Generalizable Whole Slide Image Classification with Fine-Grained Visual-Semantic Interaction

Hao Li[1] **Ying Chen**[1] **Yifei Chen**[2] **Rongshan Yu**[1*]
Wenxian Yang[3] **Liansheng Wang**[1*] **Bowen Ding**[4] **Yuchen Han**[4*]
[1]School of Informatics, Xiamen University [2]Huawei [3]Aginome Scientific, Xiamen, China
[4]Department of Pathology, Shanghai Chest Hospital, Shanghai Jiao Tong University School of Medicine
{lllih, cying2023}@stu.xmu.edu.cn, {rsyu, lswang}@xmu.edu.cn,
chenyifei14@huawei.com, wx@aginome.com, dingbowenmail@126.com, ychan@cmu.edu.cn

Abstract

Whole Slide Image (WSI) classification is often formulated as a Multiple Instance Learning (MIL) problem. Recently, Vision-Language Models (VLMs) have demonstrated remarkable performance in WSI classification. However, existing methods leverage coarse-grained pathogenetic descriptions for visual representation supervision, which are insufficient to capture the complex visual appearance of pathogenetic images, hindering the generalizability of models on diverse downstream tasks. Additionally, processing high-resolution WSIs can be computationally expensive. In this paper, we propose a novel "Fine-grained Visual-Semantic Interaction" (FiVE) framework for WSI classification. It is designed to enhance the model's generalizability by leveraging the interaction between localized visual patterns and fine-grained pathological semantics. Specifically, with meticulously designed queries, we start by utilizing a large language model to extract fine-grained pathological descriptions from various non-standardized raw reports. The output descriptions are then reconstructed into fine-grained labels used for training. By introducing a Task-specific Fine-grained Semantics (TFS) module, we enable prompts to capture crucial visual information in WSIs, which enhances representation learning and augments generalization capabilities significantly. Furthermore, given that pathological visual patterns are redundantly distributed across tissue slices, we sample a subset of visual instances during training. Our method demonstrates robust generalizability and strong transferability, dominantly outperforming the counterparts on the TCGA Lung Cancer dataset with at least 9.19% higher accuracy in few-shot experiments. The code is available at: https://github.com/ls1rius/WSI_FiVE.

*Co-corresponding authors. This work was supported by National Natural Science Foundation of China (Grant No. 62371409).

1. Introduction

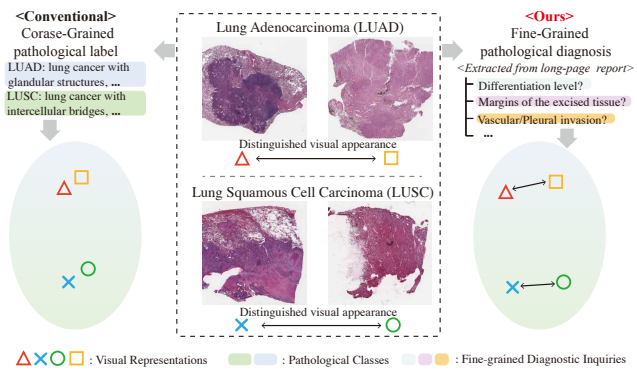

Figure 1. Challenges in WSI-text contrastive learning. Most conventional VLM approaches categorize whole slide images using category-level text descriptions, overlooking intra-class differences, leading to a decline in performance and limitations in generalization capabilities. Instead, we extract fine-grained descriptions from pathology reports as slide-level labels to develop our model, exhibiting detailed variations in each WSI.

Histological Whole Slide Image (WSI) classification plays a crucial role in computational pathology by automating disease diagnosis and subtyping. For high-resolution WSI analysis, Multiple Instance Learning (MIL) has become the dominant method. These methods treat each WSI as a "Bag" sample, and aggregate numerous patches within it as instances for thorough decision-making. Nevertheless, most existing methods [15, 16, 18, 31] focused on processing image data to conduct WSI classification, potentially not emphasizing critical pathological insights, particularly the expert textual annotations that accompany these slides.

Recently, Vision-Language Models (VLMs) [11, 25, 33] underscored the significance of integrating multimodal information for developing robust encoders. Zhang et al. [38]

exploited disease-grade text labels and extracted text insights using pre-trained language models. Qu et al. [23] utilized GPT-4 in a question-and-answer mode to obtain language prior knowledge at both instance and bag levels for VLM training. However, the challenge lies in the uniqueness and variability of content in each WSI. Existing methods developed their models with coarse-grained descriptions (*i.e.*, simplistic Category-Level text labels [38] or descriptive Category-Level text labels constructed by GPT-4 [23]), as shown in Fig. 1. They may have omitted crucial fine-grained pathological details, including differentiation level, vascular invasion, etc., which results in reduced model performance and limited generalization.

