

MUSE: Harnessing Precise and Diverse Semantics for Few-Shot Whole Slide Image Classification

Jiahao Xu¹, Sheng Huang^{1*}, Xin Zhang¹, Zhixiong Nan², Jiajun Dong¹ and Nankun Mu²

¹School of Big Data & Software Engineering, Chongqing University,

²College of Computer Science, Chongqing University

{xujiahao, zhangxin, dongjiajun}@stu.cqu.edu.cn, {huangsheng, nanzx, nankun.mu}@cqu.edu.cn

Abstract

In computational pathology, few-shot whole slide image classification is primarily driven by the extreme scarcity of expert-labeled slides. Recent vision-language methods incorporate textual semantics generated by large language models, but treat these descriptions as static class-level priors that are shared across all samples and lack sample-wise refinement. This limits both the diversity and precision of visual-semantic alignment, hindering generalization under limited supervision. To overcome this, we propose the stochastic MULTI-view Semantic Enhancement (MUSE), a framework that first refines semantic precision via sample-wise adaptation and then enhances semantic richness through retrieval-augmented multi-view generation. Specifically, MUSE introduces Sample-wise Fine-grained Semantic Enhancement (SFSE), which yields a fine-grained semantic prior for each sample through MoE-based adaptive visual-semantic interaction. Guided by this prior, Stochastic Multi-view Model Optimization (SMMO) constructs an LLM-generated knowledge base of diverse pathological descriptions per class, then retrieves and stochastically integrates multiple matched textual views during training. These dynamically selected texts serve as enriched semantic supervisions to stochastically optimize the vision-language model, promoting robustness and mitigating overfitting. Experiments on three benchmark WSI datasets show that MUSE consistently outperforms existing vision-language baselines in few-shot settings, demonstrating that effective few-shot pathology learning requires not only richer semantic sources but also their active and sample-aware semantic optimization. Our code is available at: https://github.com/JiahaoXu-god/CVPR2026_MUSE.

*Corresponding author

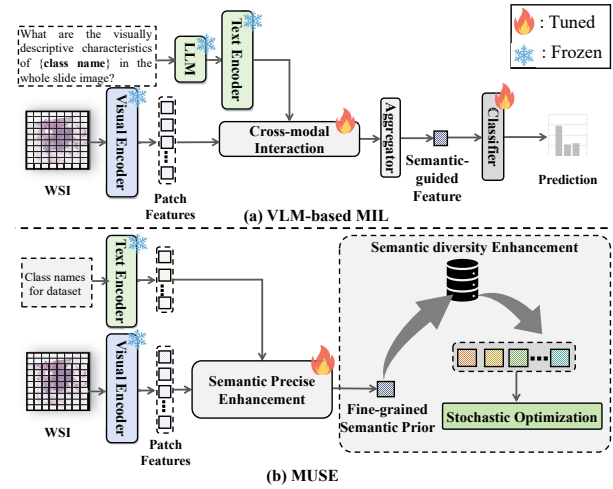


Figure 1. Comparison of MUSE with existing VLM-based MIL methods. (a) VLM-based MIL methods incorporate pathological text and enable cross-modal interaction between text and image modalities. (b) Our method performs fine-grained modeling of semantics and enables interaction between text and image modalities, while enhancing text diversity through knowledge base retrieval and stochastic optimization.

1. Introduction

Computational pathology (CPath) [7, 8, 40], a key branch of digital pathology [31], leverages advanced machine learning to enable objective and quantitative analysis of whole slide images (WSIs). These methods support automated diagnosis by interpreting complex histopathological patterns. However, the extreme scale and structural heterogeneity of WSIs present major obstacles to end-to-end learning. To circumvent the need for exhaustive pixel-level annotations, weakly supervised learning has become the de facto paradigm, most notably through the multiple instance learning (MIL) framework [23, 29, 36, 42].

Conventional MIL approaches for WSI analysis [9, 20, 26, 41, 45, 48, 49] typically follow a three-stage pipeline:

(1) patch extraction, (2) feature encoding via a pretrained backbone, and (3) bag-level aggregation under slide-level supervision. These methods have demonstrated strong performance in tasks such as cancer diagnosis [20, 41], subtyping [9, 26, 49], and survival prediction [45, 48]—but only when sufficient labeled WSIs are available.

In practice, however, acquiring slide-level annotations requires expert pathologists and is further constrained by stringent data privacy regulations. Consequently, labeled WSIs are not only scarce but often limited to just a few examples per diagnostic category. This reality necessitates a few-shot learning paradigm, where models must generalize from minimal labeled slides. Under such extreme data scarcity, visual features alone are insufficient to capture the nuanced diagnostic criteria that distinguish pathological subtypes. Instead, high-level semantic knowledge such as disease descriptions, histological terminology, and clinical narratives becomes critical. Such semantics encode invariant diagnostic principles shared across patients and institutions, offering a powerful inductive bias for generalization.

Motivated by this, recent data-efficient methods [10, 15, 33, 34, 38] leverage vision-language models (VLMs) [24, 35] pretrained on large-scale image-text corpora to align histopathological patterns with their corresponding textual semantics. By fusing or interacting visual and semantic representations, these approaches aim to solve the few-shot WSI classification (FSWC) problem, effectively using language as a proxy for expert knowledge when labeled data is unavailable.

Recent vision-language methods for computational pathology have moved beyond generic models trained on natural images by adopting pathology-specific foundation models such as PLIP [18], CONCH [28], and MUSK [46]. These models are pretrained on large-scale histopathology image-text pairs via contrastive learning [18, 19, 28, 46] and provide domain-aligned encoders for both modalities. Building on this progress, several approaches [14, 30, 38, 39] leverage large language models to generate additional textual descriptions and design cross-modal interactions between text and patch features, as shown in Figure 1(a). However, large language models are typically used only as description generators rather than semantic optimizers, resulting in static and unrefined textual prompts.

This superficial use of semantics leads to two key limitations. First, complex pathological concepts are often collapsed into a single global query, preventing disentanglement of fine-grained diagnostic attributes such as tumor grade or immune infiltration. As a result, visual-semantic alignment remains coarse and fails to attend to diagnostically relevant regions with concept-level precision. Second, the reliance on unoptimized prompts ignores the structural diversity of clinical language, including variations in abstraction level, contextual nuance, and syntactic formula-

tion. Under few-shot settings, this not only underutilizes the expressive capacity of the text encoder but also encourages overfitting to specific phrasings, degrading generalization across clinical contexts.

