ViLa-MIL: Dual-scale Vision-Language Multiple Instance Learning for Whole Slide Image Classification

Jiangbo Shi[1], Chen Li[1*], Tieliang Gong[1], Yefeng Zheng[2], Huazhu Fu[3*]

[1]School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, China
[2]Jarvis Research Center, Tencent YouTu Lab, Shenzhen, China
[3]Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore

shijiangbo@stu.xjtu.edu.cn, cli@xjtu.edu.cn, hzfu@ieee.org

Abstract

Multiple instance learning (MIL)-based framework has become the mainstream for processing the whole slide image (WSI) with giga-pixel size and hierarchical image context in digital pathology. However, these methods heavily depend on a substantial number of bag-level labels and solely learn from the original slides, which are easily affected by variations in data distribution. Recently, vision language model (VLM)-based methods introduced the language prior by pre-training on large-scale pathological image-text pairs. However, the previous text prompt lacks the consideration of pathological prior knowledge, therefore does not substantially boost the model's performance. Moreover, the collection of such pairs and the pre-training process are very time-consuming and source-intensive. To solve the above problems, we propose a dual-scale vision-language multiple instance learning (ViLa-MIL) framework for whole slide image classification. Specifically, we propose a dual-scale visual descriptive text prompt based on the frozen large language model (LLM) to boost the performance of VLM effectively. To transfer the VLM to process WSI efficiently, for the image branch, we propose a prototype-guided patch decoder to aggregate the patch features progressively by grouping similar patches into the same prototype; for the text branch, we introduce a context-guided text decoder to enhance the text features by incorporating the multi-granular image contexts. Extensive studies on three multi-cancer and multi-center subtyping datasets demonstrate the superiority of ViLa-MIL.

1. Introduction

Pathological examination serves as the gold standard for cancer diagnosis [4]. With the fast development of digital scanning devices, traditional glass slides can be rapidly
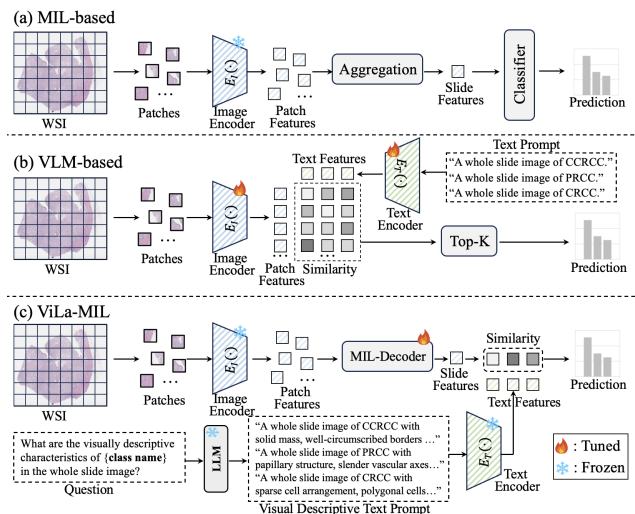

Figure 1. Comparison of our ViLa-MIL with existing MIL- and VLM-based methods. (a) MIL-based methods design various aggregation functions to generate the slide-level features; (b) VLM-based methods calculate the similarity between patches and candidate text prompts, then utilize operator like top-K to obtain the slide-level prediction; (c) Our ViLa-MIL aligns the dual-scale slide-level features and the visual descriptive text prompt to obtain the slide prediction efficiently. Note that, for simplicity, only the single-scale data stream of ViLa-MIL is visualized.

digitized into the whole slide image (WSI) [27]. WSI exhibits a hierarchical structure and huge size, which typically has about one billion pixels at its highest resolution (0.25 μm/pixel) [36]. Therefore, obtaining a substantial dataset with accurate pixel-level annotations can often be challenging and unattainable. To alleviate this problem, various weakly supervised learning methods [16, 20, 26, 32, 37] have been proposed. Among them, multiple instance learning (MIL) [9, 13, 15, 24] has gained significant popularity for tackling WSI classification tasks.

MIL utilizes small patches (*i.e.*, instances) to gener-

---
*Co-corresponding authors.

ate the slide-level (*i.e.*, bag-level) representation [25] for analysis. As shown in Figure 1(a), MIL-based methods [13,15,33] usually follow a three-step pipeline: patch cropping, feature extraction using a pretrained encoder, and slide-level feature aggregation for WSI classification. They [16,20,26,32,37] have achieved significant success in various pathological diagnostic tasks like cancer subtyping [2,16,34], staging [9,35], and tissue segmentation [18,41]. However, training these MIL-based models still heavily relies on a large number of slides with bag-level labels which are often unreachable for rare diseases. Moreover, these models only learn from the original slide, making them vulnerable to variations in data distribution and leading to suboptimal generalization performance. Recently, a series of emerging works based on visual language model (VLM) such as CLIP [30] and BLIP [17] introduced the language information, bringing new advancements to the fields of natural image classification [7,42], segmentation [21,31], object detection [40,47], etc. In the field of pathology, several VLM-based methods like MI-Zero [23], Biomed-CLIP [44], PLIP [11], and QUILT [12] have also been proposed and achieved promising results in diverse pathological diagnostic tasks and datasets. As shown in Figure 1(b), following the VLM-based framework, these methods [11,12,22,23] adopt the two-tower encoders (*i.e.*, image and text encoders), which are pre-trained on a large number of patch-text pairs collected from the Internet like Twitter or public educational resources [5,12]. Although these models have demonstrated great classification performance in various tasks, they are still subject to the following limitations:

Firstly, the text prompt does not provide effective guidance for identifying ambiguous categories, as it lacks the consideration of pathological prior knowledge. During inference, they exploit lots of handcrafted text prompt templates (*e.g.*, "An H&E image of a {class name}.") by filling in each template with the candidate category names. However, such a class-name-replacement text template only provides the class name as discriminative information, which may be insufficient for accurate classification. Introducing visual descriptive texts can help the model focus on diagnosis-related features and enhance its discriminative ability in classes with subtle differences. For example, both breast mucinous and tubular carcinoma manifest similar glandular structures. However, each subtype possesses unique features that can aid in their diagnosis. Specifically, mucinous carcinoma is characterized by mucin production, while tubular carcinoma forms small tubular structures.