WSIs accompanied by their corresponding reports (*i.e.*, WSI-report pairs) offer detailed descriptions and fine-grained information vital for WSI analysis. Furthermore, a substantial collection of these pairs is accessible in public databases, such as The Cancer Genome Atlas (TCGA) [9]. However, their full potential has not been adequately harnessed yet. The challenge in developing a Visual Language Model (VLM) using WSI-report pairs mainly lies in the diverse formats and standards of the raw reports from different hospitals, which increases the complexity of data preprocessing and standardization processes. Additionally, pathology reports often contain extraneous information, including report metadata, tissue processing descriptions, and repetitive elements, which can introduce noise to the textual data. ***How to extract useful information from raw pathology reports to construct WSI-report pairs*** is a key issue. Moreover, recent studies have demonstrated the efficacy of prompt engineering in enhancing VLMs. In contrast to natural images, WSI data encompasses extensive professional pathological information and intricate details. ***How to craft prompts to make full use of this semantic information to guide fine-grained feature learning*** is a challenging task. Besides, the high computational costs to train models with high-resolution WSIs also limits the promotion of the model, resulting in a certain resource threshold for WSI analysis.

To address these issues, we propose a novel whole slide image classification method with **Fi**ne-grained **V**isual-s**E**mantic interaction termed as FiVE, which shows robust generalizability and efficiency in computation. Firstly, we obtain WSIs with non-standardized raw pathology reports from a public database. Collaborating with professional pathologists, we craft a set of specialized prompts to standardize reports. Following this, we employ the large language model GPT-4 to automatically clean and standardize the raw report data. In addition, we propose the **T**ask-specific **F**ine-grained **S**emantic (TFS) Module, which utilizes manual-designed prompts to direct visual attention to specific pathological areas while constructing Fine-Grained Guidance to enhance the semantic relevance of model features. Considering the diffuse distribution of pathological diseases within tissue sections and the presence of numerous non-diagnostic regions in WSIs, we also incorporate a patch sampling strategy during the training phase to enhance training efficiency and reduce computational costs. The contributions of this paper are summarized as follows:

- We pioneer the utilization of the available WSI diagnostic reports with fine-grained guidance. The obtained fine-grained description labels lead to improved supervision by discriminating the visual appearances more precisely.
- We introduce a novel Task-specific Fine-grained Semantics (TFS) Module to offer fine-grained guidance, significantly enhancing the model's generalization capabilities.
- We implement a patch sampling strategy on visual instances during training to enhance computational efficiency without significantly compromising accuracy, thereby optimizing the model's training process.

2. Related Work

2.1. Whole Slide Image Analysis

Contemporary methodologies for WSI analysis predominantly employ MIL methods where each WSI is treated as a "Bag" and its extracted patches as instances within this bag. MIL methods consists of instance-based methods [3, 14, 32] and embedding representation-based methods [15, 20, 26, 29, 36]. However, the majority of existing methods [5, 18, 24, 27] almost exclusively rely on image data, neglecting vital pathological details, notably the specialist text annotations that accompany the images. Recent works [23, 38] have taken note of this issue and started to utilize text information to improve pathological image classification. They used bag-level text labels or the descriptive labels generated by GPT. However, given the unique and varied descriptions of each WSI, their methods fall short of fully leveraging the detailed textual information present in the slides.

2.2. Vision-Language Models

Recent researches have made efforts to develop Vision-Language Models (VLMs). CLIP [25] gathered 400 million image-text pairs and initiated training with synchronized vision and text encoders from the onset. LiT [35] developed a text encoder compatible with a pre-trained vision encoder. FLIP [33] integrated region-word alignment in contrastive learning to enhance detailed vision-language correlation. Coca [34] pre-trained an image-text encoder-decoder foundation model using contrastive and captioning loss. For pathological images, some research works adapted VLMs for training with pathological images and text. Lu et al. [21] built a VLM using over 1.17 million histopathology image-caption pairs based on a task-agnostic pre-training approach derived from Coca. Huang et al. [8] curated the OpenPath

