PTCMIL: Multiple Instance Learning via Prompt Token Clustering for Whole Slide Image Analysis

Beidi Zhao[1,2*], SangMook Kim[3*†], Hao Chen[4], Chen Zhou[1,5], Zu-hua Gao[1,5], Gang Wang[1,5 ‡], and Xiaoxiao Li[1,2 ‡]

[1] The University of British Columbia
[2] Vector Institute
[3] Chungnam National University
[4] The Hong Kong University of Science and Technology
[5] BC Cancer Agency
beidiz@student.ubc.ca, sangmook.kim@cnu.ac.kr, jhc@cse.ust.hk,
czhou@bccancer.bc.ca, zuhua.gao@ubc.ca, gang.wang1@bccancer.bc.ca,
xiaoxiao.li@ece.ubc.ca

Abstract. Multiple Instance Learning (MIL) has advanced WSI analysis but struggles with the complexity and heterogeneity of WSIs. Existing MIL methods face challenges in aggregating diverse patch information into robust WSI representations. While ViTs and clustering-based approaches show promise, they are computationally intensive and fail to capture task-specific and slide-specific variability. To address these limitations, we propose PTCMIL, a novel Prompt Token Clustering-based ViT for MIL aggregation. By introducing learnable prompt tokens into the ViT backbone, PTCMIL unifies clustering and prediction tasks in an end-to-end manner. It dynamically aligns clustering with downstream tasks, using projection-based clustering tailored to each WSI, reducing complexity while preserving patch heterogeneity. Through token merging and prototype-based pooling, PTCMIL efficiently captures task-relevant patterns. Extensive experiments on eight datasets demonstrate its superior performance in classification and survival analysis tasks, outperforming state-of-the-art methods. Systematic ablation studies confirm its robustness and strong interpretability. The code is released at https://github.com/ubc-tea/PTCMIL.

Keywords: Multiple Instance Learning · Prompt Learning · Clustering.

1 Introduction

Histopathology is the gold standard for cancer diagnosis, essential for tumor detection, subtyping, and survival prediction. With advancements in deep learning, digital pathology, which analyzes whole slide images (WSIs), has gained

[*] Equal contribution.

[†] Work done while the author was a postdoctoral fellow at the University of British Columbia.

[‡] Co-corresponding authors.

prominence [1,2,3]. WSIs are massive giga-pixel images that require computationally efficient processing, typically through multiple instance learning (MIL), which enables slide-level annotation without patch-level labels. However, WSIs exhibit significant inherent heterogeneity, containing diverse cell types and tissue structures with varying morphological and staining characteristics [4,9]. A key challenge in MIL is aggregating redundant patch information into robust WSI representations. Early MIL approaches treated patches independently, neglecting interactions, but recent methods leverage patch relationships for improved modeling [17,4]. The introduction of Vision Transformers (ViTs) [17,23] has enhanced global dependency modeling through self-attention. Despite their effectiveness, ViTs face computational bottlenecks and overfitting issues, limiting their scalability in MIL applications [25].

To address above issues and handle the vast diversity of patches within each WSI, recent methods incorporate clustering techniques to identify representative prototypes, thereby enhancing WSI representations [24,18]. These approaches typically follow a two-stage process [18]: (1) unsupervised clustering groups patches into prototypes, often leveraging global clustering across all patches, and (2) a pooling model trains on these prototypes for prediction over each WSI. While effective in capturing shared patterns, these methods face key limitations: (i) standalone clustering is not optimized for downstream tasks, potentially missing task-specific features, (ii) global clustering is computationally expensive, requiring patch sampling that may omit critical regions, and (iii) uniform centroids across WSIs fail to account for slide-specific variability, reducing adaptability. These challenges lead to our research question: *How can we optimize patch clustering alongside WSI-level analysis efficiently and effectively?*