Secondly, transferring the VLM-based model to the field of pathology in a parameter-efficient way is challenging. The current VLM-based methods [11,22,23] in pathology heavily rely on a large number of pathological image-text pairs, which are very time-consuming and laborious to collect. Moreover, the pre-training process requires substan-

tial computational resources and time. VLM like CLIP [30] has been pre-trained on over 400 million image-text pairs, including medical images. It may be feasible to directly apply the VLM to the field of pathology. However, given the huge size and hierarchical context of WSI, how to efficiently transfer knowledge in the VLM to process WSI is still challenging.

To address the above limitations, in this work, we propose a dual-scale vision-language multiple instance learning framework (named as ViLa-MIL) for whole slide image classification. As shown in Figure 1(c), the core idea of ViLa-MIL is to efficiently transfer the VLM to the pathology domain by better harnessing the prior knowledge of the routine pathological diagnosis. ***Firstly, we propose a dual-scale visual descriptive text prompt to boost the performance of VLM effectively***. Inspired by the routine diagnosis of pathologists, we construct our dual-scale visual descriptive text prompt, which corresponds to WSIs at different resolutions, based on the frozen large language model (LLM) [1]. Specifically, the low-scale visual descriptive text prompt mainly focuses on the global tumor structure (*e.g.*, glandular or papillary structure) presented in a WSI of low magnification (*i.e.*, low-resolution), while the high-scale visual descriptive text prompt pays more attention to local finer details (*e.g.*, clear cytoplasm and round nuclei) on a high-resolution WSI. In this way, the text prompt can guide the VLM to uncover more discriminative diagnostic-related morphological patterns from the WSI.

Secondly, we propose two lightweight and trainable decoders for image and text branches, respectively, to adapt the VLM for processing the WSI efficiently. Specifically, for the image branch, to aggregate all the patch features, we propose a prototype-guided patch decoder. For a WSI of a specific magnification, a set of learnable prototype vectors is introduced to progressively guide the fusion process of patch features by grouping similar patch features into the same prototype. In this way, each prototype captures more global contextual information for final similarity calculation. For the text branch, we propose a context-guided text decoder. By introducing the multi-granular image context (*i.e.*, local patch features and global prototype features) as the guidance, the pre-trained knowledge in the VLM text encoder can be better utilized.

Thirdly, ViLa-MIL achieves the best results compared with the current state-of-the-art MIL-based methods under the few-shot settings. In the experiment, three multi-caner and multi-center datasets are used. Specifically, ViLa-MIL outperforms the state-of-the-art methods by 1.7-7.2%, and 2.1-7.3% in area under the curve (AUC) and F1 score, respectively. Additionally, a series of ablation studies and parameter sensitivity analyses demonstrate the effectiveness of each module within ViLa-MIL.

2. Related Works

Multiple Instance Learning in WSI. Recently, MIL-based methods [9, 10, 13, 15, 23, 24, 33, 43, 46] have become the mainstream for processing the whole slide image in the field of computational pathology. The whole process mainly includes three steps: 1) a series of patches are cropped from the original WSI; 2) a pre-trained encoder is utilized to extract the patch features; 3) the patch features are aggregated to generate the final slide features. Conventional hand-crafted patch feature aggregators include non-parametric max and mean pooling operators. Later, ABMIL [13] proposes an attention-based aggregation function by learning the weight for each patch through a parameterized neural network. CLAM [24] utilizes a pre-trained image encoder for patch feature extraction and proposes a multi-branch pooling operator trained for weakly-supervised WSI classification tasks. Based on CLAM, a series of methods [15, 23, 33, 43, 46] have been proposed to explore how to aggregate the patch features effectively. For example, DSMIL [15] utilizes the multi-scale patch features as the input and aggregates it in a dual-stream architecture. TransMIL [33] explores both morphological and spatial relations through several self-attention layers. GTMIL [46] adopts a graph-based representation and vision Transformer for WSI. DT-MIL [43] utilizes a double-tier MIL framework by introducing the concept of pseudo-bags. IBMIL [20] achieves deconfounded bag-level prediction based on backdoor adjustment. Although these methods have achieved great success in many pathological diagnostic tasks, they rely solely on bag-level labels for training, thus requiring the model to learn discriminative patterns from a large quantity of WSIs, without fully utilizing pathological prior knowledge as the guideline. In this work, our ViLa-MIL introduces the dual-scale visual descriptive text prompt as the language prior to guide the training of the model effectively.

Visual Language Models in WSI. The vision language models, like CLIP [30] and FLIP [19], have shown remarkable performance in a wide variety of visual recognition tasks [7, 8, 14, 17, 21, 42]. CLIP-based methods utilize a dual-tower structure, including an image encoder and a text encoder, which are pre-trained on a large number of web-source data by aligning the image-text pairs in the same feature space. Leveraging a class-name-replacement (*e.g.*, "an image of {class name}."), CLIP classifies an image by matching it to the candidate text prompt of the highest similarity. In the field of pathology, several works (*e.g.*, MI-Zero [23] and PLIP [11]) have also been proposed by utilizing the CLIP model. For example, MI-Zero [23] aligns image and text models on pathological image patches by pre-training on 33k image-caption pairs. Later, PLIP [11] builds a large dataset of 208,414 pathological images paired with language descriptions and utilizes this dataset to fine-tune the CLIP model to the pathological field. QUILT [12] also fine-

tunes the CLIP model by creating over one million paired image-text samples based on YouTube and some public education resources. Although these works [11, 24, 45] have demonstrated significant classification performance and transferability in a new dataset, collecting a large number of image-text pairs is extremely time-consuming and labor-intensive. In the natural image field, several methods, such as CoOp [48], CLIP-Adapter [6] and TaskRes [42], have been proposed to transfer the CLIP model in a parameter-efficient fine-tuning way. However, these methods [6, 7, 42, 48] fail to take into account the hierarchical structure and large size of WSI, as well as the integration of the pathological knowledge. Recently, TOP [29] proposes a two-level prompt learning MIL framework, incorporating language prior knowledge. However, it still only utilizes the single-scale patches and the text prompt features are solely optimized based on the labels. In this work, we utilize the frozen large language model to generate the dual-scale visual descriptive text prompt, which aids the model in transferring to the target dataset with the guidance of limited labeled data. Moreover, two lightweight and trainable decoders are proposed, with one for the image branch and the other for text, to transfer the VLM model to the field of pathology efficiently.