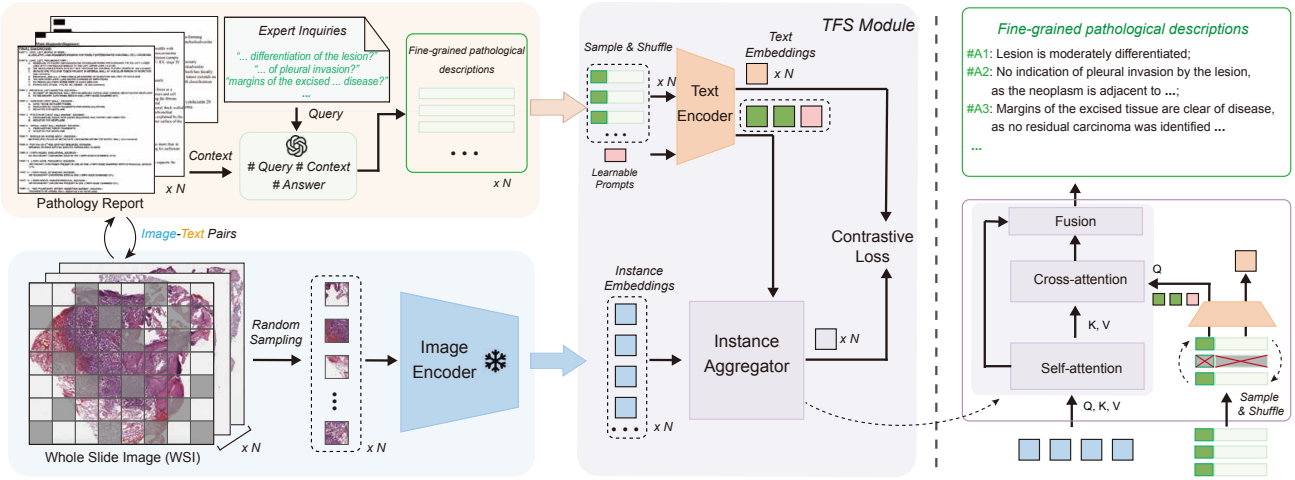

Figure 2. **Left: The structure of the FiVE framework.** The model consists of a frozen image encoder, a text encoder, and the TFS module. Whole slide images are divided into instances for embedding extraction by the image encoder. Raw pathology reports are standardized by GPT-4 into fine-grained descriptions. The fine-grained descriptions and manual prompts are sampled, shuffled, and reconstructed in pairs. These prompts aggregate instances into bag-level features, subsequently aligned with the descriptions utilizing contrastive loss. **Top Right: Fine-grained pathological descriptions.** The fine-grained pathological descriptions are generated from multiple answers based on specific queries. These descriptions undergo a process of random sampling, shuffling, and reconstruction to form a unified sentence. **Bottom Right: The Instance Aggregator module.** The instance aggregator consists of a self-attention module and a cross-attention module, fusing image instance embeddings and prompt embeddings to create bag-level features.

dataset, consisting of 208,414 pathology images with natural language descriptions from public forums, and fine-tuned a pre-trained CLIP on OpenPath. Lai et al. [13] also explored the generalization of CLIP in pathology image classification. These methods typically employ instance-level images (small patches from WSIs) and descriptions, requiring significant human and material resources. Zhang et al. [38] injected meaningful medical domain knowledge to advance pathological image embedding and classification. Qu et al. [23] employed GPT-4 to supplement image labels to enrich the information for training. However, the texts employed in their training offer only rudimentary descriptions of the images, primarily categorizing the general type of pathological slides, such as the disease category. This approach significantly constrains the model's capacity to discern fine-grained features, including the degree of differentiation, spread, and other details. Consequently, this limitation substantially restricts the model's generalizability and applicability to more nuanced diagnostic tasks.

2.3. Prompt Learning in Vision-Language Models

Drawing inspiration from prompt learning in natural language processing, some studies have proposed adapting Vision-Language models through end-to-end training of prompt tokens. CoOp [40] enhanced CLIP for few-shot transfer by optimizing a continuous array of prompt vectors within its language branch. CoCoOp [39] identified CoOp's suboptimal performance on new classes and tackled the generalization issue by conditioning prompts directly on image instances. Lu et al. [22] advocated for optimizing diverse sets of prompts by understanding their distribution. Bahng et al. [1] undertook visual prompt tuning on CLIP, focusing the prompting on the vision branch. MaPLe [12] investigated the effectiveness of multi-modal prompt learning in order to improve alignment between vision and language representations. Zhang et al. [37] adopted a set of learnable adaption prompts and prepend them to the word tokens at higher transformer layers, efficiently fine-tuning LLaMA with less cost. Furthermore, in the context of WSI classification, prompts function as valuable adjuncts, enriching contextual information and semantic interpretation. The strategic utilization of prompts substantially improved model performance [23].

3. Method

3.1. Overview

Fig. 2 shows the pipeline of our proposed FiVE method. To initiate the process, we collaborate with professional pathologists to establish a set of standards. Following this, we employ GPT-4 to automatically extract and standardize information based on these various standards. During the training phase, we construct Fine-Grained Guidance by intricately dividing and reconstructing the text description labels and manual prompts in pairs. The combination of manual prompts and learnable prompts forms the Diagnosis

Manual-Designed Standards	Fine-Grained Text Description Label Examples
1. What is the differentiation of the lesion? 2. Is there any indication of spread through air spaces around the lesion? 3. Is there any indication of vascular invasion by the lesion? 4. Is there any indication of pleural invasion by the lesion? 5. Is there any evidence of the lesion invading adjacent tissues or organs? 6. Are the margins of the excised tissue clear of disease?	**TCGA-44-6774**: Lesion differentiation is moderately to poorly differentiated; Unknown; No indication of vascular invasion by the lesion; No indication of pleural invasion by the lesion; Unknown; Margins of the excised tissue are clear of disease. **TCGA-49-4505**: Lesion differentiation is well-differentiated; Unknown; Unknown; Pleural invasion by the lesion is present, as the carcinoma extends through the visceral pleura; The lesion invades adjacent tissues or organs; Margins of the excised tissue are clear of disease.

Table 1. Manual-Designed Standards and Fine-Grained Text Description Label Examples. The answers on the right correspond to the standards on the left. "Unknown" is used as a placeholder when relevant information cannot be found.

Prompts, which are utilized to enhance the semantic relevance of the features. Subsequently, the instance aggregator module fuses instance features with fine-grained prompts, generating bag-level features, subsequently align with the corresponding fine-grained text description labels. Additionally, to reduce computational costs, we implement the patch sampling strategy, optimizing the model's training efficiency while minimizing performance loss.

