Queryable Prototype Multiple Instance Learning with Vision-Language Models for Incremental Whole Slide Image Classification

Jiaxiang Gou, Luping Ji*, Pei Liu, Mao Ye

School of Computer Science and Engineering, University of Electronic Science and Technology of China, China
goujiaxiang@std.uestc.edu.cn, jiluping@uestc.edu.cn, yuukilp@std.uestc.edu.cn, maoye@uestc.edu.cn

Abstract

Whole Slide Image (WSI) classification has very significant applications in clinical pathology, *e.g.*, tumor identification and cancer diagnosis. Currently, most research attention is focused on Multiple Instance Learning (MIL) using *static* WSI datasets. One of the most obvious weaknesses of these methods is that they cannot efficiently preserve and utilize previously learned knowledge. With any new data arriving, classification models are required to be re-trained on both previous and current new data. To overcome this shortcoming and break through traditional vision modality, this paper proposes the first *Vision-Language*-based framework with *Queryable Prototype Multiple Instance Learning* (QPMIL-VL) specially designed for incremental WSI classification. This framework mainly consists of two information processing branches: one is for generating bag-level features by prototype-guided aggregation of instance features, while the other is for enhancing class features through a combination of class ensemble, tunable vector and class similarity loss. The experiments on four public WSI datasets demonstrate that our QPMIL-VL framework is effective for incremental WSI classification and often significantly outperforms other compared methods, achieving state-of-the-art (SOTA) performance.

Code — https://github.com/can-can-ya/QPMIL-VL

Introduction

Histopathology Whole Slide Image (WSI) is crucial for the diagnosis and treatment of tumor diseases (Song et al. 2023). To model gigapixel WSIs (*e.g.*, $90,000 \times 90,000$ pixels) for training clinical-grade models, many methods based on Multiple Instance Learning (MIL) have been extensively studied in the field of computational pathology (CPATH) (Campanella et al. 2019; Lu et al. 2021; Lin et al. 2023; Liu et al. 2024c). A shared characteristic of these methods is that they are specially designed for capturing a *static* data distribution (*e.g.*, a given WSI dataset with two lung cancer subtypes), still following conventional machine learning paradigm. However, in the real-world, the distribution of WSI data could be *dynamic* due to the arrival of new datasets or the discovery of emerging cancer types (Van der Laak, Litjens, and Ciompi 2021; Derakhshani et al. 2022). As a result, a well-trained WSI classification model that is suitable for previous data often cannot adapt to new data. Therefore, it must be re-trained on the entire dataset (including both previous and new data), leading to a significant increase in training costs. This presents the necessities of new methods well-tailored to dynamic data distributions.

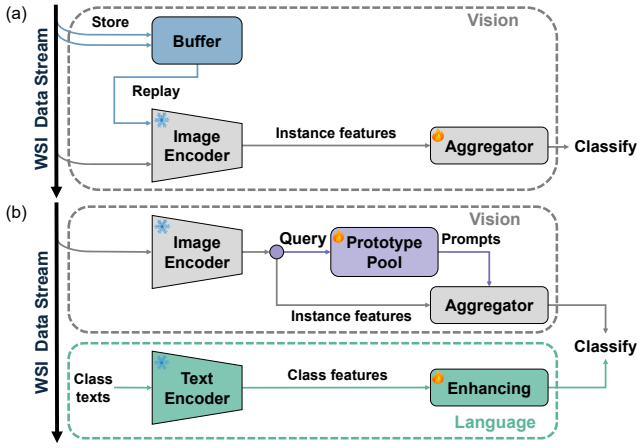

Figure 1: (a) Existing visual modality framework for incremental WSI classification with buffer dependency; (b) our proposed Vision-Language-based framework with a queryable prototype pool and class feature enhancement.

Incremental Learning (IL) (Li and Hoiem 2017), also referred to as continual learning or lifelong learning, is exactly such approach to filling the aforementioned gap. As a new learning paradigm, it can encourage models to maintain the memory stability on old data distributions while adapting to a new one (Wang et al. 2024), thus mitigating the notorious *catastrophic forgetting* problem (McCloskey and Cohen 1989). Owing to such practical benefit, this new paradigm has attracted great attention and has shown strong potential for application in the real-world (Li et al. 2019; Caccia et al. 2021; Wang, Huang, and Hong 2022). Similarly, in CPATH, IL strategies are also noticed recently and show promising performance in learning from dynamic WSI data. Concretely, ConSlide (Huang et al. 2023a) is proposed for the first time to study incremental WSI classification, as il-

*Corresponding Author.

lustrated in Fig. 1 (a). **(1)** Like most existing methods, ConSlide captures image-level features for WSI classification, relying solely on a traditional *vision modality*. **(2)** Moreover, to retain certain past data, an additional buffer is utilized to enable the model to review previous knowledge while learning new information. This may result in privacy concerns (Shokri and Shmatikov 2015) and *higher computational costs*, leaving a gap to the resource efficiency goal pursued in IL (Wang et al. 2024).

Recent works, *e.g.*, PLIP (Huang et al. 2023b), Prov-GigaPath (Xu et al. 2024) and CONCH (Lu et al. 2024), have deeply demonstrated the significant potential of Vision-Language Models (VLMs) in CPATH. Breaking through traditional vision modality, this kind of new models could effectively exploit pathology features from both *vision* and *language* modalities. All of these motivate us to leverage foundational pathology VLMs to study incremental learning for WSI classification.

To extend the pure vision frameworks, we propose a new **Vision-Language-based framework with Queryable Prototype Multiple Instance Learning** (**QPMIL-VL**) for incremental WSI classification, as shown in Fig. 1 (b). Specifically, **(1)** in the **vision** branch, inspired by L2P (Wang et al. 2022), we design a Queryable Prototype MIL (QPMIL) module comprising a prototype pool and prototype-guided aggregation. This module encourages the model to incrementally learn a set of prototypes corresponding to each dataset through a prototype-query mechanism, effectively mitigating catastrophic forgetting without relying on extra buffer. **(2)** In the **language** branch, we propose a Class Feature Enhancement (CFE) module. We employ CFE to increase the diversity of text descriptions via a class ensemble approach and further refine it with tunable vector and a class similarity loss. Our experiments show that QPMIL-VL could often surpass other state-of-the-art (SOTA) methods by a large margin in incremental WSI classification tasks. The contributions of this work are as follows:

I) To our knowledge, beyond traditional pure vision frameworks, we propose the first Vision-Language-based framework for incremental WSI classification.

II) We devise a new Queryable Prototype Multiple Instance Learning (QPMIL) strategy to alleviate catastrophic forgetting. In this strategy, instance features are matched with a set of prototypes through querying to guide the generation of WSI bag-level feature.