3. Methodology

3.1. Revisiting CLIP

Many VLMs [17, 19, 30] have shown great performance in various image classification tasks. In this work, we select the CLIP model [30] as the baseline. Specifically, CLIP adopts a two-tower model structure including an image encoder and a text encoder. For each image-text pair, the image is fed into the image encoder $E_I(\cdot)$ to generate its visual features and the text features are generated by passing the text to the text encoder $E_T(\cdot)$. The contrastive loss pulls close the image and its corresponding descriptions in a unified embedding space and pushes away the features of non-paired images and texts. During inference, CLIP classifies an image into C possible categories by calculating the similarity between the image features and all the candidate text features. Formally, let I denote the image and T denote all the text templates (*e.g.*, "an image of a {class name}"). Given the image features x and text features $\{t_i\}_{i=1}^{C}$, the predicted probability for class i is defined as

$$p(y = i|x) = \frac{\exp(\cos(x, t_i))}{\sum_{j=1}^{C} \exp(\cos(x, t_j))}, \quad (1)$$

where $\cos(\cdot, \cdot)$ denotes the cosine similarity.

3.2. Dual-scale Visual Descriptive Text Prompt

To boost the CLIP model effectively, one of the most important questions is how to construct a text prompt to
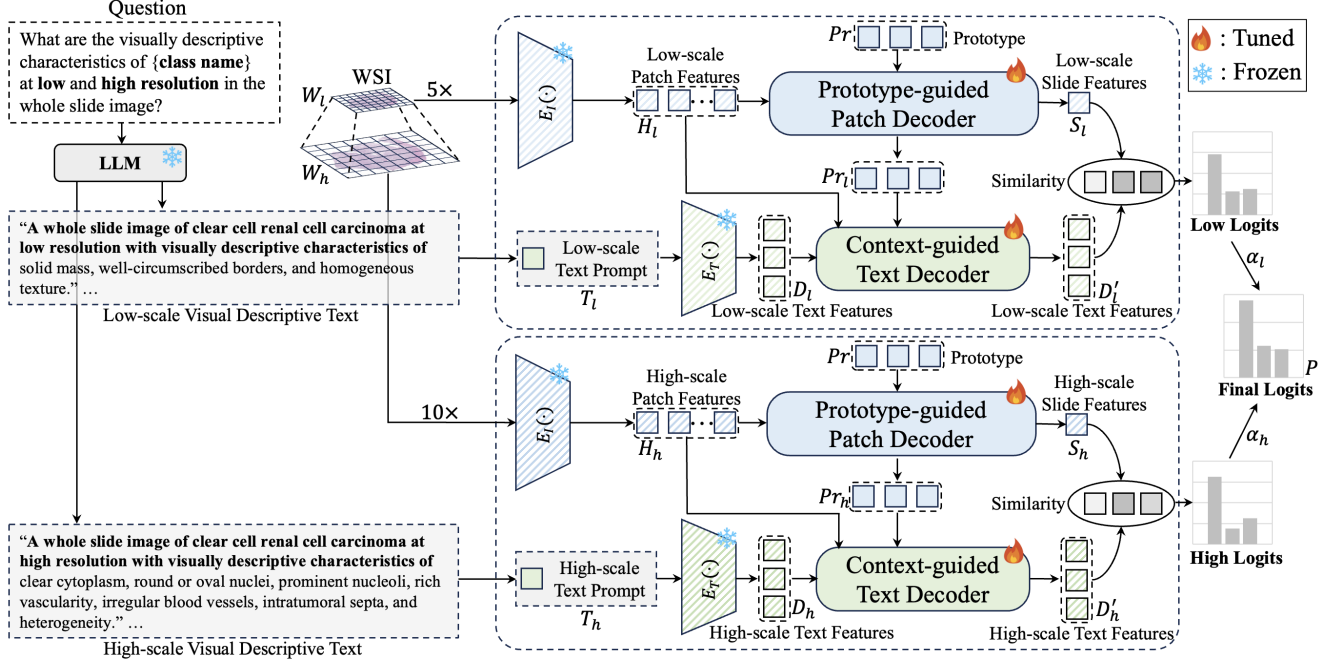
Figure 2. Pipeline of the proposed ViLa-MIL framework. The input of ViLa-MIL is a Question and WSI. The question is passed through a frozen large language model (LLM) to generate the dual-scale visual descriptive text prompt. The prototype-guided patch decoder is introduced to progressively fuse the patch features into the slide features. The context-guided text decoder is introduced to refine the text features further by utilizing the multi-granular image contexts.

describe the whole slide image. During the routine slide reviewing process, pathologists typically start by examining the overall structure of tissues at a relatively low magnification. Subsequently, they zoom in to a higher magnification to check finer details or textures, such as the size and shape of the nuclei. Combining this diagnostic prior with the multi-scale characteristics of WSI, we propose our dual-scale visual descriptive text prompt to guide the CLIP model for WSI classification, as shown in Figure 2.

The specific dual-scale visual descriptive text prompt generation process is as follows. A text question prompt is first designed for the frozen large language model: *"What are the visually descriptive characteristics of {class name} at low and high resolution in the whole slide image?"* Taking the renal cancer subtyping task as an example, by replacing the class name with the specific categories (*i.e.*, *clear cell (CCRCC), papillary (PRCC), and chromophobe renal cell carcinoma (CRCC)*), the corresponding descriptive text prompts can be generated automatically by the LLM. For example, the low-scale text prompt of PRCC mainly focuses on tissue structures present in a low-resolution WSI, like *"the papillary growth pattern"* and *"well-circumscribed borders"*. The high-scale text prompt of PRCC mainly focuses on the details in a high-resolution WSI, like *"nuclei arranged in layers"* and *"heterogeneous cytoplasm"*. These dual-scale visual descriptive texts are consistent with the daily practice of pathologists, which can

help the model distinguish subtle and fine-grained pathological morphological features and improve the model's classification performance. To better transfer the CLIP knowledge to the pathological field, inspired by the CoOp [48], we also add M learnable vectors into the dual-scale text prompts as the supplementary prompt. Formally, the dual-scale visual descriptive text prompt is denoted as T_l and T_h for low and high scales, respectively, as follows:

$$T_l = [V_l]_1[V_l]_2...[V_l]_M[Low\text{-}scale\ Text\ Prompt], \quad (2)$$

$$T_h = [V_h]_1[V_h]_2...[V_h]_M[High\text{-}scale\ Text\ Prompt], \quad (3)$$

where $[V_*]_i, i \in \{1, ..., M\}$ are the learnable vectors.

