing robust fusion mechanisms capable of addressing the heterogeneity of cross-modal representations, which is also one of the limitations of our approach.

Impact of LLM Knowledge To assess how variations in background knowledge across large language models (LLMs) affect task performance, we conduct a comparative

Dataset	Methods	1-shot			4-shot			16-shot		
		ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1
TCGA-RCC	Max pooling	46.7±13.9	60.6±12.7	35.1±12.4	71.4±8.8	87.2±8.2	65.9±10.1	92.8±0.9	98.2±0.3	91.0±1.4
	ABMIL (Ilse, Tomczak, and Welling 2018)	53.7±8.1	76.5±5.6	46.8±8.1	88.4±2.5	96.8±1.3	85.2±2.6	92.8±0.8	98.4±0.6	91.0±1.6
	TransMIL (Shao et al. 2021)	52.8±10.2	67.8±7.0	41.3±6.6	81.2±4.5	94.2±2.9	77.4±5.3	89.5±2.0	97.8±0.7	87.2±1.8
	CLAM_SB (Lu et al. 2021)	55.8±9.5	77.7±7.8	47.2±7.2	88.2±3.1	97.3±1.3	85.0±3.8	93.4±0.7	98.6±0.5	91.4±0.7
	CLAM_MB (Lu et al. 2021)	56.0±9.3	79.3±7.7	46.9±10.6	89.0±2.9	96.9±1.6	85.8±3.2	93.4±1.2	98.6±0.7	91.5±1.0
	DSMIL (Li, Li, and Eliceiri 2021)	43.7±23.6	67.6±12.6	34.2±17.9	82.0±7.6	94.8±2.2	78.0±6.8	91.5±1.3	98.2±0.7	88.9±1.9
	TOP-MIL (Qu et al. 2023)	57.3±18.7	70.6±11.4	44.5±10.0	83.5±4.3	93.5±5.3	75.3±6.3	92.0±1.4	98.4±0.5	91.3±2.0
	PAMIL (Liu et al. 2024)	60.1±18.2	77.9±11.5	54.4±16.4	88.8±3.6	97.1±2.2	85.7±4.5	93.6±0.9	98.6±0.6	92.0±0.7
	FOCUS (Guo et al. 2025)	63.2±15.5	85.0±5.4	54.4±15.9	89.7±3.2	97.6±0.9	87.3±3.8	92.2±0.9	98.4±0.5	90.4±1.4
	Libra-MIL	**71.2**±12.4	**87.2**±6.9	**62.4**±12.6	**91.4**±3.2	**98.2**±3.1	**88.5**±5.3	**93.8**±0.9	**98.8**±0.3	**92.3**±1.4
NSCLC	Max pooling	55.5±12.8	55.4±21.6	50.8±30.8	73.5±12.2	78.1±14.5	73.5±13.4	87.7±6.1	93.5±6.1	87.8±5.6
	ABMIL (Ilse, Tomczak, and Welling 2018)	59.4±6.8	64.6±9.9	43.7±22.0	74.5±11.6	80.9±12.4	74.3±13.5	89.6±1.9	95.6±1.6	89.5±1.7
	TransMIL (Shao et al. 2021)	54.5±5.3	59.0±5.7	51.0±10.5	66.1±8.2	71.6±9.7	63.2±14.4	83.4±2.5	91.1±1.8	83.1±3.0