III) Extensive experiments are performed on four public WSI datasets. The results demonstrate the superiority of our QPMIL-VL over the existing methods.

Related Work

Multiple Instance Learning for WSI Classification

Multiple Instance Learning (MIL) focuses on learning from weakly-annotated data, where only an unknown subset of instances within each input bag is relevant to the label. Due to the unique data characteristics of WSI, the MIL paradigm is widely applied in WSI classification (Campanella et al. 2019; Li, Li, and Eliceiri 2021; Shao et al. 2021; Zhang et al. 2022; Zheng et al. 2023; Liu et al. 2024b). Among these methods, ABMIL (Ilse, Tomczak, and Welling 2018) proposes an attention-based instance feature aggregation. By leveraging ResNet for instance-level feature extraction, CLAM (Lu et al. 2021) introduces an interpretable weakly-supervised learning method, focused on data-efficient WSI processing. However, most of these studies focus on *static* MIL classification tasks and rarely consider the incremental WSI classifcation with MIL.

Incremental Learning

Recently, Incremental Learning (IL) has gained significant attention, aiming to enable deep models to continually acquire knowledge like humans. IL algorithms can be classified into three main categories. Regularization-based methods (Kirkpatrick et al. 2017; Li and Hoiem 2017) aim to reduce catastrophic forgetting by limiting changes to key parameters but often fall short of optimal results. Architecture-based methods (Rusu et al. 2016; Mallya and Lazebnik 2018; Li et al. 2019; Ke, Liu, and Huang 2020) train an independent module for each task but typically apply only to task-incremental learning scenario requiring task identity during inference. Additionally, although prevalent Rehearsal-based methods (Chaudhry et al. 2018, 2019; Prabhu, Torr, and Dokania 2020; Buzzega et al. 2020; Caccia et al. 2021) could achieve SOTA performance on various benchmarks (Parisi et al. 2019; Mai et al. 2022), these methods rely on additional buffer to store past data. When buffer size is limited, the performance of such these methods deteriorates (Cha, Lee, and Shin 2021). They will even become inapplicable when historical data is unavailable (Shokri and Shmatikov 2015). Currently, most of these studies focus more on *natural images*, not medical gigapixel WSIs, except ConSlide (Huang et al. 2023a).

Vision-Language Models

Recent research has made significant success in developing Vision-Language Models (VLMs). For example, CLIP (Radford et al. 2021) learns SOTA image representations, by the training on 400 million (image, text) pairs. Coca (Yu et al. 2022) employs contrastive and captioning losses to pre-train a foundation model comprising an image-text encoder-decoder. Additionally, some specialized VLMs like PLIP (Huang et al. 2023b), Prov-GigaPath (Xu et al. 2024) and CONCH (Lu et al. 2024) have been developed in CPATH. These VLMs need large datasets for pre-training, so they could often show excellent generalizability capabilities. Moreover, some VLMs combined with certain fine-tuning methods (Zhou et al. 2022; Yu et al. 2023; Gao et al. 2024) are widely transferred to various downstream tasks (Wang et al. 2023; Qu et al. 2024; Li et al. 2024; Liu et al. 2024a). At present, to our knowledge, the potential of VLMs for incremental WSI classification remains unexplored.

Preliminaries

Problem Formulation

Incremental WSI classification requires a model to continuously learn new knowledge from sequential tasks on non-stationary datasets, while retaining the knowledge learned

from prior tasks. We define a sequence of WSI datasets $\mathcal{D} = \{\mathcal{D}_1, \cdots, \mathcal{D}_T\}$, in which the t-th dataset $\mathcal{D}_t = \{(\boldsymbol{x}_i^t, y_i^t)\}_{i=1}^{n_t}$ contains the tuples of sample $\boldsymbol{x}_i^t$ and its corresponding bag-level label y_i^t, where $1 \leq t \leq T$ and n_t is the sample number in $\mathcal{D}_t$. When training is made on current dataset $\mathcal{D}_{t_c}$, the data from previous datasets (*i.e.*, $\mathcal{D}_1, \cdots, \mathcal{D}_{t_c-1}$) may be limited or even unavailable.

For a given WSI sample $\boldsymbol{x}_i^t$, in class-incremental learning scenario, the model only uses $\boldsymbol{x}_i^t$ to predict its bag-level label y_i^t. In contrast, in task-incremental learning scenario, the model combines sample $\boldsymbol{x}_i^t$ with its corresponding task identity t together to predict label y_i^t.

VLM Pre-training and Inference

In our work, we choose a SOTA VLM in pathology, *i.e.*, CONCH (Lu et al. 2024), as our image and text encoders.

Pre-training. CONCH uses the contrastive learning on diverse histopathology images, biomedical text, and over 1.17 million image-caption pairs (not including TCGA Datasets). Based on CoCa (Yu et al. 2022), CONCH combines an image encoder $\boldsymbol{E}_{\text{img}}$ with $f(\cdot, \theta)$, a text encoder $\boldsymbol{E}_{\text{txt}}$ with $g(\cdot, \phi)$, and a multi-modal fusion decoder.

Zero-shot Patch Inference. Since CONCH is pre-trained on patch-level WSI image-text pairs, after training is completed, zero-shot inference can be fulfilled on patch-level samples. Specifically, given a patch $\boldsymbol{p}$ and class text token set $\{\boldsymbol{t}_c\}_{c=1}^{C}$, where C denotes total class number. The classification can be conducted based on the cosine similarity between patch feature $f(\boldsymbol{p}, \theta)$ and class feature $g(\boldsymbol{t}_c, \phi)$.

Proposed Method

To effectively fulfil incremental WSI classification, on traditional MIL framework, we introduce Vision-Language Model and propose a queryable prototype MIL framework, as shown in Fig. 2. It consists of two branches, one for generating bag-level feature (QPMIL), and the other for enhancing class text feature (CFE). In inference, the features from both branches will be used to compute classification probability.

Queryable Prototype Multiple Instance Learning (QPMIL)

As usual, one of the most important objectives for WSI classification is to generate discriminative bag-level feature representation. However, unlike traditional static classification, our incremental classification task requires that all subsequent model learning specifically avoids causing catastrophic impacts on the tasks learned earlier.

Moreover, it has been proved that the instances in different WSIs, even in different WSI datasets, could often appear high similarities, in terms of cell shape, staining response, and tissue arrangement (Song et al. 2024).Therefore, similar instances, even if they are from different datasets, could be clustered into the same instance prototype. In view of this, to generate more effective WSI bag-level feature, we introduce VLM and instance-level feature prototypes, designing a queryable prototype pool and prototype-guided aggregation.