3.2. Text Standardization via GPT-4

We utilize fine-grained text description labels extracted from pathology reports to align image bag-level features. Though pathology reports are readily accessible from public databases, their content exhibits significant variability depending on the source. Despite these format differences, pathology diagnoses consistently adhere to specific and well-established diagnostic standards. In our work, we develop fine-grained diagnostic criteria under the guidance of professional pathologists to standardize report data and extract fine-grained insights pertinent to pathological diagnosis, aiming to enhance its generalization capabilities substantially.

The manual-designed standards aim to extract the morphological characteristics under the microscope, such as the degree of differentiation and lesion invasion, and filter out information irrelevant to the diagnosis. Subsequently, we employ GPT-4 to automatically extract answers from the original diagnosis reports based on prompts composed of these standards. If the information queried is absent in the pathology reports, "Unknown" is used as the answer. Tab. 1 shows manual-designed standards and two fine-grained text description label examples. Then, we recombine these extracted fine-grained information and integrate it into a complete description of the case image. **More details about the prompts used for Text Standardization are provided in Supplementary Material.**

3.3. Task-specific Fine-grained Semantics Module

3.3.1 Fine-grained Guidance Construction

Due to the Text Standardization process, our data has achieved standardization. Utilizing these fine-grained text description directly as training labels can yield performance improvements. Additionally, leveraging them to generate more diverse and semantically enriched fine-grained guidance can further boost the model's performance.

During the training process we utilize these manual-designed standards as our manual-designed prompts. We divide the original fine-grained text descriptions into several parts according to manual-designed prompts, followed by random sampling and eliminating "Unknown" tags from the initial labels. Given that the staged diagnostic reports in pathological descriptions are sequence-independent, we shuffle these preliminary labels and reconstruct them into a full-sentence description. During the training, we train reconstructed text description labels and reconstructed manual prompts in pairs. For example, consider description A: ***"Lesion differentiation is moderately differentiated; Margins of the excised tissue are clear of disease."*** and description B: ***"Lesion differentiation is moderately differentiated; Margins of the excised tissue are not clear of disease, as the tumor is within the bronchial margin and parenchymal margin."***. When only the first part of each description is sampled, they would be grouped into the same category. However, when sampling the entire sections, they are considered as distinct descriptions. Changes in granularity provide diverse perspectives on the visual image, aligning visual image with text descriptions of varying granularities.

This strategy offers three key benefits: 1) Effectively alleviating the parent-child relationship in pathology categories. 2) Providing additional hierarchical semantic perspectives to enhance the text encoder's semantic comprehension ability. 3) Mitigating discrepancies in category annotation due to incomplete diagnostic information.

3.3.2 Diagnosis Prompts

We introduce Diagnosis Prompts to guide the aggregation of instance features into bag-level features. We compute the similarity between the instance features and the given manual prompts, utilizing the similarity scores as weights W for feature aggregation to improve the task-specific relevance of the features. Here we utilize the identical manual prompts as those used to standardize the raw data, as shown in the left of the Tab. 1.

In addition, manual-designed prompts may have some flaws, potentially failing to comprehensively capture the specific morphological characteristics of the lesion, and the model struggles to generalize towards unseen classes due to the late fusion through the transformer layers. Besides, fine-tuning the model may not always be feasible as it requires training a large number of parameters. Particularly in the case of low-data regimes, where the availability of training data like whole slide images is extremely limited. LLaMA-Adapter [37] and LLaMA-Adapter-v2 [6] explore the way to efficient fine-tuning of Language Models and Vision-Language Models respectively. These approaches introduced the Adaptation Prompt to gradually acquire instructional knowledge. They adopted zero-initialized attention with gating mechanisms to ensure stable training in the early stages. Inspired by these methods, we introduce learnable continuous diagnosis prompts to enrich the context information and enhance the model's transferability.

Specifically, we get the manual text prompt tokens $Q_h = [q_{h1}, q_{h2}, \cdots, q_{hn}]$, here n represents the number of the manual prompts. We concatenate the learnable continuous prompt tokens $Q_l = [q_{l1}, q_{l2}, \cdots, q_{lm}]$ on it, here m represents the number of the learnable prompt tokens. Finally we get the diagnosis prompts $Q = [q_{h1}, q_{h2}, \cdots, q_{hn}, q_{l1}, q_{l2}, \cdots, q_{lm}]$. In the training phase, part of manual prompt tokens Q_h will be sampled randomly paired with text description labels, while the whole learnable prompt tokens Q_l will be consistently retained.

Different with the traditional context learning prompts method, our approach pays attention to the acquisition of prior knowledge, similar to the methodology employed in Detection Transformer (DETR) [4]. We aim to acquire a set of appropriate query values to improve performance in subsequent feature screening processes. Additionally, it can also enable the model to quickly transfer to other tasks by fine-tuning this set of queries.

3.3.3 Instance Aggregator Module

The Instance Aggregator (IA) module is used to aggregate the fine-grained diagnosis prompts and instance features. As shown in the right of the Fig. 2, IA consists of a self-attention module and a cross-attention module.