	CLAM_SB (Lu et al. 2021)	55.4±4.7	65.1±7.5	48.1±19.8	76.0±8.6	83.6±8.6	76.0±8.9	89.6±2.1	96.1±2.2	89.3±1.9
	CLAM_MB (Lu et al. 2021)	56.4±2.6	67.7±9.4	46.8±20.9	74.9±6.3	83.6±7.3	75.9±6.0	90.7±2.0	96.2±2.1	**90.4**±1.9
	DSMIL (Li, Li, and Eliceiri 2021)	55.9±11.6	60.3±13.7	34.1±29.7	67.2±16.6	71.7±19.8	62.9±20.9	86.6±5.3	92.6±4.7	86.5±5.3
	TOP-MIL (Qu et al. 2023)	65.4±7.1	74.3±12.7	65.1±10.1	78.3±10.0	84.8±10.3	78.7±9.6	89.0±1.8	95.9±1.3	88.6±1.8
	PAMIL (Liu et al. 2024)	61.2±7.9	67.0±6.1	68.9±3.0	77.9±9.7	85.3±10.0	78.2±10.6	89.5±2.7	96.3±2.1	89.2±2.7
	FOCUS (Guo et al. 2025)	61.9±8.3	64.8±9.4	58.2±13.7	77.4±7.7	85.1±7.6	76.6±8.7	89.7±2.8	96.5±1.6	89.7±2.8
	Libra-MIL	**69.2**±9.2	**75.5**±11.9	**70.6**±9.1	**83.2**±6.5	**91.5**±4.9	**81.1**±9.7	**90.8**±1.4	**96.8**±1.1	**90.4**±1.5
CAMELYON16	Max pooling	49.8±11.7	57.0±8.0	40.5±23.3	64.7±7.7	64.3±11.0	50.3±11.2	88.7±2.9	91.9±4.2	84.4±4.0
	ABMIL (Ilse, Tomczak, and Welling 2018)	53.2±9.3	46.2±11.7	21.6±14.1	52.7±9.1	47.2±11.4	19.5±11.7	88.7±3.0	89.6±4.5	83.4±4.8
	TransMIL (Shao et al. 2021)	61.1±2.7	55.6±8.1	24.4±22.6	55.8±7.7	51.6±12.1	39.9±14.4	62.2±2.5	57.0±7.3	34.7±9.0
	CLAM_SB (Lu et al. 2021)	54.7±9.3	47.0±10.7	25.2±22.7	55.0±9.8	55.3±11.0	24.9±18.1	88.8±5.7	93.1±4.1	84.7±7.1
	CLAM_MB (Lu et al. 2021)	48.2±7.6	43.4±5.7	36.2±17.4	52.1±10.1	50.4±10.7	19.5±12.8	89.8±4.3	93.5±3.5	85.9±5.5
	DSMIL (Li, Li, and Eliceiri 2021)	51.0±7.7	54.3±6.0	37.5±21.5	51.8±9.6	51.7±11.9	31.9±21.8	63.7±6.6	62.5±7.8	52.9±7.8
	TOP-MIL (Qu et al. 2023)	53.5±9.9	49.1±10.2	42.2±20.6	59.7±15.4	60.3±14.8	45.7±17.3	89.2±2.8	93.2±1.7	84.0±5.0
	PAMIL (Liu et al. 2024)	46.8±12.2	52.3±10.3	43.5±24.4	57.1±8.7	57.0±11.2	42.4±12.2	90.5±4.5	95.0±3.1	87.3±5.3
	FOCUS (Guo et al. 2025)	51.6±6.2	47.7±8.5	46.2±6.4	61.2±14.8	61.5±15.3	56.0±15.5	90.9±1.0	93.7±2.7	89.9±1.1
	Libra-MIL	**62.0**±5.3	**64.9**±9.9	**46.6**±16.5	**65.0**±8.5	**64.8**±7.1	**57.2**±10.3	**91.5**±4.7	**95.7**±5.1	**90.6**±6.2