Prototype Pool. Specifically, the prototype pool consists of M (*key*, *prompt*) prototype pairs $\{(\boldsymbol{k}_i, \boldsymbol{P}_i)\}_{i=1}^{M}$ initialized using a *Normal Distribution*, $\boldsymbol{k}_i \in \mathbb{R}^{D_f}, \boldsymbol{P}_i \in \mathbb{R}^{L_{\boldsymbol{P}} \times D_e}$, where D_f is the output feature dimension of encoder $\boldsymbol{E}_{\text{img}}$ and $\boldsymbol{E}_{\text{txt}}$, D_e represents the embedding dimension of $\boldsymbol{E}_{\text{txt}}$ and $L_{\boldsymbol{P}}$ indicates the total number of learnable vectors in $\boldsymbol{P}_i$. In a prototype pair, the *key* works as a prototype identifier, used by WSIs for querying, while the *prompt* acts as a prototype descriptor, utilized to describe the specific visual feature of instance prototype within WSIs.

Given a WSI sample $\boldsymbol{x}$ from $\mathcal{D}_t$, it is first divided into n non-overlapping patches $\{\boldsymbol{p}_i\}_{i=1}^{n}$, *i.e.*, instances. Then, all instances are encoded by image encoder $\boldsymbol{E}_{\text{img}}$ with $f(\cdot, \theta)$, generating n instance features $\boldsymbol{Z} \in \mathbb{R}^{n \times D_f}$, as shown in Fig. 2. Next, we apply *MaxPooling* to features $\boldsymbol{Z}$ to obtain the query vector of sample $\boldsymbol{x}$, *i.e.*, $\boldsymbol{z} = MaxPooling(\boldsymbol{Z}) \in \mathbb{R}^{D_f}$. *MeanPooling* is also tried and its experimental result is provided in our Supplementary Material.

Subsequently, a query metric function $q : \mathbb{R}^{D_f} \times \mathbb{R}^{D_f} \rightarrow \mathbb{R}$ is defined to evaluate how close the query vector $\boldsymbol{z}$ matches the *key* of each prototype pair. We employs the cosine distance for this evaluation. Finally, we are able to find the most matched prototype *key*s through

$$\mathcal{K}_{\boldsymbol{x}} = Top\text{-}N^{min}\{q(\boldsymbol{z}, \boldsymbol{k}_i)\}_{i=1}^{M}, \tag{1}$$

where $Top\text{-}N^{min}$ denotes the operation of choosing the top-N prototype *key*s with minimal $\boldsymbol{z}$-*key* distances. However, in prototype matching on different WSI datasets during incremental training by Eq. 1, we observe that almost all samples tend to choose the same set of N prototypes from pool. This could usually lead to catastrophic forgetting in incremental learning. To effectively avoid this tendency, we define a penalty factor $p_i^{t_c}$ to optimize prototype choosing:

$$\begin{cases} \mathcal{K}_{\boldsymbol{x}} = Top\text{-}N^{min}\{q(\boldsymbol{z}, \boldsymbol{k}_i) \cdot p_i^{t_c}\}_{i=1}^{M} \\ P_{t_c} = \dfrac{1}{t_c - 1} \displaystyle\sum_{t=1}^{t_c-1} F_t, t_c \geqslant 2, \end{cases} \tag{2}$$

where t_c is the index of current WSI dataset $\mathcal{D}_{t_c}$, $F_t = [f_1^t, \cdots, f_M^t]$ denotes the matching frequency table of M prototype *key*s on dataset $\mathcal{D}_t$ and $p_i^{t_c}$ represents the i-th element of penalty table P_{t_c}. In essence, this penalty factor amplifies the distances of $\boldsymbol{z}$ to all the prototype *key*s that have high frequency matching on those datasets prior to $\mathcal{D}_{t_c}$.

Prototype-guided Aggregation. After *key* matching, the *prompt*s $\mathcal{P}_{\boldsymbol{x}} = \{\boldsymbol{P}_i^{\boldsymbol{x}}\}_{i=1}^{N}$ of top-N prototype pairs will be further fed into text encoder $\boldsymbol{E}_{\text{txt}}$ to generate N prototype features $\boldsymbol{F}_p \in \mathbb{R}^{N \times D_f}$.

Next, we compute the cosine similarity matrix $\boldsymbol{S} \in \mathbb{R}^{n \times N}$ between instance features $\boldsymbol{Z}$ and prototype features $\boldsymbol{F}_p$. Since each column in $\boldsymbol{S}$ exactly corresponds to a prototype, a *Softmax* normalization along column direction is performed to obtain a weighted aggregation matrix $\boldsymbol{W} \in \mathbb{R}^{n \times N}$. Finally, we obtain the prototype-guided bag-level features through $\boldsymbol{W}$, and their average yields the bag-level feature $\boldsymbol{f}_b \in \mathbb{R}^{D_f}$ for the WSI sample $\boldsymbol{x}$. The procedure can be represented as

$$\boldsymbol{f}_b = mean(Softmax(\boldsymbol{Z} \times {\boldsymbol{F}_p}^{\top})^{\top} \times \boldsymbol{Z}). \tag{3}$$

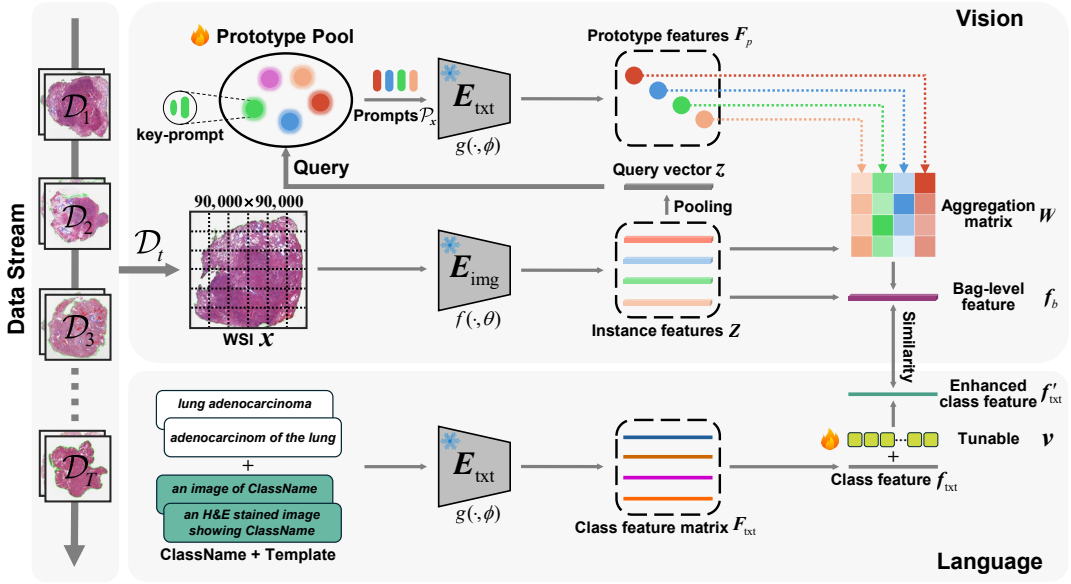

Figure 2: The framework of QPMIL-VL. The prompts in the prototype pool enable an efficient incremental learning process by gradually capturing the visual feature descriptions of instance prototypes present in the sequential WSI datasets.