To address the aforementioned limitations, we propose a stochastic **M**ulti-view **S**emantic **E**nhancement framework, abbreviated as **MUSE**. As shown in Figure 1(b), the core idea of MUSE is to jointly enhance the model’s generalization capability through precise semantic perception and enriched semantic diversity. Our framework consists of two core components: Sample-wise Fine-grained Semantic Enhancement (SFSE) and Stochastic Multi-view Model Optimization (SMMO). SFSE enhances semantic precision through compositional semantic refinement and fine-grained, query-driven sample-wise cross-modal interaction. SMMO promotes semantic diversity by retrieving multi-view textual descriptions from a contextual semantic knowledge base generated by an LLM with semantics refined by SFSE, followed by a fast stochastic optimization process for final WSI classification. These components enable MUSE to achieve strong generalization in few-shot settings. Extensive experiments on multiple widely used WSI datasets show that our method achieves superior performance compared to existing FSWC baselines. The main contributions of our paper can be summarized as follows:

- We propose the MUSE framework, which improves semantic understanding in multiple instance learning through fine-grained semantic modeling and effective exploitation of semantic diversity, significantly boosting generalization in few-shot scenarios. To the best of our knowledge, this work is the first attempt to improve few-shot WSI classification performance from the perspective of semantic optimization.
- We propose an MoE based mechanism that compositionally refines category level semantics and adapts them to individual samples through interaction with visual features. This enables the learning of sample-wise semantic priors that capture fine-grained semantic cues, thereby enhancing semantic precision beyond conventional class level representations.
- We build an LLM-generated knowledge base of multi-view and class-specific pathological descriptions, whose semantic diversity offers complementary signals for few-shot learning. Guided by SFSE-refined sample-wise priors and integrated stochastically, these multi-view semantics enhance generalization under limited labels.

2. Related Works

2.1. Multiple Instance Learning in CPath

Due to the ultra-high resolution of whole-slide images (WSIs), multiple instance learning (MIL) has become the standard framework in computational pathology. MIL-

based methods have demonstrated strong performance on WSI diagnosis, subtyping, and prognosis tasks [20, 26, 45, 48, 49]. A typical pipeline first tiles the WSI into patches, extracts features using a pretrained encoder, and then aggregates these features to predict the slide-level label. Early aggregators relied on parameter-free pooling operations [6]. Subsequent works introduced learnable mechanisms to identify diagnostically relevant patches. AB-MIL [20] uses attention to assign importance scores to individual patches. CLAM [27] enhances this with clustering constraints to localize critical regions. TransMIL [37] models global inter-patch dependencies via self-attention, while graph-based methods [16, 25, 50] incorporate spatial structure to improve contextual reasoning. However, under sparse annotations, models relying solely on visual features often underperform, highlighting the need to integrate domain knowledge for effective weakly supervised learning.

2.2. Vision-Language Models in CPath

General-purpose vision-language models such as CLIP [35] and BLIP [24] have demonstrated strong performance across a wide range of visual tasks [13, 17, 21, 47]. In computational pathology (CPath), domain-adapted foundation models including PLIP [18], CONCH [28], and MUSK [46] leverage large-scale pathological image-text data to improve diagnosis and downstream analysis. To address sparse annotations, recent works integrate these models into the MIL framework by exploiting their few-shot and zero-shot transfer capabilities. Top [33] introduces the FSWC paradigm using text-guided patch aggregation for WSI classification under data scarcity. ViLa-MIL [38] proposes a dual-scale MIL framework that fuses textual descriptions with image features at multiple resolutions. FOCUS [14] enhances representation quality through a knowledge-guided adaptive visual compression mechanism. However, these methods treat semantic prompts as static category-level descriptors, ignoring both sample-wise fine-grained semantics and the structural and perspectival diversity of clinical language, thereby limiting the expressiveness of vision-language learning.

3. Methodology

3.1. Overview

We propose a stochastic multi-view semantic enhancement framework for few-shot whole slide image classification, termed MUSE, as shown in Figure 2. The framework consists of two core components: sample-wise fine-grained semantic enhancement (SFSE) and stochastic multi-view model optimization (SMMO). Through these components, we achieve fine-grained semantic modeling and effectively harness semantic diversity, thereby enhancing the model’s understanding of the semantic modality and improving its

generalization capability under few-shot settings.

3.2. Fine-grained Semantic Enhancement

As shown in Figure 2(a), the sample-wise fine-grained semantic enhancement (SFSE) comprises two components: compositional semantic refinement (DSR) and sample-wise vision-text interaction (SVTI). DSR decomposes the input textual semantics and isolates task-relevant core semantic segments, which serve as semantic cues for cross-modal interaction. SVTI employs cross-attention to dynamically attend to visual patches using these cues as queries, selectively aggregating semantically relevant features at the sample level. This attention-driven fusion enriches the representation with fine-grained, sample-wise, and context-aware semantic information.

3.2.1. Compositional Semantic Refinement

To adapt the frozen pathology foundation model to the downstream FSWC task, we augment each category name $c \in C$ with M learnable prompt vectors. For example, in CAMELYON [3, 4], categories include “normal lymph node” and “metastatic lymph node”. The resulting text prompt for category c is formulated as:

$$T_c = [V]_1[V]_2 \dots [V]_M[c], \quad (1)$$

where $[V]_i \in \mathbb{R}^d$ ($i = 1, \dots, M$) are trainable embeddings. This prompt is encoded by the text encoder $E_T(\cdot)$ of the foundation model to produce a textual feature representation $D \in \mathbb{R}^{|C| \times d}$, where each row D_i corresponds to the encoded semantics of the i -th category.

Although prompt tuning enhances task adaptability, the resulting representation remains holistic and lacks explicit fine-grained structure. To address this, the Compositional Semantic Refinement (DSR) module refines D into task-relevant semantic cues. Specifically, inspired by the Mixture-of-Experts (MoE) paradigm, we construct R expert query matrices $\{W_i^Q\}_{i=1}^R$. A lightweight router network scores each expert based on the input semantics:

$$S = G(D) + \epsilon(D), \quad (2)$$

where $S \in \mathbb{R}^{|C| \times R}$, $G(\cdot)$ is a linear projection, and $\epsilon(D)$ injects input-dependent Gaussian noise to encourage expert diversity and prevent collapse.

For each category i , we select the top- k experts with the highest scores and retrieve their corresponding query matrices $\{W_{ij}^Q\}_{j=1}^k$. The semantic decomposition is then performed via:

$$Q_{ij} = D_i W_{ij}^Q, \quad j = 1, \dots, k. \quad (3)$$

This yields a set of refined semantic cues $\{Q_{ij}\}_{j=1}^k$ for category i , which serve as semantic queries for sample-wise vision-text interaction in the subsequent stage.