3.3. Prototype-guided Patch Decoder

To apply the CLIP model to process WSI efficiently, another important question is how to aggregate a large number of patch features. In the multiple instance learning (MIL) framework, a WSI $W = \{W_l, W_h\}$ (where W_l and W_h represent the original slides at low and high magnification, respectively.) with hierarchical structure is taken as a bag containing multiple instances as $I = \{I_l \in \mathbb{R}^{N_l \times N_0 \times N_0 \times 3}, I_h \in \mathbb{R}^{N_h \times N_0 \times N_0 \times 3}\}$ (where N_l and N_h denote the number of low- and high-resolution patches, respectively, and N_0 is the patch size). Following the current embedding-based MIL methods [13, 24], a non-overlapping sliding window method is utilized to crop patches I from the WSI. Then, the frozen CLIP image encoder $E_I(\cdot)$ is utilized to map the patches I into a feature vector $H = \{H_l \in$

$\mathbb{R}^{N_l \times d}, H_h \in \mathbb{R}^{N_h \times d}\}$ (where d is the feature dimension). To efficiently aggregate a large number of patch features into the final slide features for the calculation of similarity, we design a prototype-guided patch decoder. As shown in Figure 3(a), we randomly initialize a set of learnable prototype features $Pr \in \mathbb{R}^{N_p \times d}$ to guide the progressive fusion of patch features H_l.[1] Specifically, the defined prototypes Pr and low-scale patch features H_l are first fed into a prototype-guided attention layer, which is implemented as a cross-attention layer. Formally, we take the prototypes Pr as the query Q_l, all the low-scale patch features H_l as the key K_l and value V_l. The prototype-guided attention between the prototypes and patch features is as follows:

$$Pr_l = \text{Norm}(\text{Softmax}(\frac{Q_l K_l^\top}{\sqrt{d}})V_l) + Pr, \qquad (4)$$

where $\text{Norm}(\cdot)$ denotes the layer normalization operator and $\text{Softmax}(\cdot)$ is the activation function. By introducing the guidance of learnable prototypes, the patches with high semantic similarity will be grouped into the same prototype. Compared with the local patches with the limited receptive field, each prototype captures more global context information. Finally, to obtain the final slide-level features S_l, we utilize the attention-based feature fusion method:

$$Pr'_{l,i} = \mathbf{W_a} Pr_{l,i}, \qquad (5)$$

$$A_{l,i} = \frac{\exp\{\mathbf{W_b}^\top (\tanh(\mathbf{W_v} Pr'^\top_{l,i}))\}}{\sum_{j=1}^{N_l} \exp\{\mathbf{W_b}^\top (\tanh(\mathbf{W_v} Pr'^\top_{l,j}))\}}, \qquad (6)$$

$$S_l = \mathbf{W_c} \sum_{i=1}^{N} A_{l,i} Pr'_{l,i}, \qquad (7)$$

where $\mathbf{W_a}, \mathbf{W_c}, \mathbf{W_v}$ are trainable weight matrices with the same dimension of $d \times d$; $\mathbf{W_b} \in \mathbb{R}^{d \times 1}$ is a trainable weight matrix; $\tanh(\cdot)$ is the activation function; and A_l is the attention map that assigns the weight of each prototype in the final slide-level representation S_l. The high-scale slide-level representation S_h can be generated in a similar way.

3.4. Context-guided Text Decoder

For the text branch, the low-scale visual descriptive text prompt T_l is first fed into the frozen text encoder $E_T(\cdot)$ to generate the low-scale text features D_l. We leverage the multi-granular visual contextual information (*i.e.*, local-patch and global-prototype features) to further refine the text features. By bridging the gap between the two modalities, the model can achieve better alignment of images and texts. As shown in Figure 3(b), a context-guided attention layer is adopted to achieve this refinement process. Specifically, the low-scale prototype features Pr_l and patch features H_l are first concatenated together as the key $K_{t,l}$ and value $V_{t,l}$. The low-scale text features D_l are taken as the

---
[1]Without loss of generality, we take the low-scale (*i.e.*, low-resolution) patch feature fusion process as an example.
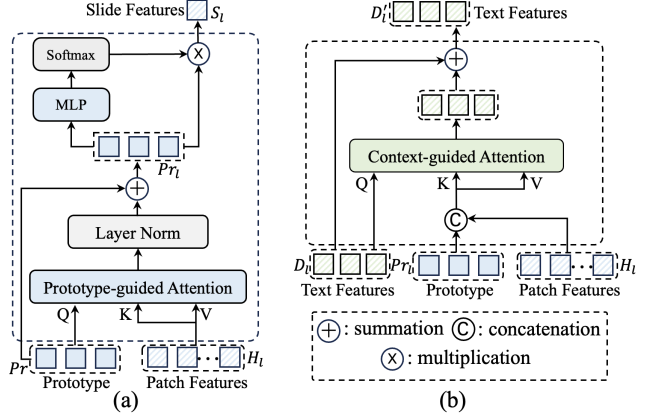

Figure 3. (a) Prototype-guided patch decoder; (b) Context-guided text decoder.

query $Q_{t,l}$. The context-guided attention is implemented with a cross-attention layer in the following manner:

$$D'_l = \text{Softmax}(\frac{Q_{t,l} K_{t,l}^\top}{\sqrt{d}})V_{t,l} + D_l. \qquad (8)$$

The updated high-scale text features D'_h of the text prompt T_h can be generated in a similar way.

3.5. Training Strategy

After obtaining the slide-level image features S_l, S_h and refined text features D'_l, D'_h for low and high scales, respectively, we can calculate the similarity between slide and text features of category i based on Eq. (1) as follows:

$$P_i = \sum_{k=\{l,h\}} \alpha_k * \frac{\exp(\cos(S_{k,i}, D'_{k,i}))}{\sum_{j=1}^{C} \exp(\cos(S_{k,i}, D'_{k,j}))}, \qquad (9)$$

where α_l and α_h control the weight of each-scale similarity. Finally, the whole model is trained end-to-end, and the cross-entropy loss is formally defined as $L = \text{CE}(P, GT)$, where $\text{CE}(\cdot, \cdot)$ denotes the cross entropy loss and GT represents the slide-level labels.