Ideally, in incremental training, different datasets will match a non-overlapped set of prototype pairs. In practice, some prototype pairs could often be shared with low frequencies by different datasets. Our method doesn't impose a strict independence of matched prototype pairs. Instead, it allows WSI datasets to choose their optimal prototypes, providing more flexibility and adaptability.

Class Feature Enhancement (CFE)

As shown in Fig. 2, class text feature enhancing branch is designed to generate an enhanced class feature $\boldsymbol{f'}_{\text{txt}}$.

Class Ensemble. For a specific subtype of cancer, such as "lung adenocarcinoma", it can also be expressed as "adenocarcinoma of the lung". Therefore, we adopt class ensemble to obtain better class feature. For each cancer subtype, we combine several common class names and templates to obtain m different class text descriptions. All text descriptions are encoded by text encoder $\boldsymbol{E}_{\text{txt}}$ with $g(\cdot,\phi)$ to obtain initial class feature matrix $\boldsymbol{F}_{\text{txt}} \in \mathbb{R}^{m\times D_f}$, and by averaging $\boldsymbol{F}_{\text{txt}}$, we get the class feature $\boldsymbol{f}_{\text{txt}} \in \mathbb{R}^{D_f}$ for this subtype. The entire calculation process can be expressed as

$$\boldsymbol{f}_{\text{txt}} = mean(g(\boldsymbol{T},\phi)), \tag{4}$$

where $\boldsymbol{T} \in \mathbb{R}^{m\times L}$ represents the text tokens corresponding to the text descriptions and L is the length of text token.

Tunable Vector. To further refine class feature $\boldsymbol{f}_{\text{txt}}$, referring to the TaskRes (Yu et al. 2023), we design a tunable (*i.e.*, learnable) vector $\boldsymbol{v} \in \mathbb{R}^{D_f}$ of the same dimension for each class. The final enhanced class feature $\boldsymbol{f'}_{\text{txt}} \in \mathbb{R}^{D_f}$ is calculated through

$$\boldsymbol{f'}_{\text{txt}} = \boldsymbol{f}_{\text{txt}} + \alpha \cdot \boldsymbol{v}, \tag{5}$$

where α (default constant 0.5) is an additional amplitude factor to control the superposition scale of vector $\boldsymbol{v}$ on feature $\boldsymbol{f}_{\text{txt}}$.

Prediction and Optimization

Class Prediction. The probability of predicting the WSI sample $\boldsymbol{x}$ as the class y can be computed as

$$p(y|\boldsymbol{x}) = \frac{e^{\tau\cdot\langle \boldsymbol{f}_b, [\boldsymbol{f'}_{\text{txt}}]_y\rangle}}{\sum_{c=1}^{C} e^{\tau\cdot\langle \boldsymbol{f}_b, [\boldsymbol{f'}_{\text{txt}}]_c\rangle}}, \tag{6}$$

where C indicates the total number of classes, $[\boldsymbol{f'}_{\text{txt}}]_c$ represents the c-th class feature, τ is a temperature parameter learned by CONCH (Lu et al. 2024) and $\langle\cdot,\cdot\rangle$ denotes the cosine similarity.

Optimization Objective. To appropriately reduce the similarity between features of different classes, we firstly define a *class similarity loss*, which is calculated through

$$\mathcal{L}_S = \frac{2}{C\cdot(C-1)} \sum_{i=1}^{C}\sum_{j=i+1}^{C} \langle [\boldsymbol{f'}_{\text{txt}}]_i, [\boldsymbol{f'}_{\text{txt}}]_j\rangle + 1. \tag{7}$$

Moreover, there are another two losses. One is *classification loss*. Based on Eq. 6, it is computed by

$$\mathcal{L}_C = -\log \frac{e^{\tau\cdot\langle \boldsymbol{f}_b, [\boldsymbol{f'}_{\text{txt}}]_y\rangle}}{\sum_{c=1}^{C} e^{\tau\cdot\langle \boldsymbol{f}_b, [\boldsymbol{f'}_{\text{txt}}]_c\rangle}}. \tag{8}$$

The other is *matching loss* to make the matched *keys* $\mathcal{K}_{\boldsymbol{x}} = \{\boldsymbol{k}_i^{\boldsymbol{x}}\}_{i=1}^{N}$ closer to the query vector $\boldsymbol{z}$ of WSI sample $\boldsymbol{x}$. It is defined as

$$\mathcal{L}_M = \frac{1}{N}\sum_{i=1}^{N} q(\boldsymbol{z}, \boldsymbol{k}_i^{\boldsymbol{x}}). \tag{9}$$

To optimize our QPMIL-VL model in training, based on the three losses above, *total loss* $\mathcal{L}_T$ is computed by

$$\mathcal{L}_T = \mathcal{L}_C + \lambda\cdot\mathcal{L}_M + \beta\cdot\mathcal{L}_S, \tag{10}$$

where balance factors λ and β are set to constant 0.5. Therefore, our optimization objective is to minimize $\mathcal{L}_T$.