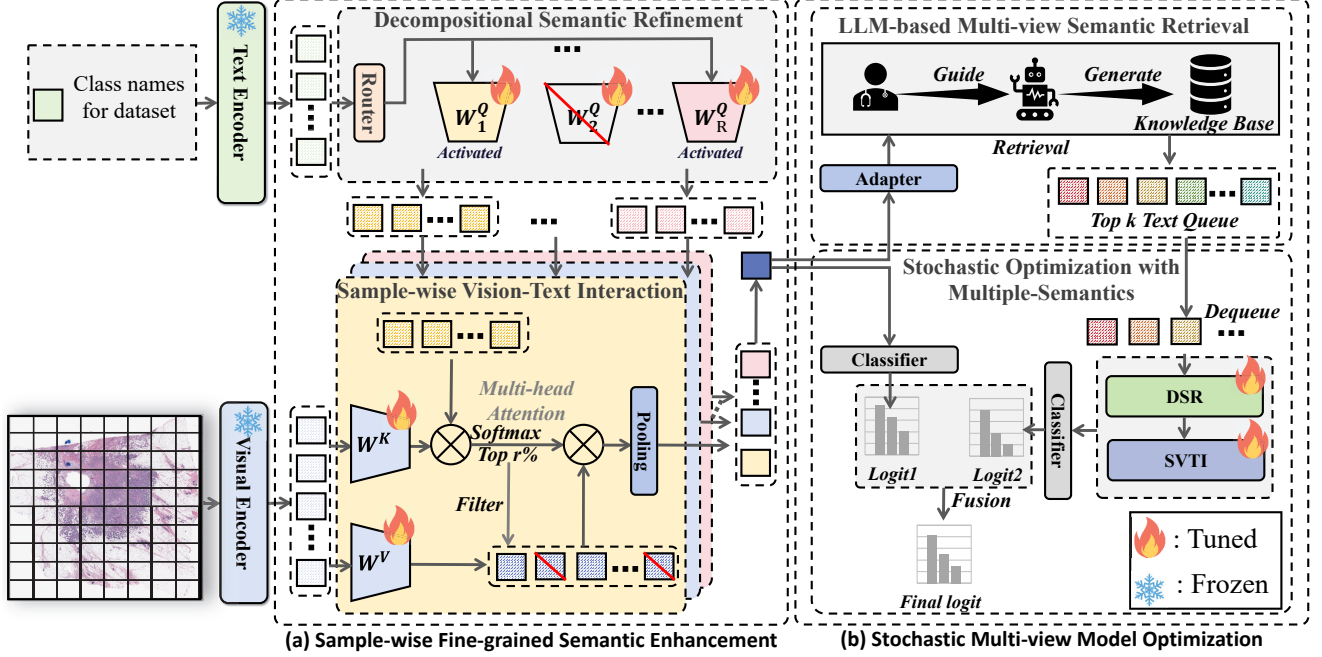


Figure 2. Overview of the proposed MUSE framework. (DSR: Decompositional Semantic Refinement. SVTI: Sample-wise Vision-Text Interaction) (a) The input semantic information is decomposed and modeled in a fine-grained manner. We then leverage the refined semantic representations to extract sample-relevant visual-semantic information and facilitate cross-modal interaction. (b) The semantic-enhanced features are used to retrieve relevant texts from the multi-view text knowledge base, and these retrieved texts are subsequently leveraged through stochastic optimization to enrich semantic diversity.

3.2.2. Sample-wise Vision-Text Interaction

For each input WSI, we first employ the visual feature extractor $E_I(\cdot)$ of the foundation model to extract patch-level features $\{b_n\}_{n=1}^N$ and form a bag representation $B \in \mathbb{R}^{N \times d}$.

Given the i -th category semantics D_i , we leverage the fine-grained semantic cues $\{Q_{ij}\}_{j=1}^k$ produced by DSR, which encode discriminative sub-concepts relevant to pathological diagnosis. The goal of SVTI is to dynamically identify and fuse the subset of cues most aligned with the visual content of the current sample through vision-text interaction, thereby constructing a sample-level semantic prior.

Specifically, we compute multi-head cross-attention between each cue Q_{ij} and the visual bag B . For attention head h , the alignment scores are:

$$A_{ij}^h = \frac{(Q_{ij}W^{Q,h})(BW^{K,h})^\top}{\sqrt{d_{\text{head}}}}, \quad (4)$$

where $W^{Q,h}, W^{K,h} \in \mathbb{R}^{d \times d_{\text{head}}}$ are learnable projection matrices.

To focus on regions highly correlated with the semantic cues, we retain only the top- $r\%$ patches with the highest attention scores for each head. Let $\mathcal{I}_{ij}^h$ denote the corresponding index set. The filtered value projection is then:

$$V_{ij}^{h,\text{filtered}} = B_{\mathcal{I}_{ij}^h} W^{V,h}, \quad (5)$$

with $W^{V,h} \in \mathbb{R}^{d \times d_{\text{head}}}$ as the value projection matrix.

We aggregate across heads to obtain a vision-text fused representation for cue j :

$$\text{head}_h = \text{Softmax}(A_{ij}^h) V_{ij}^{h,\text{filtered}}, \quad (6)$$

$$f_{ij} = \text{Concat}(\text{head}_1, \dots, \text{head}_H) W^O, \quad (7)$$

where $W^O \in \mathbb{R}^{H d_{\text{head}} \times d}$ is the output projection matrix.

Finally, we combine the k fused representations using the expert scores $\{S_{ij}\}_{j=1}^k$ from the DSR router, and obtain the final representation f through average pooling:

$$\{S_{ij}^{\text{norm}}\}_{j=1}^k = \text{Softmax}(\{S_{ij}\}_{j=1}^k), \quad (8)$$

$$f_i = \sum_{j=1}^k S_{ij}^{\text{norm}} f_{ij}, \quad f = \frac{1}{|C|} \sum_{i=1}^{|C|} f_i. \quad (9)$$

The resulting vector f serves as a fine-grained semantic prior for the input WSI with respect to the corresponding text feature D . This prior reflects the semantic sub-concepts most supported by the current sample visually, and is specifically designed to serve as a query for retrieving complementary textual knowledge from an external semantic knowledge base in the subsequent stage.

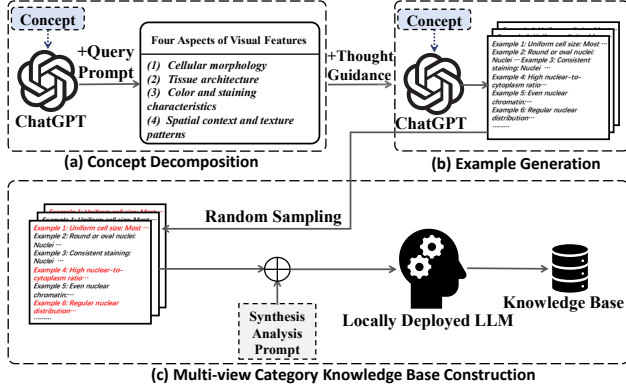


Figure 3. Our pipeline of the category related text knowledge base generation. (a) We employ ChatGPT to analyze and decompose the concepts associated with class names. (b) We leverage ChatGPT to construct concrete examples for each aspect. (c) We randomly sample examples and combine them with prompts to guide the locally deployed LLM in generating category knowledge base.

3.3. Stochastic Multi-view Model Optimization

As illustrated in Figure 2, the sample-specific semantic prior f generated by SVTI is used to query a pathology-oriented knowledge base, which is constructed offline via LLM-guided generation using chain-of-thought (CoT) [44] and in-context learning (ICL) [5] to produce diverse, multi-view textual descriptions. During training, the stochastic multi-view model optimization retrieves a set of semantically complementary texts using f and randomly samples one of them at each iteration to update the model, thereby improving generalization through exposure to diverse semantic views.

3.3.1. LLM-based Semantic Knowledge Base Generation

As shown in Figure 3, we construct a multi-view textual knowledge base for each pathological category c by leveraging category-level diagnostic concepts to guide LLM generation. The process involves three steps: concept decomposition, exemplar generation, and knowledge base assembly.

First, we prompt GPT-4 with the category name to decompose its visual diagnostic criteria into four clinically meaningful aspects: cellular morphology, tissue architecture, color-staining characteristics, and spatial-texture patterns (Figure 3(a)).

Next, for each aspect and category, GPT-4 generates 10 concrete exemplar descriptions (Figure 3(b)). To promote semantic diversity, we form synthesis prompts by randomly selecting one exemplar from each aspect and concatenating them. Due to API cost and time constraints, we use a lightweight open-source LLM deployed locally to generate 300 multi-view descriptions per category based on these prompts (Figure 3(c)).