Table 1: Comparison of classification performance under 1-, 4-, and 16-shot settings across all baselines. Best results are **bolded**, second-best are underlined. Statistical significance ($p<0.05$) is determined by paired t-tests with 95% confidence intervals over repeated runs.

Methods	1-shot			4-shot			16-shot		
	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1
w/o instance priors in T-LPG	67.0	83.6	56.1	87.8	96.9	84.3	92.0	98.5	90.7
w/o vision prototypes	66.0	84.7	55.5	87.2	97.4	83.7	90.3	98.0	87.5
SOT → concat	66.4	81.2	54.1	86.3	97.1	82.3	92.0	98.4	90.4
SOT → crossAttn	65.8	80.9	53.3	88.8	97.0	85.6	92.1	98.5	90.9
w/o bag priors in SA	55.4	82.4	47.3	88.0	97.0	88.5	92.3	98.4	90.2
Libra-MIL	**71.2**	**87.2**	**62.4**	**91.4**	**98.2**	**88.5**	**93.8**	**98.8**	**92.3**

Table 2: Ablation Studies of Libra-MIL on TCGA-RCC.

Models	1-shot			4-shot			16-shot		
	ACC	AUC	F1	ACC	AUC	F1	ACC	AUC	F1
Phikon (Filiot et al. 2023)	62.2	66.6	58.9	72.2	78.8	65.1	89.0	95.9	89.0
GigaPath (Xu et al. 2024)	65.1	72.1	66.2	71.1	77.7	63.1	87.2	94.5	87.3
UNI (Chen et al. 2024b)	65.4	68.9	63.2	72.9	79.6	71.5	88.7	95.5	88.4
CONCH (Lu et al. 2024)	69.2	75.5	70.6	83.2	91.5	81.1	90.8	96.8	90.4

Table 3: Variations in pathology foundation models markedly affect outcomes on the TCGA-NSCLC.

study using a unified prompt (see Appendix C.) across several popular LLMs, including Gemini 2.5 (Comanici et al. 2025), Qwen3 (Yang et al. 2025), OpenAI-o4mini (Menick et al. 2024), Claude4-Sonnet (Anthropic 2025), and GPT-4o (Hurst et al. 2024). Each model generated task-specific textual descriptions at both the instance and bag levels, which are then integrated into the Libra-MIL framework

and evaluated under a 4-shot setting on the TCGA-NSCLC dataset. As shown in Fig. 3, GPT-4o consistently outperformed other LLMs, while Libra-MIL maintained stable performance across models and consistently surpassed the baseline under identical conditions. These results demonstrate Libra-MIL's robustness to variations in textual semantics and indicate that more expressive, knowledge-rich LLMs can further enhance performance.

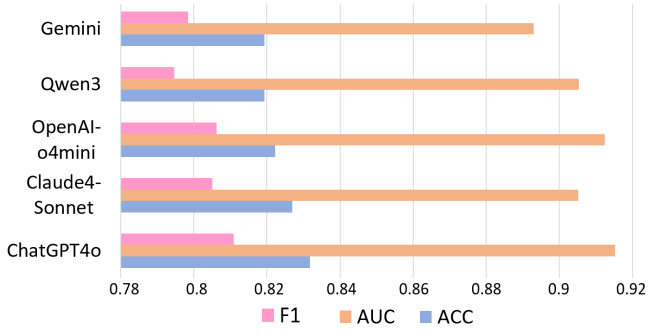

Figure 3: Results of different LLMs' priors on TCGA-NSCLC under 4-shot setting.

Vision Prototypes Number To assess the influence of the number of vision prototypes on model performance, we conducted ablation studies by varying K_v. As shown in the Table 4, the model achieves the best overall performance when $K_v = 6$ indicating an optimal trade-off between expressive capacity and generalization. When K_v is small, the model may struggle to capture the diversity of visual patterns. As K_v increases, the performance metrics generally decline, suggesting that introducing too many prototypes may lead to a sparse latent space and ambiguous class boundaries.

K_v	2	4	6	8	10	12
ACC	82.7±7.7	81.8±10.2	83.2±6.5	80.0±9.8	81.6±10.3	79.5±11.6
AUC	90.9±6.1	89.3±9.5	91.5±4.9	88.5±9.6	89.5±9.9	87.6±11.3
F1	81.8±10.5	79.3±15.2	81.1±9.7	80.3±9.1	82.5±8.9	76.4±18.8

Table 4: Performance with varying number of visual prototypes under 4-shot setting on TCGA-NSCLC.

Multimodal Prototypes with Interpretability

To llustrate the interpretability of Libra-MIL, we visualize the most representative visual patches as visual prototypes, along with their corresponding textual prototypes, under the $K_v = 6$ setting in Fig. 4. These prototypes capture meaningful histological features—such as nerve bundles, mitotic figures, and keratin beads—demonstrating clear diagnostic relevance. Additionally, we present prototype similarity matrices and attention maps of bag-level priors for two TCGA-NSCLC cases, showing that the model consistently attends to clinically significant regions. These results underscore the capability of Libra-MIL's multimodal prototypes to capture discriminative patterns and enhance interpretability.

Joint Contribution of Multimodal Prototypes

To investigate the influence of different modality-specific prototypes on the final diagnostic decision, we applied gradient-based analysis (Lundberg et al. 2020) of the Libra-MIL (Fig. 5). Specifically, in the 4-shot setting on the TCGA-RCC dataset, we backpropagated the classification output to the prototype embeddings to assess the contribution of each modality-specific prototype across feature dimensions. The results show that the average gradient magnitudes of visual and textual prototypes are comparable, indicating that both modalities play equally important roles in the decision-making process.