We employ self-attention to enable feature interaction among instance features $I_i = [e_{i1}, e_{i2}, \cdots, e_{ij}]$, resulting in the feature s_i. Subsequently, utilizing the diagnosis prompts Q to aggregate the instance features and acquire the feature z_i. Then we concatenate s_i and z_i, utilizing the learnable parameter W to fuse these features, yielding the bag-level feature v_i. The formulas are shown as follows:

$$s_i = SelfAttention(I_i, I_i) + I_i \tag{1}$$

$$z_i = CrossAttention(Q, s_i) \tag{2}$$

$$v_i = concat(mean(s_i), mean(z_i)) \cdot W \tag{3}$$

Ultimately, we acquire the image bag-level features guided by the fine-grained diagnosis prompts, which are then employed to align the fine-grained text features.

3.4. Patch Sample Strategy

Each Whole Slide Image (WSI) is partitioned into a variable number of instances, ranging from approximately 50 to 45,000. Handling such a wide range of instances markedly increases computational complexity and substantially extends the training duration. FLIP [33] reduced computation and reached higher accuracy than CLIP [25] counterpart by randomly removing a large portion of image patches during training [17]. In the case of whole slide images, pathological visual patterns are often redundantly distributed across a tissue slice. Therefore, it is feasible to sample only a subset of visual instances during training.

For the instances in each bag (*i.e.*, slide), we select a sample amounting to S_m percent of the total number p of each group of instances (patches). The required number of instances S_n is described by the following formula:

$$S_n = min(p * S_m, S_{maxn}) \tag{4}$$

Here S_{maxn} denotes the maximum number of sample instances. To achieve this, we evenly divide each group of instances into S_n chunks and randomly select one instance from each chunk. Since the different whole slide images could sample different S_n, we pad the rest of the space with the same padding.

3.5. Encoder and Loss Function

We divide each WSI into instances x_k and encode these instances into embeddings $e_k \in R^D$ using pre-trained vision encoder E_{img}, composed with ResNet structure following [15], here D represents the dimension of the embeddings. Then we send the instance embeddings into the TFS Module to aggregate the instance features and prompts, and obtain the bag-level embeddings $v_i \in R^D$. The formulas are shown as follows:

$$e_k = E_{img}(x_k) \tag{5}$$

$$I_i = [e_{i1}, e_{i2}, \cdots, e_{ik}] \tag{6}$$

$$v_i = IA(I_i, Q) \tag{7}$$

Besides, we generate pathologically meaningful text embeddings for each WSI, represented as $t_i \in R^D$, by leveraging the fine-tuned text encoder BioClinicalBERT [30].

$$t_i^c = E_{txt}(x_{txt}^c) \tag{8}$$

where E_{txt} denotes the text encoder, and $x_{txt}^c (c \in [1, C])$ where C denotes the number of categories. Here we use the same embedding dimension D as the vision encoder, suitable for contrastive learning. For text encoder E_{txt}, we adopt the Low-Rank Adaptation (LoRA) [7] approach for efficient fine-tuning.

Subsequently, the bag-level embeddings v_i are aligned with the text embeddings t_i^c to complete the training process. In this case, prediction $\hat{y}$ is obtained by applying softmax on scaled cosine similarities between the image embeddings and text embeddings:

$$p(\hat{y} = c|I) = \frac{exp(sim(t_i^c, v_i)/\tau)}{\sum_{c'=1}^{C} exp(sim(t_i^{c'}, v_i)/\tau)} \tag{9}$$

where $sim(\cdot, \cdot)$ refers to cosine similarity and τ is the temperature parameter.

The fine-grained training loss is computed as the cross-entropy between the logits and soft targets as:

$$L^{v \to t} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} y_{ij} log(p_{ij}) \tag{10}$$

here N corresponds to the batch size.

Likewise, we can compute $L^{t \to v}$ and serve L as the final training objective.

$$L = \frac{L^{v \to t} + L^{t \to v}}{2} \tag{11}$$

4. Experiments and Results

4.1. Datasets

We evaluated our method on public histopathology WSI datasets: The Cancer Genome Atlas Lung (TCGA Lung) Cancer[1] and Camelyon16 [2].

TCGA Lung Cancer. The TCGA Lung Cancer dataset comprises two cancer subtypes: Lung Adenocarcinoma (LUAD) and Lung Squamous Cell Carcinoma (LUSC). It includes diagnostic slides with 541 slides from 478 LUAD cases and 512 slides from 478 LUSC cases. For WSI preprocessing, following [15], we cropped each WSI into 256 × 256 non-overlapping patches and removed the background region. The dataset encompasses approximately 5.2 million patches at 20× magnification, averaging about 5,000 patches per WSI. Following [28], we randomly split the dataset into training, validation, and testing sets with a ratio of 65:10:25 on the patient level and adopted 4-fold cross-validation. We collected corresponding pathology reports [2] and cleaned them by GPT-4 to produce fine-grained text description labels. To ensure professionalism and accuracy, we invited professional pathologists to check and correct the textual labels. Additionally, to evaluate the generalizability of the model and perform zero-shot classification, we utilized a dataset of histological subtype labels for TCGA-LUAD from the cBioPortal database[3], **More details about subtype labels are provided in Supplementary Material.**

Camelyon16. The Camelyon16 dataset [2] consists of 399 Hematoxylin and Eosin (H&E) stained slide images, utilized for metastasis detection in breast cancer. We preprocessed each WSI by segmenting it into 256 × 256 non-overlapping patches, excluding background regions. In total, this process yielded approximately 2.8 million patches at a 20× magnification level, with about 7,200 patches per bag. We adopted 3-times 3-fold cross-validation.