Finally, all generated texts are encoded by the text en-

coder $E_T(\cdot)$ into feature vectors, forming the category-specific knowledge base $\mathcal{B}_c$. Additional details are provided in the Supplementary Material.

3.3.2. LLM-based Multi-view Semantic Retrieval

Given the sample-wise semantic prior f generated by SVTI and the ground-truth category label c , we retrieve semantically relevant text features from the pre-constructed knowledge base $\mathcal{B}_c$. The retrieval is performed by computing cosine similarity between an adapted representation of f and all entries in $\mathcal{B}_c$:

$$T_m = \text{Top-}m_{t \in \mathcal{B}_c} (\text{Sim}(A(f), t)), \quad (10)$$

where $A(\cdot)$ is a lightweight adapter that aligns the feature space of f with that of the text encoder, and $\text{Sim}(\cdot, \cdot)$ denotes cosine similarity.

To facilitate subsequent stochastic utilization, the retrieved set T_m is randomly shuffled and stored in a queue structure, enabling sequential popping of text features in a randomized order during optimization.

3.3.3. Stochastic Optimization with Multiple-Semantics

Building upon the retrieved text features T_m , we stochastically incorporate diverse textual views to enrich semantic representation. The set T_m is stored in a shuffled queue to enable random access during training.

At each iteration, a single text feature t is dequeued from T_m . We apply Decompositional Semantic Refinement (DSR) to decompose t into fine-grained semantic queries $\{Q_j^t\}_{j=1}^k$, which are then processed by the Sample-wise Vision-Text Interaction (SVTI) module together with the input bag B to produce an auxiliary sample-wise prior f_{aux}^t . This prior introduces semantic information from additional textual views, thereby compensating for the semantic representation of the primary prior f .

Both f and f_{aux}^t are projected into logits via a shared MLP:

$$z = \text{MLP}(f), \quad z_{\text{aux}}^t = \text{MLP}(f_{\text{aux}}^t), \quad (11)$$

and fused by summation:

$$z_{\text{final}}^t = \frac{z + z_{\text{aux}}^t}{2}. \quad (12)$$

The model is trained with cross-entropy loss:

$$\mathcal{L}^t = \text{CE}(z_{\text{final}}^t, GT), \quad (13)$$

where GT denotes the sample-wise (slide-level) ground-truth label.

By stochastically exposing the model to multiple textual views from T_m , this procedure realizes multi-view semantic compensation, thereby improving generalization in few-shot scenarios. The full training algorithm is provided in the Supplementary Material.

4. Experiments

4.1. Experimental Settings

4.1.1. Datasets

To validate the effectiveness of MUSE, we conduct experiments on three datasets: CAMELYON [3, 4], TCGA-NSCLC [43], and TCGA-BRCA [43]. CAMELYON, designed for breast cancer metastasis detection, consists of 591 normal slides and 308 slides containing metastases. TCGA-NSCLC contains slides from two lung cancer subtypes: lung adenocarcinoma (LUAD, 541 slides) and lung squamous cell carcinoma (LUSC, 512 slides). TCGA-BRCA contains two breast cancer subtypes: invasive ductal carcinoma (IDC, 826 slides) and invasive lobular carcinoma (ILC, 211 slides). Each dataset is split into training, validation, and test sets with a 6:2:2 ratio. In the few-shot setting, we randomly sample k slides per class from the training set, with $k = 4, 8, 16$.

4.1.2. Implementation Details

The original whole slide image (WSI) at $40\times$ magnification is processed by the CLAM toolset [27] into 512×512 -pixel patches. We utilize the text encoder and visual encoder of CONCH [28] to extract textual features and patch features, respectively. To construct the multi-perspective textual knowledge base, we first employ GPT-4 [1] for concept decomposition and example generation, and then build the knowledge base using a locally deployed Qwen-7B model [2]. All experiments were conducted using PyTorch [32] on a single NVIDIA RTX 3090 GPU. Additional implementation details are provided in the Supplementary Material.

4.1.3. Evaluation Metrics

We use three evaluation metrics to assess model performance: accuracy (ACC), area under the receiver operating characteristic curve (AUC), and F1 score. To mitigate the impact of dataset split variability in few-shot settings, we repeat each experiment ten times and report the results as mean $\pm$ standard deviation.

4.2. Main Results

We evaluate MUSE under few-shot settings with 4, 8, and 16 labeled slides per class on three histopathology datasets: CAMELYON, TCGA-NSCLC, and TCGA-BRCA. We compare against a comprehensive suite of state-of-the-art methods, including traditional MIL approaches (ABMIL [20], DSMIL [22], CLAM-SB/MB [27], TransMIL [37], WiKG [25]), vision-language models (VLMs) (Top [33], ViLa-MIL [38], FOCUS [14]), and simple aggregation baselines (mean/max pooling). All methods use CONCH [28] features for fair comparison. As reported in Table 1, MUSE achieves consistently strong performance

and establishes new state-of-the-art results across most settings. Notably, MUSE exhibits the largest gains when labeled data is most limited. In the 4-shot setting, it improves ACC over the best baseline by 6.73%, 0.76%, and 2.34% on CAMELYON, TCGA-NSCLC, and TCGA-BRCA, respectively. In the 8-shot setting, MUSE outperforms the best baseline in F1-score by 4.23%, 1.24%, and 1.60% on the same datasets. Across all benchmarks, the performance gap between MUSE and existing methods widens as the number of shots decreases. This confirms that our framework effectively leverages semantic priors to mitigate the scarcity of visual supervision, which is a key challenge in medical few-shot learning. We further observe that VLM-based methods generally surpass conventional MIL approaches under low-shot conditions, highlighting the value of semantic information in whole-slide image classification. Crucially, MUSE consistently outperforms all VLM baselines across all metrics in both 4-shot and 8-shot settings, validating the effectiveness of our proposed semantic enhancement.

4.3. Ablation Studies

4.3.1. Component Analysis

To evaluate the contribution of each module in MUSE to the overall performance, we conduct an ablation study on CAMELYON using four variants under 4, 8, and 16-shot settings. The results are reported in Table 2. BM (Base MIL) denotes classification based solely on aggregated patch features. TI (Traditional Interaction) refers to feature aggregation after cross-attention between textual prompts and patch features. SFSE and SMMO correspond to the modules shown in Figure 2 (a) and (b), which perform semantic fine-grained modeling and semantic diversity enhancement, respectively. The results show that performance improves incrementally as each component is added. Notably, the SMMO module yields substantial gains under the 4 and 8-shot settings, highlighting its critical role in low-data regimes.

4.3.2. Retrieval Strategy Ablation

To investigate the impact of retrieval strategies on performance, we evaluate three approaches: Random, L2-norm, and Cosine. Random selects text entries uniformly at random from the knowledge base. L2-norm and Cosine retrieve the nearest text entries based on Euclidean distance and cosine similarity in the feature space, respectively. Results in Table 3 show that similarity-based retrieval consistently outperforms random selection. Moreover, cosine similarity yields the best performance under both 4 and 16-shot settings, demonstrating its effectiveness for semantic alignment in few-shot WSI classification.

4.3.3. Optimization Strategy Ablation

We investigate the impact of different optimization strategies on model performance. *Multi-mean* optimizes the net-

Table 1. Few-shot weakly supervised learning results (presented in %) on CAMELYON, TCGA-NSCLC, and TCGA-BRCA under 4-shot, 8-shot, and 16-shot settings are presented. The best performance is highlighted in **bold**, and the second-best is underlined.