IL Type	Method	ACC (↑)	Upper-bound Ratio (↑)	Forgetting (↓)	BWT (↑)	Params	Masked ACC (↑)
Baseline	JointTrain (Upper)	0.908 ± 0.022	1.000 ± 0.000	-	-	0.502	0.937 ± 0.022
	FineTune (Lower)	0.308 ± 0.045 ***	0.336 ± 0.045	0.841 ± 0.054	-0.841 ± 0.054		0.844 ± 0.038 ***
Regularization-based	EWC (Kirkpatrick et al. 2017)	0.309 ± 0.051 ***	0.338 ± 0.055	0.836 ± 0.064	-0.836 ± 0.064	0.502	0.840 ± 0.035 ***
	LwF (Li and Hoiem 2017)	0.378 ± 0.081 ***	0.412 ± 0.083	0.731 ± 0.116	-0.731 ± 0.116		0.916 ± 0.031
Rehearsal-based	A-GEM/30 (Chaudhry et al. 2018)	0.436 ± 0.058 ***	0.477 ± 0.070	0.670 ± 0.072	-0.670 ± 0.072	0.502	0.879 ± 0.045 **
	ER-ACE/30 (Caccia et al. 2021)	0.666 ± 0.049 ***	0.732 ± 0.051	0.075 ± 0.057	-0.070 ± 0.059		0.917 ± 0.023
	DER++/30 (Buzzega et al. 2020)	0.749 ± 0.055 ***	0.823 ± 0.059	0.219 ± 0.059	-0.214 ± 0.062		0.904 ± 0.029 **
	ER/30 (Chaudhry et al. 2019)	0.790 ± 0.040 ***	0.870 ± 0.049	0.186 ± 0.044	-0.186 ± 0.044		0.894 ± 0.036 **
	ConSlide/30 (Huang et al. 2023a)	0.659 ± 0.022	-	0.076 ± 0.030	-0.075 ± 0.030	39.446	0.861 ± 0.017
	ER/100 (Chaudhry et al. 2019)	0.827 ± 0.028 ***	0.910 ± 0.034	0.143 ± 0.034	-0.142 ± 0.032	0.502	0.918 ± 0.028
Pre-trained VLM-based	AttriCLIP (Wang et al. 2023)	0.616 ± 0.056 ***	0.677 ± 0.071	0.285 ± 0.059	-0.285 ± 0.059	**0.111**	0.844 ± 0.026 ***
	MI-Zero (Lu et al. 2023)	0.839 ± 0.034 ***	0.927 ± 0.044	-	-	-	0.909 ± 0.019 **
	QPMIL-VL (ours)	$\mathbf{0.890 \pm 0.021}$	$\mathbf{0.982 \pm 0.032}$	$\mathbf{0.027 \pm 0.014}$	$\mathbf{-0.027 \pm 0.014}$	0.365	$\mathbf{0.930 \pm 0.018}$

Table 1: Main results in class-incremental learning scenario on forward-order training. The best performances are highlighted as **bold**. "Params" represents the number of learnable parameters (M). Gray numbers are directly cited from published paper. "/30" and "/100" mean buffer 30 WSIs and buffer 100 WSIs, respectively. We present the Masked ACC reflecting task-incremental learning scenario for reference. */**/*** denote there are significant differences (paired t-test p-value $< 0.05/0.01/0.001$) between the best performance and others.

Experiments

Following the setup of previous representative work (Van de Ven and Tolias 2019), we mainly evaluate QPMIL-VL in the class-incremental learning scenario and additionally report the results in the task-incremental learning scenario.

Experimental Details

Datasets. Following ConSlide (Huang et al. 2023a), we use four public WSI datasets from The Cancer Genome Atla (TCGA) repository: non-small cell lung carcinoma (NSCLC), invasive breast carcinoma (BRCA), renal cell carcinoma (RCC) and esophageal carcinoma (ESCA). With the arrival of each dataset, we add two additional classes to the model for training and evaluation.

We use CLAM (Lu et al. 2021) to crop non-overlapping 256×256 patches from the segmented tissue at $10\times$ magnification. Then, pre-trained image encoder $\boldsymbol{E}_{\text{img}}$ in CONCH (Lu et al. 2024) is used to extract instance features.

Evaluation Metrics. We assess all incremental learning methods with Average Accuracy (ACC), Upper-bound Ratio, Forgetting and Backward Transfer (BWT) (Lopez-Paz and Ranzato 2017; Hayes et al. 2018; Fini et al. 2022). Besides, we provide the Masked Average Accuracy (Masked ACC) metric, which is used for reference only to evaluate performance in task-incremental learning scenario by masking the logits of irrelevant classes. All results are reported by the mean and standard deviation of ten-fold cross-validation.

Training Details. In all experiments, we set the number of training epoch to 12 for each dataset and use the Adam (Kingma and Ba 2014) optimizer with $\text{lr} = 0.001$ and $\text{weight\_decay} = 0.0005$. The mini-batch sizes on forward-order and reverse-order training are set to 16 and 8, respectively. For QPMIL-VL, we set $M = 20, N = 5$ and $L_{\boldsymbol{P}} = 24$. All experiments are conducted on a machine with two NVIDIA GeForce RTX 3090 GPUs.

Refer to our Supplementary Material for more details.

Comparison Results

Compared Methods. It is necessary to compare with the upper and lower bounds of IL methods. We employ the classic ABMIL (Ilse, Tomczak, and Welling 2018) as the aggregator for the vision branch, serving as a **Baseline**. Its performances by jointly training on all datasets (JointTrain, *non-incremental*) are taken as upper bound, and those by naively fine-tuning sequential datasets (FineTune) are taken as lower bound. In addition, we compare **Regularization-based** methods: EWC (Kirkpatrick et al. 2017) and LwF (Li and Hoiem 2017); **Rehearsal-based** methods: A-GEM (Chaudhry et al. 2018), ER (Chaudhry et al. 2019), DER++ (Buzzega et al. 2020), ER-ACE (Caccia et al. 2021) and ConSlide (Huang et al. 2023a); and **Pre-trained VLM-based** methods: AttriCLIP (Wang et al. 2023) and MI-Zero (Lu et al. 2023). To ensure a fair comparison, all methods are adapted to VLM. Specially, they use the same instance features and text encoder $\boldsymbol{E}_{\text{txt}}$ in CONCH.

Results on Forward-order Training. This set of comparisons on NSCLC→BRCA→RCC→ESCA is shown in Tab. 1. From this table, we have three empirical findings.

(1) Our QPMIL-VL method is consistently superior to all other methods, with parameters only slightly larger than the smallest one. QPMIL-VL achieves an ACC of 0.890 ± 0.021 with an Upper-bound Ratio as high as 0.982 ± 0.032, winning SOTA performance metrics. Its Upper-bound Ratio is even 5.5% higher than that of MI-Zero, the second-best method. Additionally, our model has a small learnable parameter size of 0.365M, which is only slightly larger than the smallest size of 0.111M in AttriCLIP. However, in terms of performance metrics, our method is far superior to the lowest-parameter method, *i.e.*, AttriCLIP. Besides, in task-incremental experiments, our Masked ACC could reach up to 0.930 ± 0.018, which is also significantly higher than any achieved by other compared methods.