Visual Prompting (VP) [12], adapted from natural language processing, enables ViTs to focus on specific tasks without extensive retraining. Prior work [21] highlights that learnable prompt tokens enhance flexibility and efficiency across visual tasks. To address our research question, we propose **PTCMIL**, a Prompt Token Clustering-based ViT for MIL aggregation, integrating clustering, prototyping, and downstream tasks in an end-to-end manner. Unlike traditional two-stage clustering methods, PTCMIL introduces prompt tokens to dynamically guide task-relevant clustering, capturing WSI patch heterogeneity while optimizing WSI-level analysis. We introduce projection-based clustering tailored to each WSI, reducing complexity compared to global clustering [18] while adding minimal parameters. PTCMIL improves prototype representations and WSI prediction. Our contributions are: (1) an end-to-end framework dynamically aligning clustering for prototyping with WSI-level analysis to enhance feature relevance for downstream tasks and improving performance and interpretability, (2) efficient token clustering using prompt tokens and projection-based merging tailored to each WSI, preserving heterogeneity while reducing complexity, and (3) extensive experiments demonstrating effectiveness across multiple WSI tasks, with systematic ablations validating design robustness.

2 Methodology

This section introduces the overall pipeline of our method, which can simultaneously learn downstream task-related prototypes of WSI during the training (Fig. 1). The architecture of our PTCMIL consists of three parts: 1) Learnable prompt token-based clustering; 2) Prototype merging over clusters; 3) Pooling over prototypes to get the WSI representation for different downstream tasks.

2.1 Learnable Prompt Token-based Clustering

PTCMIL builds on the ViT-based MIL aggregation. The introduced learnable prompts (▲ in Fig. 1) are appended alongside patch tokens (patch feature and represented as ■ in Fig. 1) and class tokens (represented as ● in Fig. 1) as the inputs of the ViT base model[§]. Denote N as the number of patches of the given WSI. Denote C as the desired number of clusters, a hyperparameters in our pipeline, and we have $C \ll N$. We represent the token embeddings of prompts, patches, and class tokens after the linear embedding layer (blue block in Fig. 1) as $\mathbf{P}_0 = [\mathbf{p}_0^1, \cdots, \mathbf{p}_0^C] \in \mathbb{R}^{C \times D}$, $\mathbf{E}_0 = [\mathbf{e}_0^1, \cdots, \mathbf{e}_0^N] \in \mathbb{R}^{N \times D}$, and $\mathbf{cls} \in \mathbb{R}^D$.

These tokens are fed into a global Transformer layer (yellow block in Fig. 1) to enhance contextual understanding of the global (WSI-level) information and improved feature representation :

$$[\mathbf{cls}_1, \mathbf{P}_1, \mathbf{E}_1] = f_{\text{global}}([\mathbf{cls}, \mathbf{P}_0, \mathbf{E}_0]), \tag{1}$$

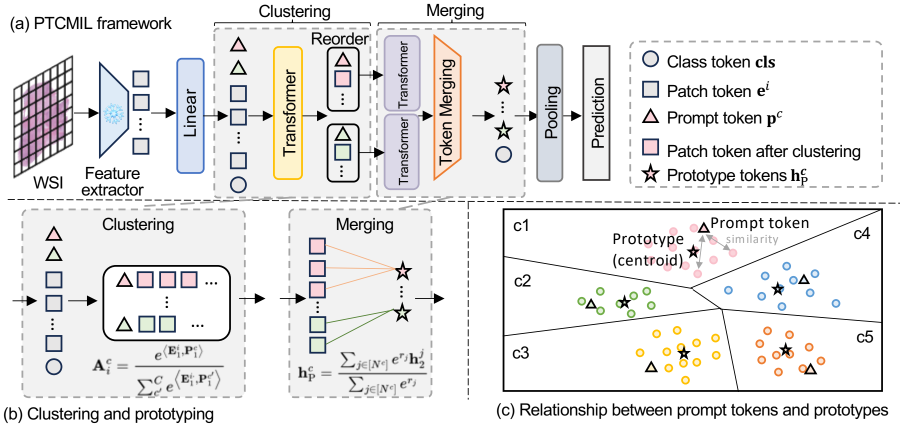
where $[\mathbf{cls}_1, \mathbf{P}_1, \mathbf{E}_1]$ are the output of the global Transformer layer f_{global}.