Dataset	Model	4-shot			8-shot			16-shot		
		ACC	AUC	F1 Score	ACC	AUC	F1 Score	ACC	AUC	F1 Score
CAMELYON	Mean Pooling	59.07 $\pm$ 3.03	51.19 $\pm$ 4.81	47.17 $\pm$ 5.19	64.05 $\pm$ 4.14	61.53 $\pm$ 5.07	55.14 $\pm$ 2.90	65.83 $\pm$ 1.96	62.50 $\pm$ 4.56	55.04 $\pm$ 7.53
	Max Pooling	62.93 $\pm$ 4.20	62.50 $\pm$ 8.88	55.58 $\pm$ 7.61	70.37 $\pm$ 5.00	74.63 $\pm$ 4.38	67.64 $\pm$ 4.44	74.46 $\pm$ 6.46	78.38 $\pm$ 8.12	72.15 $\pm$ 6.45
	ABMIL [20]	62.04 $\pm$ 2.77	58.98 $\pm$ 7.37	49.61 $\pm$ 8.85	71.52 $\pm$ 10.41	75.05 $\pm$ 11.03	69.69 $\pm$ 10.15	82.37 $\pm$ 7.10	85.48 $\pm$ 6.22	81.04 $\pm$ 7.36
	DSMIL [22]	59.44 $\pm$ 9.00	54.45 $\pm$ 10.00	53.01 $\pm$ 9.16	73.19 $\pm$ 6.93	74.63 $\pm$ 9.69	70.20 $\pm$ 8.79	84.72 $\pm$ 5.11	87.80 $\pm$ 4.69	83.60 $\pm$ 5.35
	CLAM-SB [27]	57.59 $\pm$ 6.18	55.42 $\pm$ 9.81	53.67 $\pm$ 6.09	71.67 $\pm$ 7.50	73.56 $\pm$ 10.40	67.16 $\pm$ 10.43	81.30 $\pm$ 9.97	84.65 $\pm$ 10.30	79.91 $\pm$ 10.34
	CLAM-MB [27]	60.22 $\pm$ 4.56	54.90 $\pm$ 7.69	50.38 $\pm$ 7.11	72.41 $\pm$ 10.11	73.88 $\pm$ 12.15	68.48 $\pm$ 12.69	80.03 $\pm$ 8.03	84.01 $\pm$ 9.48	78.13 $\pm$ 9.61
	TransMIL [37]	58.95 $\pm$ 4.08	52.07 $\pm$ 3.26	47.01 $\pm$ 3.34	66.91 $\pm$ 2.71	62.72 $\pm$ 7.56	55.27 $\pm$ 5.33	68.81 $\pm$ 6.13	68.24 $\pm$ 10.04	61.23 $\pm$ 11.19
	WiKG [25]	62.15 $\pm$ 3.48	56.60 $\pm$ 6.96	51.13 $\pm$ 5.53	65.98 $\pm$ 3.63	65.10 $\pm$ 6.05	57.84 $\pm$ 6.17	71.00 $\pm$ 6.90	70.73 $\pm$ 10.08	66.58 $\pm$ 8.68
	Top [33]	63.04 $\pm$ 3.99	56.39 $\pm$ 7.22	49.28 $\pm$ 7.43	66.39 $\pm$ 4.36	64.98 $\pm$ 7.13	54.85 $\pm$ 10.01	77.54 $\pm$ 9.07	76.76 $\pm$ 12.51	72.11 $\pm$ 13.26
	ViLa-MIL [38]	65.20 $\pm$ 5.30	62.96 $\pm$ 8.16	53.91 $\pm$ 11.15	76.87 $\pm$ 7.17	79.15 $\pm$ 8.19	72.79 $\pm$ 9.30	84.60 $\pm$ 4.27	88.00 $\pm$ 4.23	83.24 $\pm$ 4.55
	FOCUS [14]	68.13 $\pm$ 8.16	68.84 $\pm$ 10.32	63.40 $\pm$ 7.85	80.33 $\pm$ 8.64	83.95 $\pm$ 9.03	78.19 $\pm$ 10.51	88.62 $\pm$ 1.96	91.51 $\pm$ 2.38	87.49 $\pm$ 2.19
	MUSE	74.86 $\pm$9.22	76.65 $\pm$14.92	68.66 $\pm$15.50	84.01 $\pm$4.46	88.32 $\pm$4.29	82.42 $\pm$4.62	89.70 $\pm$1.57	92.52 $\pm$1.57	88.59 $\pm$1.91
TCGA-NSCLC	Mean Pooling	71.32 $\pm$ 7.78	77.87 $\pm$ 10.46	70.72 $\pm$ 8.55	78.13 $\pm$ 4.14	85.89 $\pm$ 4.61	78.06 $\pm$ 4.16	83.60 $\pm$ 3.97	91.44 $\pm$ 3.09	83.51 $\pm$ 4.05
	Max Pooling	68.22 $\pm$ 8.41	75.22 $\pm$ 10.51	67.54 $\pm$ 9.03	76.64 $\pm$ 6.92	83.27 $\pm$ 7.42	76.58 $\pm$ 6.93	85.28 $\pm$ 3.07	92.93 $\pm$ 2.74	85.25 $\pm$ 3.08
	ABMIL [20]	78.38 $\pm$ 8.60	84.93 $\pm$ 9.47	78.25 $\pm$ 8.68	85.88 $\pm$ 4.40	93.33 $\pm$ 3.85	85.84 $\pm$ 4.42	<u>89.24 $\pm$2.76</u>	<u>96.08 $\pm$1.31</u>	<u>89.22 $\pm$2.76</u>
	DSMIL [22]	77.31 $\pm$ 9.08	84.77 $\pm$ 10.67	77.19 $\pm$ 9.16	85.44 $\pm$ 4.90	92.33 $\pm$ 3.75	85.42 $\pm$ 4.90	89.14 $\pm$ 2.27	95.66 $\pm$ 1.58	89.12 $\pm$ 2.27
	CLAM-SB [27]	76.45 $\pm$ 8.63	84.15 $\pm$ 10.68	76.05 $\pm$ 9.36	85.28 $\pm$ 3.57	92.17 $\pm$ 3.78	85.25 $\pm$ 3.58	88.73 $\pm$ 4.12	95.38 $\pm$ 2.67	88.68 $\pm$ 4.20
	CLAM-MB [27]	78.00 $\pm$ 6.81	86.69 $\pm$ 7.49	77.73 $\pm$ 7.03	84.36 $\pm$ 3.47	91.79 $\pm$ 3.01	84.31 $\pm$ 3.50	87.87 $\pm$ 2.80	95.27 $\pm$ 1.75	87.84 $\pm$ 2.83