(2) Two regularization-based methods are almost ineffective in WSI class-incremental learning scenario. For example, as a regularization-based method, LwF only achieves an ACC of 0.378 ± 0.081, which is slightly higher

IL Type	Method	ACC (↑)	Upper-bound Ratio (↑)	Forgetting (↓)	BWT (↑)	Masked ACC (↑)
Baseline	JointTrain (Upper)	0.908 ± 0.022	1.000 ± 0.000	-	-	0.937 ± 0.022
	FineTune (Lower)	0.234 ± 0.008 ***	0.262 ± 0.010	0.927 ± 0.023	-0.927 ± 0.023	0.803 ± 0.060 ***
Regularization-based	EWC (Kirkpatrick et al. 2017)	0.235 ± 0.011 ***	0.264 ± 0.011	0.928 ± 0.020	-0.928 ± 0.020	0.833 ± 0.069 ***
	LwF (Li and Hoiem 2017)	0.236 ± 0.016 ***	0.265 ± 0.012	0.908 ± 0.030	-0.908 ± 0.030	0.900 ± 0.041
Rehearsal-based	A-GEM/30 (Chaudhry et al. 2018)	0.536 ± 0.047 ***	0.591 ± 0.053	0.527 ± 0.067	-0.527 ± 0.067	0.872 ± 0.026 ***
	ER-ACE/30 (Caccia et al. 2021)	0.703 ± 0.049 ***	0.777 ± 0.053	0.281 ± 0.062	-0.279 ± 0.063	0.889 ± 0.041 *
	DER++/30 (Buzzega et al. 2020)	0.684 ± 0.055 ***	0.755 ± 0.063	0.310 ± 0.069	-0.307 ± 0.072	0.910 ± 0.043
	ER/30 (Chaudhry et al. 2019)	0.644 ± 0.028 ***	0.711 ± 0.033	0.387 ± 0.050	-0.386 ± 0.049	0.901 ± 0.035
	ConSlide/30 (Huang et al. 2023a)	0.499 ± 0.025	-	0.058 ± 0.032	-0.021 ± 0.039	0.854 ± 0.039
Pre-trained VLM-based	AttriCLIP (Wang et al. 2023)	0.694 ± 0.058 ***	0.766 ± 0.061	0.207 ± 0.063	-0.207 ± 0.063	0.861 ± 0.019 ***
	MI-Zero (Lu et al. 2023)	0.839 ± 0.034	0.927 ± 0.044	-	-	0.909 ± 0.019
	QPMIL-VL (ours)	$\mathbf{0.859 \pm 0.032}$	$\mathbf{0.946 \pm 0.028}$	$\mathbf{0.064 \pm 0.031}$	$\mathbf{-0.064 \pm 0.031}$	$\mathbf{0.925 \pm 0.018}$

Table 2: Main results in class-incremental learning scenario on reverse-order training.

QPMIL	CFE			ACC (↑)	Forgetting (↓)
	TV	CE	CSL		
				0.308 ± 0.045	0.841 ± 0.054
✓				0.720 ± 0.029	0.158 ± 0.025
✓	✓			0.747 ± 0.039	0.149 ± 0.037
✓	✓	✓		0.874 ± 0.033	0.052 ± 0.030
✓	✓	✓	✓	$\mathbf{0.890 \pm 0.021}$	$\mathbf{0.027 \pm 0.014}$

Table 3: Ablation studies on main components (TV: Tunable Vector, CE: Class Ensemble, CSL: Class Similarity Loss).

Ablation	ACC (↑)	Forgetting (↓)
w/o key	0.675 ± 0.029	0.332 ± 0.034
w/o matching penalty	0.659 ± 0.029	0.356 ± 0.048
QPMIL-VL	$\mathbf{0.890 \pm 0.021}$	$\mathbf{0.027 \pm 0.014}$

Table 4: Ablation studies on prototype pool.

than the lower bound of 0.308 ± 0.045.

(3) Most rehearsal-based and pre-trained VLM-based methods could achieve good classification results. The rehearsal-based DER++ acquires an ACC of 0.749 ± 0.055 with an Upper-bound Ratio of 0.823 ± 0.059, and the pre-trained VLM-based MI-Zero obtains an ACC of 0.839 ± 0.034 with an Upper-bound Ratio of 0.927 ± 0.044. Their Upper-bound Ratio results exceed 80%. Therefore, these two types of methods are believed to be more effective than the former, consistent with natural image observations. Notably, although AttriCLIP performs well in the natural image-based IL setting, its performance in this task is lower than some rehearsal-based methods, reflecting the unique challenges of analyzing gigapixel WSI. We consider the main reasons are the poor bag-level feature (obtained through *MaxPooling*) and the overly similar class features derived from the same set of prompts.

Results on Reverse-order Training. In incremental classification, the training order of the dataset could influence the model's performance. The results on reverse-order training (ESCA→RCC→BRCA→NSCLC) are shown in Tab. 2. By comparison, we have three observations, which are almost the same as those from Tab. 1.

(1) Our QPMIL-VL method is consistently superior to all other methods in both class-incremental and task-incremental learning scenarios, maintaining a SOTA position. **(2)** Regularization-based methods are still ineffective in class-incremental learning scenario due to suffering from severe catastrophic forgetting. **(3)** Most rehearsal-based and pre-trained VLM-based methods often achieve good classification performance.

Comparison between Forward-order and Reverse-order Training. Additionally, we observe that on reverse-order training, the performance of most methods is noticeably lower than on forward-order training. For example, our QPMIL-VL decreases to an ACC of 0.859 ± 0.032. In contrast, its ACC is 0.890 ± 0.021 on forward-order training (as shown in Tab. 1). One possible explanation to this decline is that in IL, later datasets can affect earlier ones. ESCA, with only 150 slides, is the first to be learned, while NSCLC, with 965 slides, is the last. NSCLC is dominant in the total number of training samples, which could cause a greater influence on the earlier datasets, leading to more forgetting.

Ablation and Analysis

Ablation Studies. The following are several ablation analyses of QPMIL-VL.

I) on Main Components To analyze the impact of the main components on performance, we conduct a set of ablation studies on forward-order training. The results are shown in Tab. 3, where the first row represents our baseline (FineTune). Two obvious findings can be observed as follows.

(1) Each component in QPMIL-VL contributes to the performance improvement. For example, starting from the baseline, QPMIL can significantly boost the ACC from 0.308 ± 0.045 to 0.720 ± 0.029. Similarly, incrementally applying Class Similarity Loss raises the ACC from 0.874 ± 0.033 to 0.890 ± 0.021. **(2)** Among all components, both QPMIL and Class Ensemble have a more significant impact on performance than the other two. This is because the prototype query mechanism in QPMIL effectively mitigates catastrophic forgetting, while the Class Ensemble significantly optimizes class feature encoding in the VLM by increasing the diversity of class text descriptions. These comparisons imply that each component is effective.

II) on Prototype Pool Furthermore, we conduct another set of ablation studies specifically on the queryable prototype pool. Our experimental results are shown in Tab. 4.