4.2. Implementation Details

In experiments, we employed ResNet following [15] as image encoder to extract image features, while pre-trained BioClinicalBERT from [30] as text encoder to generate text features. LoRA [7] was adopted for fine-tuning the text encoder, with an alpha value of 32 and a rank value of 8. We divided the whole slide image into patches with 256×256, then applied a random crop with size 224×224. We adopted AdamW [19] with beta (0.9, 0.98), eps 1e-8, learning rate 3e-6, warmup ratio 0.1, weight decay 1e-4 as our optimizer. Additionally, we used the batch size of 1 with an accumulation step of 8 and trained for 150 epochs. We utilized mixed-precision training on 4 NVIDIA-A800 GPUs.

4.3. Zero-Shot Histological Subtype Classification

Prior researches [23, 28, 38] have predominantly concentrated on classifying primary cancer categories. Our approach extends beyond this by attempting to classify detailed histological subtypes. It is crucial to emphasize that this task poses a significant challenge, often proving difficult for even skilled pathologists to make direct judgments.

Since only the LUAD's histological subtype dataset was provided on the online database, we conducted zero-shot subtype classification evaluation on LUAD subtype datasets, with the model being pre-trained on TCGA-LUAD or TCGA-LUSC. Throughout the training phase, subtype label information was deliberately excluded, ensuring that all experiments are conducted solely with fine-grained labels. This aims to evaluate the model's ability to identify novel diagnostic categories without specific training on subtypes. We extended the morphological appearance text de-

[1] http://www.cancer.gov/tcga

[2] https://github.com/tatonetti-lab/tcga-path-reports

[3] https://www.cbioportal.org/

scription labels of the target data using GPT-4. After the text encoder encodes the labels, similarity calculation with image features achieved zero-shot classification.

Since existing zero-shot learning methods cannot be used in WSI subtype classification, we constructed three Linear-Probe method baselines: Mean pooling, Max pooling, and Attention pooling. As shown in Tab. 2, FiVE attains 65.23% top-1 accuracy and 95.18% top-5 accuracy when pre-trained with TCGA-LUAD fine-grained labels. When pre-trained with TCGA-LUSC fine-grained labels, it achieves 62.02% top-1 accuracy and 94.36% top-5 accuracy. Moreover, the zero-shot performance of FiVE notably exceeds that of the baseline. This capability in classifying LUAD subtypes is attributed to the focus on fine-grained pathological features during training.

Method	TCGA-LUAD		TCGA-LUSC	
	Top-1	Top-5	Top-1	Top-5
Mean-pooling	40.82	83.46	31.86	82.96
Max-pooling	45.05	88.77	36.36	86.09
Attention-pooling	58.36	93.53	54.41	92.50
FiVE (Ours)	**65.23** +6.87	**95.18** +1.65	**62.02** +7.61	**94.36** +1.86

Table 2. Zero-Shot performance on histological subtype classification. We pre-trained the model using fine-grained labels of TCGA-LUAD and TCGA-LUSC, then applied zero-shot classification to the histological subtypes of TCGA-LUAD.

4.4. Few-Shot Classification

Our model demonstrates adaptability to various tasks even in scenarios with limited data availability. Few-shot experiments were conducted to demonstrate its transferability to downstream tasks. We initialized the networks with pre-trained weights derived from the model trained on TCGA image-report pairs, and subsequently fine-tuned the model on downstream datasets for few-shot image classification. We followed [23] and conducted experiments with 1, 2, 4, 8, 16, and additional 0 shot on the downstream dataset. The results are summarized in Tab. 3.

Our model exhibites remarkable performance in zero-shot classification, achieving an accuracy of 71.26%, even surpassing the SOTA method's one-shot experiment. Upon the introduction of training data, our models display exceptional transferability, outperforming the SOTA by 12.90% in the one-shot setting. The model's performance improves accordingly with the number of shot. Upon reaching 16-shot, our model reaches an impressive accuracy of 91.25%, showcasing a notable 9.19% improvement close to the fully supervised performance level.

Method	16-shot	8-shot	4-shot	2-shot	1-shot	0-shot
Mean-pool	65.33	53.89	44.85	52.93	45.34	\
Max-pool	48.48	49.55	44.22	48.39	49.03	\
Attn-pool	72.50	65.79	62.47	58.36	56.23	\
CoOp [40]	78.35	67.99	67.60	67.54	67.81	\
TOP [23]	82.06	80.51	75.41	72.38	71.01	\
FiVE (Ours)	**91.25** +9.19	**90.80** +10.29	**88.10** +12.69	**85.51** +13.13	**83.91** +12.90	**71.26**

Table 3. Few-shot classification performance on TCGA Lung Cancer. Mean-pool, Max-pool, and Attn-pool correspond to Linear-Probe implementations with Mean-pooling, Max-pooling, and Attention-pooling, respectively.

4.5. Performance Comparison with Existing Works

We compared FiVE with ABMIL [10], DSMIL [15], CLAM-SB [20], CLAM-MB [20], TransMIL [26], DTFD-MIL [36], and MHIM-MIL [28], all of which are attention-based MIL methods. In addition, we included two traditional MIL pooling operations, Max-pooling and Mean-pooling, for comparison. The results of all other methods are reproduced using the official code they provide under the same settings.