	TransMIL [37]	71.93 $\pm$ 8.35	78.54 $\pm$ 9.95	71.19 $\pm$ 9.54	80.34 $\pm$ 4.14	87.78 $\pm$ 4.18	80.26 $\pm$ 4.21	86.51 $\pm$ 2.85	93.98 $\pm$ 2.26	86.49 $\pm$ 2.86
	WiKG [25]	71.86 $\pm$ 6.37	80.80 $\pm$ 7.08	71.43 $\pm$ 6.82	78.48 $\pm$ 4.09	86.43 $\pm$ 4.83	78.36 $\pm$ 4.14	84.43 $\pm$ 3.80	92.24 $\pm$ 3.08	84.40 $\pm$ 3.80
	Top [33]	69.74 $\pm$ 7.97	74.41 $\pm$ 9.98	69.37 $\pm$ 8.33	81.44 $\pm$ 3.27	83.84 $\pm$ 9.17	70.29 $\pm$ 9.34	84.30 $\pm$ 5.14	91.68 $\pm$ 4.14	84.26 $\pm$ 5.15
	ViLa-MIL [38]	74.96 $\pm$ 8.36	82.80 $\pm$ 10.16	74.55 $\pm$ 8.90	83.57 $\pm$ 5.04	91.00 $\pm$ 4.97	83.49 $\pm$ 5.05	88.41 $\pm$ 3.50	95.12 $\pm$ 2.67	88.37 $\pm$ 3.56
	FOCUS [14]	79.14 $\pm$ 6.96	87.04 $\pm$ 8.06	78.96 $\pm$ 7.32	86.04 $\pm$ 3.37	93.69 $\pm$ 2.30	85.96 $\pm$ 3.48	88.38 $\pm$ 3.02	95.63 $\pm$ 1.67	88.35 $\pm$ 3.05
	MUSE	79.90 $\pm$7.23	87.57 $\pm$7.94	79.85 $\pm$7.23	87.27 $\pm$3.22	94.27 $\pm$2.59	87.20 $\pm$3.29	89.74 $\pm$2.87	96.82 $\pm$1.38	89.70 $\pm$2.90
TCGA-BRCA	Mean Pooling	76.65 $\pm$ 6.78	75.94 $\pm$ 19.57	66.90 $\pm$ 10.54	79.48 $\pm$ 5.93	82.19 $\pm$ 7.98	71.03 $\pm$ 7.89	84.63 $\pm$ 3.65	89.60 $\pm$ 2.45	77.38 $\pm$ 3.94
	Max Pooling	77.17 $\pm$ 5.12	69.55 $\pm$ 13.59	60.51 $\pm$ 8.24	76.55 $\pm$ 6.25	76.07 $\pm$ 6.41	64.63 $\pm$ 7.32	80.54 $\pm$ 5.87	84.89 $\pm$ 6.01	72.82 $\pm$ 6.50
	ABMIL [20]	79.42 $\pm$ 6.98	83.43 $\pm$ 11.34	71.75 $\pm$ 8.94	79.22 $\pm$ 4.45	85.90 $\pm$ 6.69	72.39 $\pm$ 5.24	86.39 $\pm$ 2.86	91.87 $\pm$ 2.23	80.69 $\pm$ 3.37
	DSMIL [22]	78.71 $\pm$ 11.11	83.94 $\pm$ 9.66	68.62 $\pm$ 12.23	82.63 $\pm$ 4.61	82.58 $\pm$ 10.35	72.51 $\pm$ 9.36	<u>88.03 $\pm$2.67</u>	91.05 $\pm$ 3.50	<u>81.38 $\pm$4.35</u>
	CLAM-SB [27]	79.80 $\pm$ 12.21	84.96 $\pm$ 12.08	73.40 $\pm$ 12.41	81.92 $\pm$ 5.63	87.24 $\pm$ 5.32	74.69 $\pm$ 6.44	85.85 $\pm$ 3.64	92.03 $\pm$ 1.78	80.05 $\pm$ 4.04
	CLAM-MB [27]	80.45 $\pm$ 5.14	82.09 $\pm$ 13.43	68.55 $\pm$ 13.13	80.70 $\pm$ 6.33	86.07 $\pm$ 6.84	73.43 $\pm$ 6.45	86.94 $\pm$ 2.16	92.59 $\pm$1.77	81.32 $\pm$ 2.61
	TransMIL [37]	79.93 $\pm$ 4.73	77.06 $\pm$ 10.51	63.71 $\pm$ 11.39	80.90 $\pm$ 6.88	81.07 $\pm$ 12.36	69.15 $\pm$ 11.25	83.89 $\pm$ 2.49	89.61 $\pm$ 2.92	75.88 $\pm$ 4.50
	WiKG [25]	79.83 $\pm$ 3.41	83.79 $\pm$ 6.00	67.51 $\pm$ 8.66	81.44 $\pm$ 4.28	81.28 $\pm$ 8.40	68.63 $\pm$ 10.19	85.91 $\pm$ 3.21	90.48 $\pm$ 2.62	77.10 $\pm$ 9.43
	Top [33]	80.96 $\pm$ 3.71	78.45 $\pm$ 12.48	65.73 $\pm$ 10.86	81.44 $\pm$ 3.27	83.38 $\pm$ 9.17	70.29 $\pm$ 9.34	85.65 $\pm$ 3.66	89.50 $\pm$ 3.62	75.54 $\pm$ 11.25
	ViLa-MIL [38]	<u>81.80 $\pm$4.82</u>	82.62 $\pm$ 6.54	68.90 $\pm$ 8.35	<u>83.92 $\pm$4.23</u>	84.54 $\pm$ 10.03	72.25 $\pm$ 11.97	87.78 $\pm$ 2.77	<u>92.36 $\pm$2.32</u>	81.31 $\pm$ 3.82
	FOCUS [14]	80.16 $\pm$ 8.08	82.81 $\pm$ 9.74	68.76 $\pm$ 12.63	83.02 $\pm$ 3.03	84.28 $\pm$ 8.07	68.94 $\pm$ 13.01	86.78 $\pm$ 2.52	92.14 $\pm$ 2.07	80.51 $\pm$ 2.52
	MUSE	84.14 $\pm$2.55	86.66 $\pm$8.66	73.81 $\pm$10.67	84.37 $\pm$3.97	88.39 $\pm$5.17	76.29 $\pm$6.58	88.23 $\pm$2.03	91.82 $\pm$ 2.60	81.44 $\pm$3.40