4. Experiments

4.1. Settings

Datasets. We evaluate our ViLa-MIL on three real-world WSI subtyping datasets, namely TIHD-RCC, TCGA-RCC, and TCGA-Lung,[2] under the few-shot setting. Note that TIHD-RCC is the in-house dataset collected by our team. Each dataset is split into the training, validation, and test sets with a ratio of 4:3:3. To build the few-shot setting, we then randomly select 16 samples for each category in the training set as training samples.

Implementation Details. The original WSI is processed by Otsu's binarization algorithm first to filter out all the non-tissue regions. The stain of patches is pre-processed

---
[2]https://portal.gdc.cancer.gov.

Dataset	TIHD-RCC			TCGA-RCC			TCGA-Lung		
Metric	AUC	F1	ACC	AUC	F1	ACC	AUC	F1	ACC
Max-pooling	52.5 ± 4.9	32.2 ± 3.5	37.5 ± 2.3	67.4 ± 4.9	46.7 ± 11.6	54.1 ± 4.8	53.0 ± 6.0	45.8 ± 8.9	53.3 ± 3.4
Mean-pooling	67.2 ± 4.6	42.4 ± 10.1	47.9 ± 3.9	83.3 ± 6.0	60.9 ± 8.5	62.3 ± 7.4	67.4 ± 7.2	61.1 ± 5.5	61.9 ± 5.5
ABMIL [13]	65.6 ± 9.6	46.9 ± 7.1	47.9 ± 6.7	83.6 ± 3.1	64.4 ± 4.2	65.7 ± 4.7	60.5 ± 15.9	56.8 ± 11.8	61.2 ± 6.1
CLAM-SB [24]	68.9 ± 14.4	50.5 ± 16.3	53.1 ± 13.8	90.1 ± 2.2	75.3 ± 7.4	77.6 ± 7.0	66.7 ± 13.6	59.9 ± 13.8	64.0 ± 7.7
CLAM-MB [24]	71.8 ± 9.8	54.8 ± 9.5	55.4 ± 10.2	90.9 ± 4.1	76.2 ± 4.4	78.6 ± 4.9	68.8 ± 12.5	60.3 ± 11.1	63.0 ± 9.3
TransMIL [33]	71.2 ± 4.2	53.9 ± 3.9	54.9 ± 4.1	89.4 ± 5.6	73.0 ± 7.8	75.3 ± 7.2	64.2 ± 8.5	57.5 ± 6.4	59.7 ± 5.4
DSMIL [15]	69.9 ± 9.4	49.9 ± 6.9	50.0 ± 7.4	87.6 ± 4.5	71.5 ± 6.6	72.8 ± 6.4	67.9 ± 8.0	61.0 ± 7.0	61.3 ± 7.0
GTMIL [46]	77.1 ± 3.5	61.4 ± 4.3	62.2 ± 4.3	88.1 ± 13.3	71.1 ± 15.7	76.1 ± 12.9	66.0 ± 15.3	61.1 ± 12.3	63.8 ± 9.9
DTMIL [43]	70.3 ± 10.3	55.6 ± 9.6	55.8 ± 9.9	90.0 ± 4.6	74.4 ± 5.3	76.8 ± 5.2	67.5 ± 10.3	57.3 ± 11.3	$66.6 \pm 7.5^*$
IBMIL [20]	71.5 ± 6.3	57.2 ± 7.6	56.4 ± 4.8	90.5 ± 4.1	75.1 ± 5.2	77.2 ± 4.2	69.2 ± 7.4	57.4 ± 8.3	$66.9 \pm 6.5^*$
ViLa-MIL Low	79.5 ± 3.6	61.2 ± 7.1	63.1 ± 3.0	90.9 ± 2.9	77.3 ± 4.0	79.5 ± 3.7	71.9 ± 6.2	64.1 ± 7.8	65.8 ± 5.8
ViLa-MIL High	83.0 ± 5.6	65.3 ± 6.4	66.2 ± 6.8	92.0 ± 1.3	77.8 ± 3.7	80.0 ± 3.0	72.3 ± 7.6	66.6 ± 6.5	66.9 ± 6.3
ViLa-MIL	$\mathbf{84.3 \pm 4.6}$	$\mathbf{68.7 \pm 7.3}$	$\mathbf{68.8 \pm 7.3}$	$\mathbf{92.6 \pm 3.0}$	$\mathbf{78.3 \pm 6.9}$	$\mathbf{80.3 \pm 6.2}$	$\mathbf{74.7 \pm 3.5}$	$\mathbf{67.0 \pm 4.9}$	$\mathbf{67.7 \pm 4.4}$

Table 1. Results (presented in %) on three datasets under the 16-shot setting. The top result is in bold, the second best result is underlined, and the comparable performance is denoted by superscript * based on a paired t-test (p-value > 0.05).

with the z-score normalization. The cropped patch size is 256×256 pixels (*i.e.*, $N_0 = 256$). GPT-3.5 is taken as the frozen large language model (LLM). ResNet-50 is taken as the image encoder $E_I(\cdot)$, and the corresponding CLIP Transformer is used as the text encoder $E_T(\cdot)$. The feature dimension d is 1024. The number of prototype vectors Pr is 16 and the number of context tokens M is 16. The regularization parameters α_l and α_h are both 1. The code and datasets are available at https://github.com/Jiangbo-Shi/ViLa-MIL.

Evaluation Metrics. The area under the curve (AUC) score, F1 score (F1), and accuracy (ACC) are utilized as the evaluation metrics in the experiment. We conduct the experiment for each method five times. Specifically, for each run, we randomly split the dataset according to the prescribed proportion and then select 16 samples for each category in the training set to train the model. The mean and standard deviation values are calculated based on the five results. The paired t-test as a statistical method is also adopted to determine whether the two paired results are comparable. When the p-value is larger than 0.05, these two results are statistically comparable.