To evaluate the performance of FiVE, we conducted fine-grained pre-training exclusively on the TCGA dataset, followed by fully supervised experiments on Camelyon16 and TCGA Lung Cancer datasets for WSI classification. Note that the test data is not used in pre-training. Results are shown in Tab. 4. It can be found that our method outperforms all the other baselines by a great margin, which fully demonstrates the significance of our fine-grained training scheme in improving performance on downstream tasks. Besides, since the pre-training data mainly comes from TCGA data, on the other hand, the task-specific prompt design is more suitable for TCGA data. This results in the performance improvement of our method for TCGA data being greater than Camelyon16 data.

4.6. Ablation Studies

4.6.1 Effectiveness of TFS Module

We focused on evaluating the impact of the TFS module on the model's overall performance with the TCGA Lung Cancer. The results detailed in Tab. 5 reveal significant improvements at each stage of feature enhancement. Initially, the model with only Self Attention attained ACC of 89.77%, AUC of 92.85%, and F1-score of 89.95%. The incorporation of fine-grained labels led to increases of 1.44% in ACC, 1.51% in AUC, and 1.43% in F1-score. Subsequent integration of additional fine-grained guidance further improved performance. Ultimately, full framework with the TFS Module achieved the highest performance of 94.62% ACC, 96.33% AUC, and 93.89% F1-score. These ablation

Method	Camelyon16			TCGA Lung Cancer		
	ACC	AUC	F1-score	ACC	AUC	F1-score
Max-pooling	78.95±2.28	81.28±3.74	71.06±2.59	81.49±1.24	86.45±0.71	80.56±1.09
Mean-pooling	76.69±0.20	80.07±0.78	70.41±0.16	84.14±2.97	90.13±2.40	83.39±3.14
ABMIL [10]	90.06±0.60	94.00±0.83	87.40±1.05	88.03±2.19	93.17±2.05	87.41±2.42
DSMIL [15]	90.17±1.02	94.57±0.40	87.65±1.18	88.32±2.70	93.71±1.82	87.90±2.50
CLAM-SB [20]	90.32±0.12	94.65±0.30	87.89±0.59	87.74±2.22	93.67±1.64	87.36±2.24
CLAM-MB [20]	90.14±0.85	94.70±0.76	88.10±0.63	88.73±1.62	93.69±0.54	88.28±1.58
TransMIL [26]	89.22±2.32	93.51±2.13	85.10±4.33	87.08±1.97	92.51±1.76	86.40±2.08
DTFD-MIL [36]	90.22±0.36	95.15±0.14	87.62±0.59	88.23±2.12	93.83±1.39	87.71±2.04
MHIM-MIL [28]	92.48±0.35	96.49±0.65	90.75±0.73	89.93±3.37	95.53±1.74	89.71±2.92
FiVE (Ours)	**94.25±0.33**	**97.56±0.59**	**93.24±0.72**	**94.62±2.13**	**96.33±1.21**	**93.89±1.90**

Table 4. Comparison Performance of slide classification on Camelyon16 and TCGA Lung Cancer.

SA	FGL	FGG	LDP	ACC	AUC	F1-score
✓				89.77	92.85	89.58
✓	✓			91.21	94.36	91.01
✓	✓	✓		93.56	96.01	93.17
✓	✓	✓	✓	**94.62**	**96.33**	**93.89**

Table 5. The ablation experiments of the pre-trained model fine-tuned on the TCGA Lung Cancer. SA, FGL, FGG, and LDP represent Self Attention, Fine-Grained Labels, Fine-Grained Guidance, and Learnable Diagnosis Prompts, respectively.

experiments highlight the benefits of integrating these methods in the WSI classification task.

4.6.2 Effectiveness of Patch Sample Strategy

We verified the impact of different sampling strategies on model performance on TCGA data. Here, we assumed that there are two main indicators that affect model performance, sample ratio and max sample threshold (represented as MAXN). At the same time, in order to differentiate the experimental results, we used the unfrozen image encoder for experimental verification.

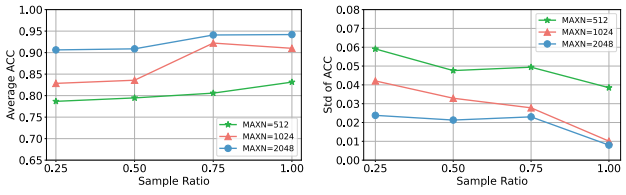

Figure 3. Classification performance on TCGA Lung Cancer with diverse sampling strategies, presenting average and standard deviation (std) ACC values.

As shown in Fig. 3, when MAXN is sufficiently large ($\geq$ 2048), the correlation between the model's performance and the sample ratio is not significant. Even with a small sample ratio ($\leq$ 0.5), the model can still effectively align with the original data distribution. Conversely, when MAXN is not large enough ($<$2048), the primary change in the model's performance depends on whether the magnitude of the sample ratio can match the original data distribution. When the sample ratio increases to 0.75, a steep performance improvement can be observed, suggesting that the sampled data at this point conforms to the original data distribution, after which the model's performance stabilizes.