work by using the average of retrieved text features across multiple runs. *Stochastic*, as described in Section 3.3.3, optimizes the model using individual retrieved text instances in a stochastic manner. As shown in Table 4, the Stochastic strategy consistently outperforms Multi-mean across nearly all evaluation metrics, demonstrating the benefit of preserving semantic diversity during optimization.

4.3.4. Influence of LLM Choice on Knowledge Base

We study how the choice of large language model for knowledge base construction affects the performance of MUSE. Specifically, we generate text-based knowledge bases using three locally deployed large language models: Deepseek-R1-Distill-Qwen-7B [12], Llama-3.2-1B [11], and Qwen2-7B [2]. All models follow the knowledge generation pipeline described in Section 3.3.1. For Deepseek-R1-Distill-Qwen-7B, we disable reasoning traces to produce concise responses.

As shown in Table 5, MUSE achieves the best performance when using the knowledge base generated by Qwen2-7B, outperforming other variants under both 4-shot and 16-shot settings. This demonstrates that the linguistic quality and semantic richness of the generated knowledge directly influence downstream few-shot classification accuracy. The results highlight the importance of selecting capable large language models when constructing knowledge bases for medical vision-language learning.

4.4. Visualization Analysis

To investigate the alignment between fine-grained semantics and sample embeddings, we jointly map the raw semantics and fine-grained semantic priors into the unified space. The visualization in Figure 4 (a) shows that, in the latent space, the fine-grained semantic priors refined by SFSE are closer to the sample embeddings than the raw semantics. Moreover, we compare the behavior of text feature retrieval

Table 2. Ablation experiments (presented in %) of key modules of MUSE on CAMELYON under few-shot learning settings.(BM: Base MIL. TI: Traditional Interaction.)

	BM	TI	SFSE	SMMO	ACC	AUC	F1 Score
16-shot	✓	×	×	×	82.37	85.48	81.04
	✓	✓	×	×	86.61	90.13	85.41
	✓	✓	✓	×	86.72	90.20	85.33
	✓	✓	×	✓	87.99	90.95	86.70
	✓	✓	✓	✓	89.70	92.52	88.59
8-shot	✓	×	×	×	71.52	75.05	69.69
	✓	✓	×	×	79.10	82.48	77.04
	✓	✓	✓	×	80.85	84.05	78.55
	✓	✓	×	✓	83.68	88.30	82.08
	✓	✓	✓	✓	84.01	88.32	82.42
4-shot	✓	×	×	×	62.04	58.98	49.61
	✓	✓	×	×	65.13	65.84	60.40
	✓	✓	✓	×	65.16	67.30	60.43
	✓	✓	×	✓	71.18	72.75	63.09
	✓	✓	✓	✓	74.86	76.65	68.66

Table 3. Performance comparison (presented in %) of different retrieval methods in SMMO on CAMELYON.

	Retrieval Method	ACC	AUC	F1 Score
16-shot	Random	88.17	90.86	87.08
	L2-norm	87.06	90.94	85.93
	Cosine	89.70	92.52	88.59
8-shot	Random	84.20	87.66	82.70
	L2-norm	84.49	88.40	83.07
	Cosine	84.01	88.32	82.42
4-shot	Random	72.23	72.58	65.26
	L2-norm	73.38	75.51	69.75
	Cosine	74.86	76.65	68.66

Table 4. Performance comparison (presented in %) of different optimization strategies in SMMO on CAMELYON under few-shot learning settings.

	Optimization Strategy	ACC	AUC	F1 Score
16-shot	Multi-mean	87.36	90.23	86.23
	Stochastic	89.70	92.52	88.59
8-shot	Multi-mean	84.08	89.14	81.70
	Stochastic	84.01	88.32	82.42
4-shot	Multi-mean	70.33	73.53	65.01
	Stochastic	74.86	76.65	68.66

from the knowledge base across different samples, with visualization shown in Figure 4 (b). This demonstrates that our method is capable of retrieving diverse and semantically rich textual features from the knowledge base, with retrieved features exhibiting distinctiveness across different

Table 5. Performance comparison (presented in %) of knowledge bases generated by different locally deployed LLMs.

	Method	ACC	AUC	F1 Score
16-shot	Deepseek-R1-Distill-Qwen-7B	87.65	90.93	86.47
	Llama-3.2-1B	87.17	90.23	85.95
	Qwen2-7B	89.70	92.52	88.59
8-shot	Deepseek-R1-Distill-Qwen-7B	84.94	89.04	83.44
	Llama-3.2-1B	84.20	87.66	82.70
	Qwen2-7B	84.01	88.32	82.42
4-shot	Deepseek-R1-Distill-Qwen-7B	72.63	75.68	66.83
	Llama-3.2-1B	70.44	74.21	66.57
	Qwen2-7B	74.86	76.65	68.66

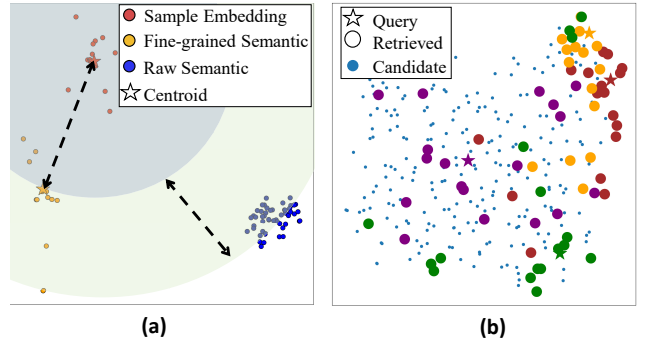


Figure 4. The visualization analysis of fine-grained semantic priors via SFSE and sample-based retrieval. (a) Visualization of fine-grained semantic in sample embedding space. (b) Comparison of text features retrieved by different samples. (Different colors represent different samples.)

samples. Additional details on visualizations and the properties of MUSE are provided in the Supplementary Material.

5. Conclusion

In this study, we propose MUSE, a stochastic multi-view semantic enhancement framework for few-shot whole slide image classification. We design Sample-wise Fine-grained Semantic Enhancement, which leverages a Mixture-of-Experts architecture to decompose category semantics and achieve fine-grained alignment between visual and textual representations. Furthermore, we introduce Stochastic Multi-view Model Optimization, which leverages the large language models to construct a multi-view textual knowledge base per pathological category and exploits the diversity of generated semantics through retrieval and stochastic optimization. Across multiple pathology benchmarks, our framework consistently demonstrates superior performance in few-shot scenarios, showing that enhanced semantic precision and enriched semantic diversity can effectively improve the model’s ability to understand and generalize complex pathological features.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62176030, in part by the Key Projects of the Special Program for Technological Innovation and Applied Development in Chongqing Municipality under Grant CSTB2023TIADKPX0060, and in part by the Fundamental Research Funds for the Central Universities under Grant 2023CDJYGRH-YB18.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 6
- [2] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 6, 7
- [3] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermesen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE transactions on medical imaging*, 38(2): 550–560, 2018. 3, 6
- [4] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermesen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017. 3, 6
- [5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 5
- [6] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miralflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019. 3
- [7] Didem Cifci, Gregory P Veldhuizen, Sebastian Foersch, and Jakob Nikolas Kather. Ai in computational pathology of cancer: improving diagnostic workflows and clinical outcomes? *Annual Review of Cancer Biology*, 7(1):57–71, 2023. 1
- [8] Miao Cui and David Y Zhang. Artificial intelligence and computational pathology. *Laboratory Investigation*, 101(4):412–422, 2021. 1
- [9] Heng Fang, Sheng Huang, Wenhao Tang, Luwen Huangfu, and Bo Liu. Sam-mil: A spatial contextual aware multiple instance learning approach for whole slide image classification. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 6083–6092, 2024. 1, 2
- [10] Kexue Fu, Linhao Qu, Shuo Wang, Ying Xiong, Ilias Maglogiannis, Longxiang Gao, Manning Wang, et al. Fast: A dual-tier few-shot learning paradigm for whole slide image classification. *Advances in Neural Information Processing Systems*, 37:105090–105113, 2024. 2
- [11] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 7
- [12] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 7
- [13] Zixian Guo, Bowen Dong, Zhilong Ji, Jinfeng Bai, Yiwu Guo, and Wangmeng Zuo. Texts as images in prompt tuning for multi-label image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2808–2817, 2023. 3
- [14] Zhengrui Guo, Conghao Xiong, Jiabo Ma, Qichen Sun, Lishuang Feng, Jinzhao Wang, and Hao Chen. Focus: Knowledge-enhanced adaptive visual compression for few-shot whole slide image classification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15590–15600, 2025. 2, 3, 6, 7
- [15] Minghao Han, Linhao Qu, Dingkan Yang, Xukun Zhang, Xiaoying Wang, and Lihua Zhang. Mscpt: Few-shot whole slide image classification with multi-scale and context-focused prompt tuning. *IEEE Transactions on Medical Imaging*, 2025. 2
- [16] Wentai Hou, Lequan Yu, Chengxuan Lin, Helong Huang, Rongshan Yu, Jing Qin, and Liansheng Wang. H²-mil: exploring hierarchical representation with heterogeneous multiple instance learning for whole slide image analysis. In *Proceedings of the AAAI conference on artificial intelligence*, pages 933–941, 2022. 3