4.2. Comparisons with State-of-the-Art

We compare our ViLa-MIL with state-of-the-art pathological image classification methods, including Max-pooling, Mean-pooling, ABMIL [13], CLAM [24], TransMIL [33], DSMIL [15], GTMIL [46], DTMIL [43], and IBMIL [20]. DSMIL [15] and our ViLa-MIL utilize the patches cropped from $5\times$ and $10\times$ slides. The other methods are single-scale approaches and use the patches cropped from $10\times$ slides. We also compare with two single-scale variants of our method: ViLa-MIL Low, which utilizes $5\times$ slide and low-scale text prompt, and ViLa-MIL High, which utilizes the $10\times$ slide and high-scale text prompt. All the comparison methods utilize the ResNet50-based frozen CLIP image encoder to extract the patch features to ensure a fair comparison.
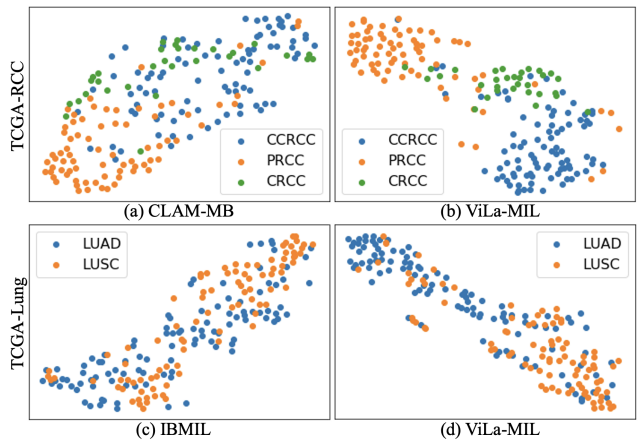

Figure 4. Slide-level feature clustering results of different methods on the TCGA-RCC (top) and TCGA-Lung (bottom) datasets.

The experiment results are reported in Table 1. Compared with baselines (*i.e.*, ABMIL), our ViLa-MIL achieves significant performance improvement across all metrics on all three datasets. Moreover, compared with the current best MIL-based methods, ViLa-MIL achieves an improvement of 1.7-7.2% in AUC, 2.1-7.3% in F1, and 0.8-6.6% in ACC across three datasets. Traditional MIL-based methods rely on a large number of slides with bag-level labels to train the model. This training paradigm is unable to learn discriminative visual features under the few-shot setting. In this work, even though the training data is limited, our ViLa-MIL can still learn discriminative morphological patterns from the WSI. Specifically, with the guidance of the visual descriptive text prompt, the single-scale ViLa-MILs already demonstrate promising performance, while the performance is boosted further with the dual-scale text prompt and image features. Moreover, the proposed two decoders transfer the CLIP knowledge efficiently for the WSI subtyping task, which further improves the model's classification performance.

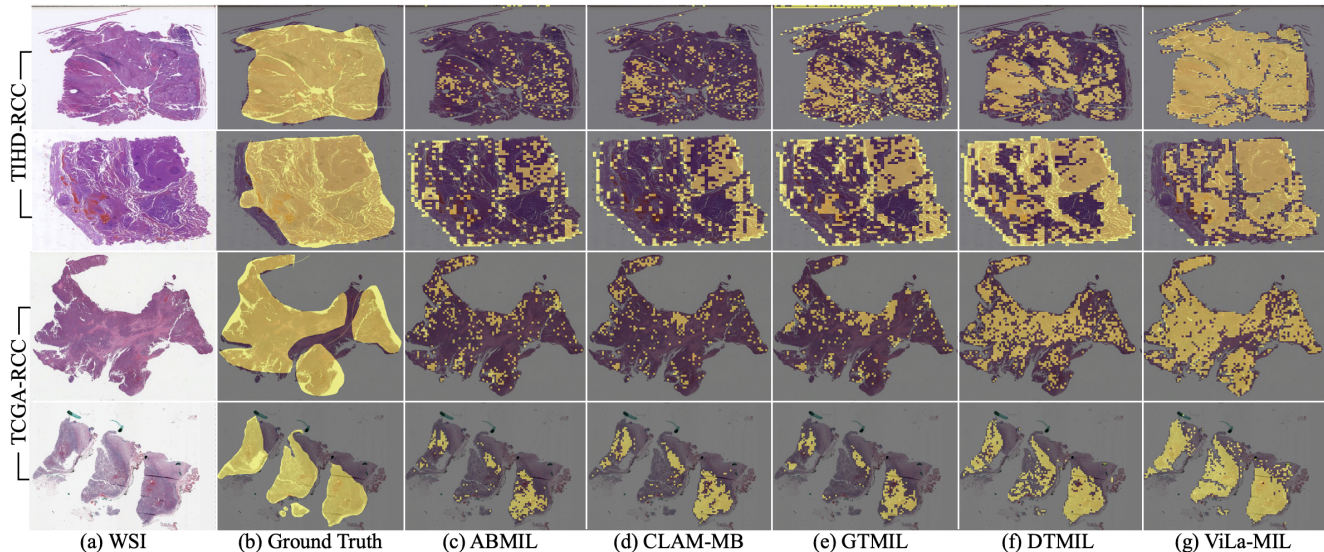In Figure 4, we visualize the slide-level feature cluster-

Figure 5. Interpretability analysis (yellow for cancer) of several exemplars from the TIHD-RCC and TCGA-RCC datasets.

Setting	TIHD → TCGA			TCGA → TIHD		
Metric	AUC	F1	ACC	AUC	F1	ACC
Mean	61.9±9.2	31.9±7.7	24.6±3.6	69.1±5.9	43.9±7.1	34.0±6.5
ABMIL	61.6±7.0	26.9±2.7	24.7±3.3	66.8±6.1	43.1±6.9	34.1±8.6
CLAM-MB	77.1±6.9	36.1±2.0	35.3±2.6	68.2±8.7	46.9±8.2	35.0±6.6
TransMIL	71.1±5.9	50.3±6.3	47.3±4.8	63.2±7.9	44.6±7.5	30.9±8.4
DSMIL	74.3±7.3	24.6±4.6	21.9±6.7	68.1±6.0	42.1±6.0	32.3±6.3
GTMIL	76.7±9.2	46.7±4.8	43.1±8.6	72.7±9.7	51.0±7.4*	37.4±5.7*
DTMIL	77.9±5.6	30.0±6.8	30.9±5.7	69.8±8.3	38.0±6.6	35.4±7.0
IBMIL	78.2±4.7	37.0±6.5	36.9±4.6	70.1±7.4	47.0±6.4	35.5±7.2
ViLa-MIL	**83.4±3.3**	**51.7±5.6**	**50.3±5.8**	**75.4±4.9**	**51.4±4.7**	**38.3±5.4**

Table 2. Results (presented in %) of multi-center cross-evaluation between TIHD-RCC and TCGA-RCC for renal cell carcinoma subtyping under the 16-shot setting. The best result is in bold and its comparable performance is denoted by superscript * based on a paired t-text (p-value > 0.05).

ing results by t-SNE [39] on the validation sets of TCGA-RCC and TCGA-Lung. Compared with the best MIL-based methods, ViLa-MIL learns an embedding space that exhibits a higher level of intra-class compactness and inter-class separability, which further demonstrates its superiority. In Figure 5, we also visualize qualitative prediction results for further comparison. Compared with several current best methods, ViLa-MIL demonstrates better localization of the cancer region.