Prompt-based Clustering. With the refined prompt token embeddings $\mathbf{P}_1$ and $\mathbf{E}_1$, we dynamically group patches based on their distance to the C prompt embeddings via projection. The previous token clustering methods in ViT for image tasks [15,26] cluster tokens based on their pairwise similarities. However, these approaches become impractical in WSI settings, where the number of image tokens can exceed ten thousands in our cases, making the computation of pairwise distances prohibitively expensive due to high computational costs and resource constraints. To address this challenge, we introduce a novel approach leveraging learnable prompt tokens, each associated with a cluster as a proxy, for efficient clustering by projecting patch tokens onto the prompt tokens. We define an assignment matrix $\mathbf{A} \in \mathbb{R}^{N \times C}$, where each row corresponds to a patch token and each column to a cluster. The entry $\mathbf{A}_i^c$ indicates the probability that the i-th patch token belongs to cluster c and is calculated follows:

$$\mathbf{A}_i^c = \frac{e^{\left\langle \mathbf{E}_1^{i\cdot}, \mathbf{P}_1^{c\cdot} \right\rangle}}{\sum_{c'}^{C} e^{\left\langle \mathbf{E}_1^{i\cdot}, \mathbf{P}_1^{c'\cdot} \right\rangle}}, \tag{2}$$

where $\langle \cdot, \cdot \rangle$ is the inner product, $\mathbf{E}_1$ is the output of token feature after the global Transformer layer's self-attention module.

[§] We keep the class token following general VPT design [12].


(a) PTCMIL framework
Clustering
Merging
Reorder
WSI
Feature extractor
Linear
Transformer
Transformer
Transformer
Token Merging
Pooling
Prediction
Class token **cls**
Patch token $\mathbf{e}^i$
Prompt token $\mathbf{p}^c$
Patch token after clustering
Prototype tokens $\mathbf{h}_{\mathrm{P}}^c$
Clustering
$\mathbf{A}_i^c = \frac{e^{\langle \mathbf{E}_1^i, \mathbf{P}_1^c \rangle}}{\sum_{c'}^{C} e^{\langle \mathbf{E}_1^i, \mathbf{P}_1^{c'} \rangle}}$
Merging
$\mathbf{h}_{\mathrm{P}}^c = \frac{\sum_{j \in [N^c]} e^{r_j} \mathbf{h}_2^j}{\sum_{j \in [N^c]} e^{r_j}}$
c1
c2
c3
c4
c5
Prototype (centroid)
Prompt token
similarity

(b) Clustering and prototyping

(c) Relationship between prompt tokens and prototypes

Fig. 1: Overview of PTCMIL. (a) Overall framework, with patch feature tokens, prompt tokens and class token as input, and the objective prediction goal as output; (b) Interpretation of clustering and prototyping based on the clusters; (c) Interpretation of the relationship between prototypes and prompt tokens.

Prompt Updating. Xavier uniform initialization [8] is used to randomly initialize C uniform prompt tokens $\mathbf{P}_0 \in \mathbb{R}^{C \times D}$ to prevent clustering collapse. Then, the Gram-Schmidt process ensures orthogonality for $i \in [C]$:$\mathbf{u}_i = \mathbf{X}_{i \cdot} - \sum_{j=1}^{i-1} \frac{\langle \mathbf{u}_j, \mathbf{X}_{i \cdot} \rangle}{\langle \mathbf{u}_j, \mathbf{u}_j \rangle} \mathbf{u}_j, \mathbf{P}_{0i \cdot} = \frac{\mathbf{u}_i}{\|\mathbf{u}_i\|_2}$, where $\mathbf{X} = [\mathbf{X}_1, \cdots, \mathbf{X}_C]$ is a set of randomly initialized vectors with Xavier uniform initialization and $\mathbf{u} = [\mathbf{u}_1, \cdots, \mathbf{u}_C]$ is a set of orthonormal vectors. Furthermore, we introduce a soft constraint as the regularization loss function to prevent prompt collapse across clusters during training, in addition to updates based on the downstream task loss (see Sec. 2.3). Explicitly, we minimize the difference between $\mathbf{P}_1^T \mathbf{P}_1$ and the identity matrix $\mathbf{I}$:

$$\mathcal{L}_{\mathrm{reg}} = \|\mathbf{P}_1^T \mathbf{P}_1 - \mathbf{I}\|_2. \tag{3}$$