(1) "**w/o key**" indicates the removal of the *key* from prototype pair, implying that incremental training always uses the

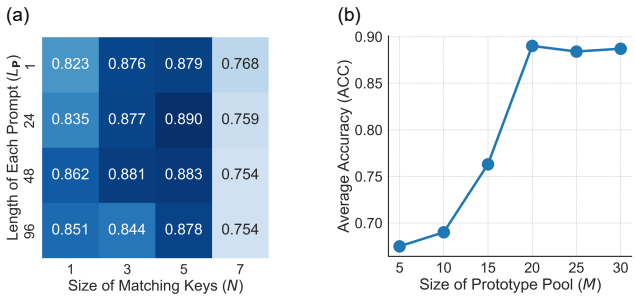

Figure 3: (a) ACC w.r.t length of each prompt (L_P) and size of matching keys (N), size of prototype pool $M = 20$; (b) ACC w.r.t M, $L_P = 24$ and $N = 5$.

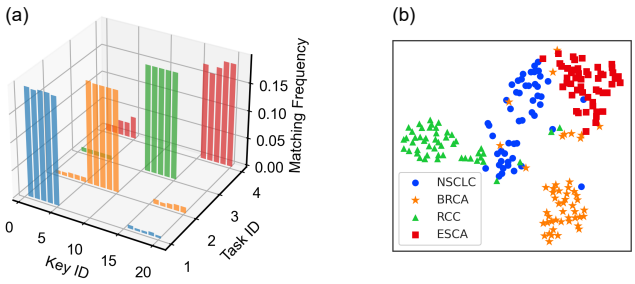

Figure 4: (a) Prototype key matching frequency histogram; (b) prototype feature visualization.

same set of *prompt*s. In this situation, the knowledge from different datasets is encoded into the same set of *prompt*s, leading to catastrophic forgetting and a significant performance decline (ACC dropping to 0.675 ± 0.029). **(2) "w/o matching penalty"** presents that only Eq. 1 is used for matching prototype pairs. Under this ablation, the ACC metric decreases to 0.659 ± 0.029 because different datasets tend to match the same set of prototypes during training, which consequently results in catastrophic forgetting. Only when both key and matching penalty are applied together in our QPMIL-VL does the ACC reach its highest value of 0.890 ± 0.021. This ablation demonstrates that both are beneficial for incremental classification.

Further Performance Analysis. Additionally, we conduct additional experiments to further analyze our method.

I) Effect of Hyper-parameters As shown in Fig. 3, there are three important parameters in the prototype pool. The results indicate that inappropriate prototype pool capacity ((a): $N = 7$, (b): $M < 20$) may result in knowledge interference and too few *prompt*s ((a): $N = 1$) make it difficult to describe the various instance prototype visual features present in WSIs.

II) Prototype Key Matching To analyze the matching results of *key*s on incremental datasets, we compute the matching frequency of each prototype pair in the first fold of experiments, as illustrated in Fig. 4 (a). It can be observed that each dataset tends to stably match 5 *key*s. Moreover, the 5 *key*s predominantly matched by different datasets are often non-overlapping due to the matching penalty.

III) Prototype Feature Visualization As depicted in Fig. 4 (b), we present the distribution visualization of 200 prototype features using t-SNE (Van der Maaten and Hinton 2008). These features are derived from the ten-fold cross-validation experiments conducted on four datasets, *i.e.*, $200 = 10 \times 4 \times 5$. Their distributions show that most prototypes learned from the same dataset exhibit clear clustering properties. From another perspective, they also imply that our model can learn distinct prototype features from different datasets. Additionally, from Fig. 4 (a) and (b), we observe a small number of outliers. They can be regarded as shared prototypes.

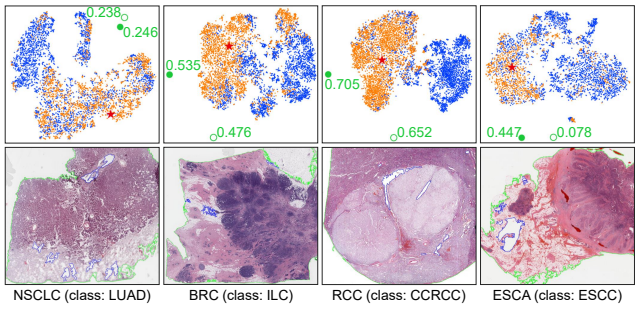

Figure 5: Class features visualization, only one class per dataset. Red ⋆ is the center, *i.e.*, the average of features, for the instances of interest (orange points, determined by the cosine similarity between instance features and learned prototype features). Blue points are non-interest instance features. Green ○ and ● denote the cosine similarities before and after feature enhancement, respectively.

IV) Class Feature Enhancement Analysis To show the effectiveness of CFE, we visualize four classes of WSI samples and their corresponding class features, as shown in Fig. 5. More experimental results can be found in our Supplementary Material.

From the alteration in feature similarity, we can observe that the class features are noticeably improved by the enhancement. For example, in the case of the class sample ESCA-ESCC, the initial cosine similarity between the center point feature and the class feature is only 0.078, but it increases to 0.447 after enhancement, resulting in an improvement of 0.369. This alteration suggests that the CFE is effective in enhancing the quality of class features.

Conclusions

This paper proposes the first Vision-Language-based framework for incremental WSI classification, QPMIL-VL, breaking through traditional vision modality. By a pool of learnable prototype pairs, it aggregates the instance features from pre-trained image encoder to generate bag-level feature under the guidance of matched prototype *prompt*s. Class probability is predicted by the cosine similarity between bag-level feature and enhanced class feature. Extensive experiments demonstrate that our method and its components are effective. It obviously outperforms other methods in incremental WSI classification, achieving SOTA results.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (NSFC) under Grant No.62476049.

References

Buzzega, P.; Boschini, M.; Porrello, A.; Abati, D.; and Calderara, S. 2020. Dark experience for general continual learning: a strong, simple baseline. *Advances in neural information processing systems*, 33: 15920–15930.

Caccia, L.; Aljundi, R.; Asadi, N.; Tuytelaars, T.; Pineau, J.; and Belilovsky, E. 2021. New insights on reducing abrupt representation change in online continual learning. *arXiv preprint arXiv:2104.05025*.

Campanella, G.; Hanna, M. G.; Geneslaw, L.; Miraflor, A.; Werneck Krauss Silva, V.; Busam, K. J.; Brogi, E.; Reuter, V. E.; Klimstra, D. S.; and Fuchs, T. J. 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8): 1301–1309.

Cha, H.; Lee, J.; and Shin, J. 2021. Co2L: Contrastive Continual Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 9516–9525.

Chaudhry, A.; Ranzato, M.; Rohrbach, M.; and Elhoseiny, M. 2018. Efficient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420*.