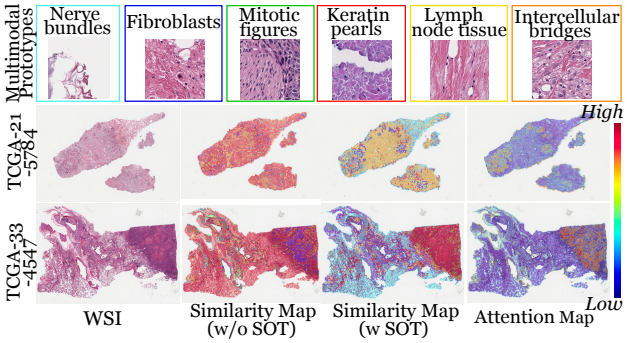

Figure 4: Case study of multimodal prototypes with different histological morphologies and semantic attention on WSIs.

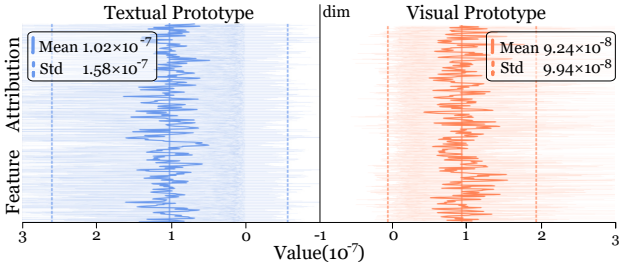

Figure 5: Gradient-based Contribution of textual and visual prototypes.

Conclusion

We present Libra-MIL, a multimodal prototype-based MIL framework designed for computational pathology under limited supervision. To overcome key limitations of existing VLMIL methods—such as biased LLM-generated descriptions and unidirectional alignment—we introduce task-specific textual priors and propose a bidirectional fusion strategy based on Stereoscopic Optimal Transport. By constructing both visual and textual prototypes, Libra-MIL enables interpretable and structure-aware cross-modal reasoning. Few-shot experiments on three cancer datasets demonstrate its superior performance and enhanced interpretability. Future work will explore rapid domain-adaptive tuning, broader modality integration, and extensions to more diverse tasks such as survival regression analysis.

Acknowledgments

This work was supported by National Natural Science Foundation of China (Grant No. 62371409) and Fujian Provincial Natural Science Foundation of China (Grant No. 2023J01005).

References

Anthropic. 2025. System Card: Claude Opus 4 & Claude Sonnet 4. https://www-cdn.anthropic.com/07b2a3f9902ee19fe39a36ca638e5ae987bc64dd.pdf.

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

Brenier, Y. 1991. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4): 375–417.

Cancer Genome Atlas Research Network; et al. 2013. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, 45: 1113–1120.

Chen, R. J.; Ding, T.; Lu, M. Y.; Williamson, D. F.; Jaume, G.; Chen, B.; Zhang, A.; Shao, D.; Song, A. H.; Shaban, M.; et al. 2024a. Towards a General-Purpose Foundation Model for Computational Pathology. *Nature Medicine*.

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. 2024b. Towards a general-purpose foundation model for computational pathology. *Nature medicine*, 30(3): 850–862.

Chen, R. J.; Lu, M. Y.; Weng, W.-H.; Chen, T. Y.; Williamson, D. F.; Manz, T.; Shady, M.; and Mahmood, F. 2021. Multimodal Co-Attention Transformer for Survival Prediction in Gigapixel Whole Slide Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4025.

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Cuturi, M. 2013. Sinkhorn Distances: Lightspeed Computation of Optimal Transportation Distances. *Advances in Neural Information Processing Systems*, 26: 2292–2300.

Ding, K.; Zhou, M.; Metaxas, D. N.; and Zhang, S. 2023. Pathology-and-genomics multimodal transformer for survival outcome prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 622–631. Springer.

Filiot, A.; Ghermi, R.; Olivier, A.; Jacob, P.; Fidon, L.; Camara, A.; Mac Kain, A.; Saillard, C.; and Schiratti, J.-B. 2023. Scaling self-supervised learning for histopathology with masked image modeling. *MedRxiv*, 2023–07.

Gadermayr, M.; and Tschuchnig, M. 2024. Multiple instance learning for digital pathology: A review of the state-of-the-art, limitations & future potential. *Computerized Medical Imaging and Graphics*, 112: 102337.