4.3. Generalization on Domain Shift

To verify the generalization capabilities of ViLa-MIL, we report the cross-evaluation results for renal cell carcinoma between TIHD-RCC and TCGA-RCC. As shown in Table 2, the first two columns denote the results of training on TIHD-RCC and validating on TCGA-RCC; the last two columns show the results of exchanging the training and validation datasets. Due to the differences in data distribution, the performance of all methods has decreased. ViLa-MIL still achieves an improvement of 5.5% in the

AUC metric compared with the best baseline, indicating its superior cross-domain adaptability and robustness. Although the data distribution may vary across different institutions, the diagnostic criteria and visual description features remain constant. Our dual-scale visual descriptive text prompt contains comprehensive pathological diagnostic information, ensuring ViLa-MIL to have a better generalization ability across multiple centers.

4.4. Ablation Studies

Effect of Each Module in ViLa-MIL. To evaluate the effect of each module in ViLa-MIL, we perform an ablation study on the TIHD-RCC dataset under the 16-shot setting. For more details on module ablation settings, please refer to Supplementary Material. The experiment results are reported in Table 3. Specifically, the first two rows are the baseline results that utilize the single-scale text prompt and attention-based patch aggregation method (*i.e.*, ABMIL [13]). After replacing the ABMIL [13] with our proposed prototype-guided patch decoder (*i.e.*, the third and fourth rows), all three metrics have increased, which indicates that our proposed patch decoder is more efficient in aggregating these large number of patch features by grouping similar patch features into the same prototype and progressively generating the final slide features. After introducing our dual-scale text prompt (*i.e.*, the fifth row), we observe a sensible improvement for all three metrics. This indicates that our dual-scale visual descriptive text prompt can boost the WSI classification performance by effectively utilizing the complementary image features at multiple magnifications. The last row is the results of our ViLa-MIL, by introducing the context-guided text decoder. The model's performance also shows a certain amount of improvement, which denotes that the multi-granular patch and prototype features can re-

Method	AUC	F1	ACC
ABMIL + Low-scale	76.8 ± 3.1	57.2 ± 3.9	61.4 ± 2.8
ABMIL + High-scale	79.3 ± 3.2	62.7 ± 3.6	63.3 ± 3.1
Patch Decoder + Low-scale	79.4 ± 2.1	60.8 ± 5.1	62.8 ± 3.9
Patch Decoder + High-scale	82.9 ± 2.6	64.8 ± 5.1	65.6 ± 4.7
Patch Decoder + Dual-scale	83.6 ± 2.7	67.8 ± 4.5	68.3 ± 4.1
ViLa-MIL	**84.3 ± 4.6**	**68.7 ± 7.3**	**68.8 ± 7.3**

Table 3. Ablation experiment (presented in %) on the TIHD-RCC dataset under 16-shot setting.

Metric	AUC	F1	ACC
Mean Pooling	80.8 ± 4.5	66.6 ± 5.2	67.6 ± 4.9
Attention-based Pooling	81.1 ± 2.9	62.9 ± 7.5	64.9 ± 4.4
Self-attention-based Pooling	80.1 ± 3.2	62.3 ± 4.2	62.8 ± 3.9
ViLa-MIL	**84.3 ± 4.6**	**68.7 ± 7.3**	**68.8 ± 7.3**

Table 4. Results (presented in %) of different patch feature aggregation methods on the TIHD-RCC dataset under 16-shot setting.

fine the text features further.

Effect of Patch Aggregation. To verify the effect of patch aggregation methods, we compare our prototype-guided patch decoder with commonly used feature aggregation methods (*i.e.*, mean pooling, attention-based pooling, and self-attention-based pooling). Specifically, we replace our proposed patch decoder with these three methods, respectively. As shown in Table 4, our decoder achieves the best results under all three metrics on the TIHD-RCC dataset, which means that our progressive aggregation method is more effective by grouping similar patch features into the same prototype, then fusing all the prototypes to generate the final slide features.

Effect of Text Prompt. To verify the effect of different text prompts, we compare three text prompts (*i.e.*, "Class-name-replacement", "Diagnostic Guidelines", and "Large Language Model"). Specifically, the "Class-name-replacement" just replaces the class name in the text prompt "A WSI of {class name}"; "Diagnostic Guideline" utilizes the guideline of "WHO Classification of Tumours" for each subtyping, which is also a handcrafted template with clinical prior; "Large Language Model" adopts the large language model (*i.e.*, GPT-3.5) to generate a visual descriptive text prompt but does not differentiate between multi-scale visual features. As shown in Table 5, compared with the "Class-name-replacement", the "Diagnostic Guideline" achieves better performance by introducing more diagnostic-related texts. Our ViLa-MIL significantly outperforms all the other methods, indicating its capability to learn more discriminative visual features and improve the model's classification ability for challenging and ambiguous samples.

Effect of Large Language Model. To verify the effect of different large language models, we compare our adopted GPT-3.5 with three current popular LLMs (*i.e.*, PaLM-2 [3], LLaMA-2 [38] and GPT-4 [28]). These LLMs utilize the same question prompt to generate the dual-scale visual de-

Method	AUC	F1	ACC
Class-name-replacement	66.4 ± 2.5	40.0 ± 4.4	44.3 ± 3.0
Diagnostic Guideline	78.0 ± 2.8	59.5 ± 2.2	61.1 ± 1.2
Large Language Model	80.5 ± 1.5	63.3 ± 4.5	64.6 ± 4.8
ViLa-MIL	**84.3 ± 4.6**	**68.7 ± 7.3**	**68.8 ± 7.3**

Table 5. Results (presented in %) of different text prompts on the TIHD-RCC dataset under 16-shot setting.