Based on comprehensive experimental results, we established the sample ratio and MAXN as 0.5 and 2048 as suitable hyperparameters for the unfrozen image encoder. As for the frozen image encoder experiment, taking into account the performance constraints imposed by the frozen image encoder on the model's performance upper limit [38], we recommend the sample ratio and MAXN to be set at 0.5 and 16384 based on our experiments.

5. Conclusion

In this paper, we introduce FiVE, a novel framework that demonstrates robust generalization and strong transferability for WSI classification. Our work pioneers the utilization of non-standardized WSI-report pairs from public databases to develop a VLM. To capture the complexities and diversity within these reports, we introduce the Task-specific Fine-grained Semantics (TFS) module. This module reconstructs fine-grained labels and diagnosis prompts during training, enhancing the semantic relevance of its features by introducing diagnosis prompts. Furthermore, considering that pathological visual patterns are redundantly distributed across tissue slices, we employ a sampling strategy to reduce computational costs. Our experiments demonstrate the robust generalizability and computational efficiency of the proposed framework, which can also be easily adapted to other tasks with minimal fine-tuning. We aspire to provide empirical insights and contribute to AI pathology research.

References

[1] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Visual prompting: Modifying pixel space to adapt pre-trained models. *arXiv preprint arXiv:2203.17274*, 3:11–12, 2022. 3

[2] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, 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. 6

[3] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, 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. 2

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 5

[5] Tsai Hor Chan, Fernando Julio Cendra, Lan Ma, Guosheng Yin, and Lequan Yu. Histopathology whole slide image analysis with heterogeneous graph representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15661–15670, 2023. 2

[6] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. 5

[7] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 6

[8] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature Medicine*, pages 1–10, 2023. 2

[9] Carolyn Hutter and Jean Claude Zenklusen. The cancer genome atlas: creating lasting value beyond its data. *Cell*, 173(2):283–285, 2018. 2

[10] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018. 7, 8

[11] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 1

[12] 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

[13] Zhengfeng Lai, Zhuoheng Li, Luca Cerny Oliveira, Joohi Chauhan, Brittany N Dugger, and Chen-Nee Chuah. Clipath: Fine-tune clip with visual feature fusion for pathology image analysis towards minimizing data collection efforts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2374–2380, 2023. 3

[14] Marvin Lerousseau, Maria Vakalopoulou, Marion Classe, Julien Adam, Enzo Battistella, Alexandre Carré, Théo Estienne, Théophraste Henry, Eric Deutsch, and Nikos Paragios. Weakly supervised multiple instance learning histopathological tumor segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23*, pages 470–479. Springer, 2020. 2

[15] 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. 1, 2, 5, 6, 7, 8

[16] 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

[17] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23390–23400, 2023. 5

[18] 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

[19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6

[20] 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. 2, 7, 8

[21] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Andrew Zhang, Long Phi Le, et al. Towards a visual-language foundation model for computational pathology. *arXiv preprint arXiv:2307.12914*, 2023. 2

[22] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5206–5215, 2022. 3

[23] Linhao Qu, Xiaoyuan Luo, Kexue Fu, Manning Wang, and Zhijian Song. The rise of ai language pathologists:

Exploring two-level prompt learning for few-shot weakly-supervised whole slide image classification. *arXiv preprint arXiv:2305.17891*, 2023. 2, 3, 6, 7

[24] Linhao Qu, Zhiwei Yang, Minghong Duan, Yingfan Ma, Shuo Wang, Manning Wang, and Zhijian Song. Boosting whole slide image classification from the perspectives of distribution, correlation and magnification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21463–21473, 2023. 2

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, 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. 1, 2, 5

[26] 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. 2, 7, 8

[27] Zhuchen Shao, Yifeng Wang, Yang Chen, Hao Bian, Shaohui Liu, Haoqian Wang, and Yongbing Zhang. Lnpl-mil: Learning from noisy pseudo labels for promoting multiple instance learning in whole slide image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21495–21505, 2023. 2

[28] 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. 6, 7, 8

[29] Xinggang Wang, Yongluan Yan, Peng Tang, Xiang Bai, and Wenyu Liu. Revisiting multiple instance neural networks. *Pattern Recognition*, 74:15–24, 2018. 2

[30] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019. 6

[31] Jinxi Xiang and Jun Zhang. Exploring low-rank property in multiple instance learning for whole slide image classification. In *The Eleventh International Conference on Learning Representations*, 2022. 1

[32] Gang Xu, Zhigang Song, Zhuo Sun, Calvin Ku, Zhe Yang, Cancheng Liu, Shuhao Wang, Jianpeng Ma, and Wei Xu. Camel: A weakly supervised learning framework for histopathology image segmentation. In *Proceedings of the IEEE/CVF International Conference on computer vision*, pages 10682–10691, 2019. 2

[33] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021. 1, 2, 5

[34] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 2

[35] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022. 2

[36] Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E Coupland, and Yalin Zheng. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18802–18812, 2022. 2, 7, 8

[37] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023. 3, 5

[38] Yunkun Zhang, Jin Gao, Mu Zhou, Xiaosong Wang, Yu Qiao, Shaoting Zhang, and Dequan Wang. Text-guided foundation model adaptation for pathological image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 272–282. Springer, 2023. 1, 2, 3, 6, 8

[39] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 3

[40] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 3, 7