Chaudhry, A.; Rohrbach, M.; Elhoseiny, M.; Ajanthan, T.; Dokania, P. K.; Torr, P. H.; and Ranzato, M. 2019. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*.

Derakhshani, M. M.; Najdenkoska, I.; van Sonsbeek, T.; Zhen, X.; Mahapatra, D.; Worring, M.; and Snoek, C. G. 2022. Lifelonger: A benchmark for continual disease classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 314–324. Springer.

Fini, E.; Da Costa, V. G. T.; Alameda-Pineda, X.; Ricci, E.; Alahari, K.; and Mairal, J. 2022. Self-supervised models are continual learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9621–9630.

Gao, P.; Geng, S.; Zhang, R.; Ma, T.; Fang, R.; Zhang, Y.; Li, H.; and Qiao, Y. 2024. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2): 581–595.

Hayes, T. L.; Kemker, R.; Cahill, N. D.; and Kanan, C. 2018. New metrics and experimental paradigms for continual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2031–2034.

Huang, Y.; Zhao, W.; Wang, S.; Fu, Y.; Jiang, Y.; and Yu, L. 2023a. 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*, 21349–21360.

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

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

Ke, Z.; Liu, B.; and Huang, X. 2020. Continual learning of a mixed sequence of similar and dissimilar tasks. *Advances in neural information processing systems*, 33: 18493–18504.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13): 3521–3526.

Li, B.; Li, Y.; and Eliceiri, K. W. 2021. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14318–14328.

Li, H.; Chen, Y.; Chen, Y.; Yu, R.; Yang, W.; Wang, L.; Ding, B.; and Han, Y. 2024. Generalizable Whole Slide Image Classification with Fine-Grained Visual-Semantic Interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11398–11407.

Li, X.; Zhou, Y.; Wu, T.; Socher, R.; and Xiong, C. 2019. Learn to grow: A continual structure learning framework for overcoming catastrophic forgetting. In *International conference on machine learning*, 3925–3934. PMLR.

Li, Z.; and Hoiem, D. 2017. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12): 2935–2947.

Lin, T.; Yu, Z.; Hu, H.; Xu, Y.; and Chen, C.-W. 2023. Interventional bag multi-instance learning on whole-slide pathological images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19830–19839.

Liu, P.; Ji, L.; Gou, J.; Fu, B.; and Ye, M. 2024a. Interpretable Vision-Language Survival Analysis with Ordinal Inductive Bias for Computational Pathology. *arXiv preprint arXiv:2409.09369*.

Liu, P.; Ji, L.; Ye, F.; and Fu, B. 2024b. Advmil: Adversarial multiple instance learning for the survival analysis on whole-slide images. *Medical Image Analysis*, 91: 103020.

Liu, P.; Ji, L.; Zhang, X.; and Ye, F. 2024c. Pseudo-Bag Mixup Augmentation for Multiple Instance Learning-Based Whole Slide Image Classification. *IEEE Transactions on Medical Imaging*, 43(5): 1841–1852.

Lopez-Paz, D.; and Ranzato, M. 2017. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30.

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

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

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

Mai, Z.; Li, R.; Jeong, J.; Quispe, D.; Kim, H.; and Sanner, S. 2022. Online continual learning in image classification: An empirical survey. *Neurocomputing*, 469: 28–51.

Mallya, A.; and Lazebnik, S. 2018. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 7765–7773.

McCloskey, M.; and Cohen, N. J. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, 109–165. Elsevier.

Parisi, G. I.; Kemker, R.; Part, J. L.; Kanan, C.; and Wermter, S. 2019. Continual lifelong learning with neural networks: A review. *Neural networks*, 113: 54–71.

Prabhu, A.; Torr, P. H.; and Dokania, P. K. 2020. Gdumb: A simple approach that questions our progress in continual learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, 524–540. Springer.

Qu, L.; Fu, K.; Wang, M.; Song, Z.; et al. 2024. 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.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Rusu, A. A.; Rabinowitz, N. C.; Desjardins, G.; Soyer, H.; Kirkpatrick, J.; Kavukcuoglu, K.; Pascanu, R.; and Hadsell, R. 2016. Progressive neural networks. *arXiv preprint arXiv:1606.04671*.

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

Shokri, R.; and Shmatikov, V. 2015. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 1310–1321.

Song, A. H.; Chen, R. J.; Ding, T.; Williamson, D. F.; Jaume, G.; and Mahmood, F. 2024. Morphological prototyping for unsupervised slide representation learning in computational pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11566–11578.

Song, A. H.; Jaume, G.; Williamson, D. F.; Lu, M. Y.; Vaidya, A.; Miller, T. R.; and Mahmood, F. 2023. Artificial intelligence for digital and computational pathology. *Nature Reviews Bioengineering*, 1(12): 930–949.

Van de Ven, G. M.; and Tolias, A. S. 2019. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*.

Van der Laak, J.; Litjens, G.; and Ciompi, F. 2021. Deep learning in histopathology: the path to the clinic. *Nature medicine*, 27(5): 775–784.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, L.; Zhang, X.; Su, H.; and Zhu, J. 2024. A comprehensive survey of continual learning: theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wang, R.; Duan, X.; Kang, G.; Liu, J.; Lin, S.; Xu, S.; Lü, J.; and Zhang, B. 2023. Attriclip: A non-incremental learner for incremental knowledge learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3654–3663.

Wang, Y.; Huang, Z.; and Hong, X. 2022. S-prompts learning with pre-trained transformers: An occam's razor for domain incremental learning. *Advances in Neural Information Processing Systems*, 35: 5682–5695.

Wang, Z.; Zhang, Z.; Lee, C.-Y.; Zhang, H.; Sun, R.; Ren, X.; Su, G.; Perot, V.; Dy, J.; and Pfister, T. 2022. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 139–149.

Xu, H.; Usuyama, N.; Bagga, J.; Zhang, S.; Rao, R.; Naumann, T.; Wong, C.; Gero, Z.; González, J.; Gu, Y.; et al. 2024. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 1–8.

Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; and Wu, Y. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Yu, T.; Lu, Z.; Jin, X.; Chen, Z.; and Wang, X. 2023. Task residual for tuning vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10899–10909.

Zhang, H.; Meng, Y.; Zhao, Y.; Qiao, Y.; Yang, X.; Coupland, S. E.; and Zheng, Y. 2022. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18802–18812.

Zheng, Y.; Li, J.; Shi, J.; Xie, F.; Huai, J.; Cao, M.; and Jiang, Z. 2023. Kernel attention transformer for histopathology whole slide image analysis and assistant cancer diagnosis. *IEEE Transactions on Medical Imaging*, 42(9): 2726–2739.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.