Method	AUC	F1	ACC
PaLM-2 [3]	83.5 ± 3.6	65.8 ± 3.6	66.3 ± 3.9
LLaMA-2 [38]	81.5 ± 3.3	66.2 ± 7.3	66.8 ± 3.5
GPT-3.5 [1]	84.3 ± 4.6	**68.7 ± 7.3**	68.8 ± 7.3
GPT-4 [28]	**85.6 ± 2.5**	68.0 ± 4.5	**69.2 ± 4.8**

Table 6. Results (presented in %) of different large language models on the TIHD-RCC dataset under 16-shot setting.

scriptive text prompt like GPT-3.5. The results are reported in Table 6. All variants significantly outperform other baselines and superior performances have been achieved with different LLMs, indicating high robustness of ViLa-MIL to different kinds of frozen LLMs. Compared with GPT-3.5, GPT-4 achieves superior performances on metrics of AUC and ACC. This implies that the LLM with enhanced language expression capabilities has the potential to further improve the performance of ViLa-MIL.

5. Conclusion

In this work, we proposed a dual-scale vision-language multiple instance learning framework (ViLa-MIL) for whole slide image classification. Specifically, inspired by the diagnostic process of pathologists, we utilized the frozen LLM to generate dual-scale visual descriptive text prompts, which correspond to the hierarchical image contexts in WSIs. To transfer the VLM to process the WSI efficiently, a prototype-guided patch decoder was proposed to progressively aggregate the patch features. A context-guided text decoder was also proposed to refine the text prompt features further by utilizing the multi-granular image contexts. Extensive comparative and ablation experiments on three cancer subtyping datasets demonstrated that ViLa-MIL achieved state-of-the-art results for whole slide image classification. We believe that this work will inspire further research by introducing language prior information to efficiently transfer knowledge from large pre-trained models to the field of pathology.

Acknowledgements

References

[1] 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. 2, 8

[2] 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

[3] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. PaLM: Scaling language modeling with Pathways. *arXiv preprint arXiv:2204.02311*, 2022. 8

[4] Miao Cui and David Y Zhang. Artificial intelligence and computational pathology. *Laboratory Investigation*, 101(4):412–422, 2021. 1

[5] Jevgenij Gamper and Nasir Rajpoot. Multiple instance captioning: Learning representations from histopathology textbooks and articles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16549–16559, 2021. 2

[6] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. CLIP-Adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, pages 1–15, 2023. 3

[7] Zixian Guo, Bowen Dong, Zhilong Ji, Jinfeng Bai, Yiwen 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. 2, 3

[8] Cheng Han, Qifan Wang, Yiming Cui, Zhiwen Cao, Wenguan Wang, Siyuan Qi, and Dongfang Liu. E^{2}VPT: An effective and efficient approach for visual prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17491–17502, 2023. 3

[9] Wentai Hou, Lequan Yu, Chengxuan Lin, Helong Huang, Rongshan Yu, Jing Qin, and Liansheng Wang. H^2-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. 1, 2, 3

[10] Yanyan Huang, Weiqin Zhao, Shujun Wang, Yu Fu, Yuming Jiang, and Lequan Yu. ConSlide: Asynchronous hierarchical interaction Transformer with breakup-reorganize rehearsal for continual whole slide image analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21349–21360, 2023. 3

[11] 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, 3

[12] Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir 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, 2024. 2, 3

[13] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the International Conference on Machine Learning*, pages 2127–2136, 2018. 1, 2, 3, 4, 6, 7

[14] 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

[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, 3, 6

[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, 2

[17] 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

[18] Kailu Li, Ziniu Qian, Yingnan Han, I Eric, Chao Chang, Bingzheng Wei, Maode Lai, Jing Liao, Yubo Fan, and Yan Xu. Weakly supervised histopathology image segmentation with self-attention. *Medical Image Analysis*, 86:102791, 2023. 2

[19] 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. 3

[20] 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, 6

[21] Yuqi Lin, Minghao Chen, Wenxiao Wang, Boxi Wu, Ke Li, Binbin Lin, Haifeng Liu, and Xiaofei He. CLIP is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15305–15314, 2023. 2, 3

[22] 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

[23] Ming Y Lu, Bowen Chen, Andrew Zhang, Drew FK Williamson, Richard J Chen, Tong Ding, Long Phi Le, Yung-

Sung Chuang, and Faisal Mahmood. Visual language pre-trained multiple instance zero-shot transfer for histopathology images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19764–19775, 2023. 2, 3

[24] 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. 1, 3, 4, 6

[25] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *Advances in Neural Information Processing Systems*, 10:570–576, 1997. 2

[26] 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. pages 11547–11557, 2023. 1, 2

[27] 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

[28] OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 8

[29] 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, 2024. 3

[30] 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, 2021. 2, 3

[31] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. DenseCLIP: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18082–18091, 2022. 2

[32] 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, 2

[33] 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, 3, 6

[34] Jiangbo Shi, Lufei Tang, Zeyu Gao, Yang Li, Chunbao Wang, Tieliang Gong, Chen Li, and Huazhu Fu. MG-Trans: Multi-scale graph transformer with information bottleneck for whole slide image classification. *IEEE Transactions on Medical Imaging*, 42(12):3871–3883, 2023. 2

[35] Jiangbo Shi, Lufei Tang, Yang Li, Xianli Zhang, Zeyu Gao, Yefeng Zheng, Chunbao Wang, Tieliang Gong, and Chen Li. A structure-aware hierarchical graph-based multiple instance learning framework for pT staging in histopathological image. *IEEE Transactions on Medical Imaging*, 42(10):3000–3011, 2023. 2

[36] 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*, pages 1–20, 2023. 1

[37] 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

[38] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 8

[39] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11):2579–2605, 2008. 7

[40] Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. CORA: Adapting CLIP for open-vocabulary detection with region prompting and anchor pre-matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7031–7040, 2023. 2

[41] 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

[42] Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang. Task residual for tuning vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10899–10909, 2023. 2, 3

[43] 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. 3, 6

[44] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, et al. Large-scale domain-specific pre-training for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*, 2023. 2

[45] 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. 3

[46] 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, 6

[47] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. RegionCLIP: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022. 2

[48] 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, 4