Gou, J.; Ji, L.; Liu, P.; and Ye, M. 2025. Queryable Prototype Multiple Instance Learning with Vision-Language Models for Incremental Whole Slide Image Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 3158–3166.

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.

Hou, W.; Yu, L.; Lin, C.; Huang, H.; Yu, R.; Qin, J.; and Wang, L. 2022. Hˆ 2-MIL: exploring hierarchical representation with heterogeneous multiple instance learning for whole slide image analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, 933–941.

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.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Ilse, M.; Tomczak, J.; and Welling, M. 2018. Attention-based deep multiple instance learning. In *International conference on machine learning*, 2127–2136. PMLR.

Jaume, G.; Vaidya, A.; Chen, R. J.; Williamson, D. F.; Liang, P. P.; and Mahmood, F. 2024. Modeling dense multimodal interactions between biological pathways and histology for survival prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11579–11590.

Li, B.; Li, Y.; and Eliceiri, K. W. 2021. 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*, 14318–14328.

Li, H.; Zhu, C.; Zhang, Y.; Sun, Y.; Shui, Z.; Kuang, W.; Zheng, S.; and Yang, L. 2023. Task-specific fine-tuning via variational information bottleneck for weakly-supervised pathology whole slide image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7454–7463.

Liu, J.; Mao, A.; Niu, Y.; Zhang, X.; Gong, T.; Li, C.; and Gao, Z. 2024. Pamil: Prototype attention-based multiple instance learning for whole slide image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 362–372. 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: 863–874.

Lu, M. Y.; Williamson, D. F.; Chen, T. Y.; Chen, R. J.; Barbieri, M.; and Mahmood, F. 2021. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6): 555–570.

Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; and Lee, S.-I. 2020. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1): 2522–5839.

Menick, J.; Lu, K.; Zhao, S.; Wallace, E.; Ren, H.; Hu, H.; Stathas, N.; and Such, F. P. 2024. GPT-4o mini: advancing cost-efficient intelligence. *Open AI: San Francisco, CA, USA*.

Omar, M.; Ullanat, V.; Loda, M.; Marchionni, L.; and Umeton, R. 2024. ChatGPT for digital pathology research. *The Lancet Digital Health*, 6(8): e595–e600.

Qu, L.; Fu, K.; Wang, M.; Song, Z.; et al. 2023. The rise of ai language pathologists: Exploring two-level prompt learning for few-shot weakly-supervised whole slide image classification. *Advances in Neural Information Processing Systems*, 36: 67551–67564.

Qu, L.; Yang, D.; Huang, D.; Guo, Q.; Luo, R.; Zhang, S.; and Wang, X. 2024. Pathology-knowledge enhanced multi-instance prompt learning for few-shot whole slide image classification. In *European conference on computer vision*, 196–212. Springer.

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.

Rymarczyk, D.; Pardyl, A.; Kraus, J.; Kaczyńska, A.; Skomorowski, M.; and Zieliński, B. 2022. Protomil: Multiple instance learning with prototypical parts for whole-slide image classification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 421–436. Springer.

Sahoo, P.; Singh, A. K.; Saha, S.; Jain, V.; Mondal, S.; and Chadha, A. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

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

Shi, J.; Li, C.; Gong, T.; Zheng, Y.; and Fu, H. 2024. ViLa-MIL: 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.

Snell, J.; Swersky, K.; and Zemel, R. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

Sun, S.; Tessier, L.; Meeuwsen, F.; Grisi, C.; van Midden, D.; Litjens, G.; and Baumgartner, C. F. 2025. Label-free Concept Based Multiple Instance Learning for Gigapixel Histopathology. *arXiv preprint arXiv:2501.02922*.

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.

Xu, Y.; and Chen, H. 2023. Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, 21241–21251.

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*.

Yang, L.; Mehta, D.; Liu, S.; Mahapatra, D.; Di Ieva, A.; and Ge, Z. 2023. Tpmil: Trainable prototype enhanced multiple instance learning for whole slide image classification. *arXiv preprint arXiv:2305.00696*.

Zhang, H.; Meng, Y.; Zhao, Y.; Qiao, Y.; Yang, X.; Coupland, S. E.; and Zheng, Y. 2022. 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*, 18802–18812.