- [17] Sheng Huang, Jiexuan Yan, Beiyang Liu, Bo Liu, and Richang Hong. Dual-view alignment learning with hierarchical-prompt for class-imbalance multi-label image classification. *IEEE Transactions on Image Processing*, 2025. 3
- [18] Zhi Huang, Federico Bianchi, Mert Yuksekogun, Thomas J Montine, and James Zou. A visual-language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9):2307–2316, 2023. 2, 3
- [19] Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatmir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1m: One million image-text pairs for histopathology. *Advances in neural information processing systems*, 36:37995–38017, 2023. 2
- [20] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018. 1, 2, 3, 6, 7
- [21] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19113–19122, 2023. 3
- [22] Bin Li, Yin Li, and Kevin 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*, pages 14318–14328, 2021. 6, 7
- [23] Honglin Li, Chenglu Zhu, Yunlong Zhang, Yuxuan Sun, Zhongyi Shui, Wenwei Kuang, Sunyi Zheng, and Lin Yang. 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*, pages 7454–7463, 2023. 1
- [24] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 2, 3
- [25] Jiawen Li, Yuxuan Chen, Hongbo Chu, Qiehe Sun, Tian Guan, Anjia Han, and Yonghong He. Dynamic graph representation with knowledge-aware attention for histopathology whole slide image analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11323–11332, 2024. 3, 6, 7
- [26] Tiancheng Lin, Zhimiao Yu, Hongyu Hu, Yi Xu, and Chang-Wen Chen. Interventional bag multi-instance learning on whole-slide pathological images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19830–19839, 2023. 1, 2, 3
- [27] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021. 3, 6, 7
- [28] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature medicine*, 30(3):863–874, 2024. 2, 3, 6
- [29] Ramin Nakhli, Puria Azadi Moghadam, Haoyang Mi, Hossein Farahani, Alexander Baras, Blake Gilks, and Ali Bashashati. Sparse multi-modal graph transformer with shared-context processing for representation learning of giga-pixel images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11547–11557, 2023. 1
- [30] Anh-Tien Nguyen, Duy Minh Ho Nguyen, Nghiem Tuong Diep, Trung Quoc Nguyen, Nhat Ho, Jacqueline Michelle Metsch, Miriam Cindy Maurer, Daniel Sonntag, Hanibal Bohnenberger, and Anne-Christin Hauschild. Few-shot whole slide pathology classification with multi-granular vision-language models. In *ICLR 2025 Workshop on Foundation Models in the Wild*, 2025. 2
- [31] Muhammad Khalid Khan Niazi, Anil V Parwani, and Metin N Gurcan. Digital pathology and artificial intelligence. *The lancet oncology*, 20(5):e253–e261, 2019. 1
- [32] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6
- [33] Linhao Qu, Kexue Fu, Manning Wang, Zhijian Song, et al. 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, 2023. 2, 3, 6, 7
- [34] Linhao Qu, Dingkan Yang, Dan Huang, Qin hao Guo, Rongkui Luo, Shaoting Zhang, and Xiaosong Wang. Pathology-knowledge enhanced multi-instance prompt learning for few-shot whole slide image classification. In *European conference on computer vision*, pages 196–212. Springer, 2024. 2
- [35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya

- Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sas-try, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. [2](#), [3](#)
- [36] Jeongun Ryu, Aaron Valero Puche, JaeWoong Shin, Seonwook Park, Biagio Brattoli, Jinhee Lee, Wonkyung Jung, Soo Ick Cho, Kyunghyun Paeng, Chan-Young Ock, et al. Ocelot: Overlapped cell on tissue dataset for histopathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23902–23912, 2023. [1](#)
- [37] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021. [3](#), [6](#), [7](#)
- [38] Jiangbo Shi, Chen Li, Tieliang Gong, Yefeng Zheng, and Huazhu Fu. 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*, pages 11248–11258, 2024. [2](#), [3](#), [6](#), [7](#)
- [39] Anabia Sohail, Iyyakutti Iyappan Ganapathi, Basit Alawode, Sajid Javed, Mohammed Bennamoun, and Arif Mahmood. Convlm: Context-guided vision-language model for fine-grained histopathology image classification. *Information Fusion*, page 103737, 2025. [2](#)
- [40] Andrew H Song, Guillaume Jaume, Drew FK Williamson, Ming Y Lu, Anurag Vaidya, Tiffany R Miller, and Faisal Mahmood. Artificial intelligence for digital and computational pathology. *Nature Reviews Bioengineering*, 1(12):930–949, 2023. [1](#)
- [41] Wenhao Tang, Sheng Huang, Xiaoxian Zhang, Fengtao Zhou, Yi Zhang, and Bo Liu. Multiple instance learning framework with masked hard instance mining for whole slide image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4078–4087, 2023. [1](#), [2](#)
- [42] Wenhao Tang, Sheng Huang, Xiaoxian Zhang, Fengtao Zhou, Yi Zhang, and Bo Liu. Multiple instance learning framework with masked hard instance mining for whole slide image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4078–4087, 2023. [1](#)
- [43] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. Review the cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia*, 2015(1):68–77, 2015. [6](#)
- [44] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. [5](#)
- [45] Zhuoyu Wen, Shidan Wang, Donghan M Yang, Yang Xie, Mingyi Chen, Justin Bishop, and Guanghua Xiao. Deep learning in digital pathology for personalized treatment plans of cancer patients. In *Seminars in diagnostic pathology*, pages 109–119. Elsevier, 2023. [1](#), [2](#), [3](#)
- [46] Jinxi Xiang, Xiyue Wang, Xiaoming Zhang, Yinghua Xi, Feyisope Eweje, Yijiang Chen, Yuchen Li, Colin Bergstrom, Matthew Gopaulchan, Ted Kim, et al. A vision–language foundation model for precision oncology. *Nature*, 638(8051):769–778, 2025. [2](#), [3](#)
- [47] Jiexuan Yan, Sheng Huang, NanKun Mu, Luwen Huangfu, and Bo Liu. Category-prompt refined feature learning for long-tailed multi-label image classification. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2146–2155, 2024. [3](#)
- [48] Jiawen Yao, Xinliang Zhu, Jitendra Jonnagaddala, Nicholas Hawkins, and Junzhou Huang. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical image analysis*, 65:101789, 2020. [1](#), [2](#), [3](#)
- [49] Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E Coupland, and Yalin Zheng. Dtf-d-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*, pages 18802–18812, 2022. [1](#), [2](#), [3](#)
- [50] Yi Zheng, Rushin H Gindra, Emily J Green, Eric J Burks, Margrit Betke, Jennifer E Beane, and Vijaya B Kolachalama. A graph-transformer for whole slide image classification. *IEEE transactions on medical imaging*, 41(11):3003–3015, 2022. [3](#)