In the MIL problem, the batch size is typically set to 1 due to memory limits to process vast amount patches in each WSI. To facilitate stable prompt updating, we utilize moving average strategy to update prompts: $\bar{\mathbf{P}}_{1m} = \theta \bar{\mathbf{P}}_{1(m-1)} + (1 - \theta) \mathbf{P}_{1m}$, where m is the number of steps in one epoch, $\theta \in [0, 1]$ is the decay factor that controls how fast the prompts are updated, and $\bar{\mathbf{P}}$ indicate the averaged prompt over iterations. This approach smooths out prompt updates across batches, reducing sensitivity to individual slide differences.

2.2 Merging to Obtain Prototypes over Clusters

Next, we aim to learn the prototypes in each cluster. According to the assignment matrix obtained in Eq. (2), we have cluster index vector $\mathbf{a} = \arg\max(\mathbf{A}_{i \cdot}) = [a_1, a_2, \ldots, a_N]^T$, for $i \in [C]$. With $\mathbf{a}$, we conduct cluster-wise re-index to patch tokens and get $\mathbf{H} = [\mathbf{H}_1, \mathbf{H}_2, \ldots, \mathbf{H}_C]$, where $\mathbf{H}_i$ is the concatenation of features in cluster i. The tokens in each cluster are another local Transformer layer (denoted as f_{local} with output dimension d, shown as purple blocks in Fig. 1, parameters shared for efficiency) to learn cluster-wise local context:

$$[\bar{\mathbf{p}}_2^c, \mathbf{H}_2^c] = f_{\text{local}}([\bar{\mathbf{p}}_1^c, \mathbf{H}_1^c]), \quad c = 1, 2, \ldots, C. \tag{4}$$

where $\bar{\mathbf{p}}_2^c \in \mathbb{R}^d, \mathbf{H}_2^c \in \mathbb{R}^{N_i \times d}$ are the output from the cth Transformer for cluster i, which contains N^c patches.

To reduce redundancy before passing it to the pooling module, clustering-based MIL methods summarize representative information from each cluster as *prototypes*. In clustering task, centroid of the data in the cluster is commonly used as prototype that provide good approximation of the entire cluster [11]. Although we introduce learned prompts as proxies for clusters to enable efficient, learnable clustering, using $\bar{\mathbf{p}}_2$ as a candidate to represent clusters may deviate significantly from the actual cluster centers in practice (as illustrated in Fig. 1(b)). To address this, we propose to calculate the centroid among token embeddings $\mathbf{H}_2^c$ via merging to represent prototypes (represented as the orange block in Fig. 1). Additionally, following [16], we introduce learnable weights $\mathbf{r}_i = [r_1, \ldots, r_{N^c}]^T \in \mathbb{R}^{N^c}$, for $c \in [C]$, to explicitly represent the averaging weights of the patch token features. Hence, the prototype $\mathbf{h}_{\text{P}}^c$ for cluster i is written as via weighted averaging: $\mathbf{h}_{\text{P}}^c = \frac{\sum_{j \in [N^c]} e^{r_j} \mathbf{h}_2^j}{\sum_{j \in [N^c]} e^{r_j}} \in \mathbb{R}^d$, and we have $\mathbf{H}_{\text{P}} = [\mathbf{h}_{\text{P}}^1, \cdots, \mathbf{h}_{\text{P}}^C]$.

2.3 Global Pooling for Downstream Tasks

Focusing on the most common WSI analysis tasks: classification and survival analysis, we detail the pooling module for these downstream applications.

Classification. Due to the heterogeneity of contents in WSI, it is not suitable to only use a single class token to summarize the information over the whole WSI [5]. With the help of prototypes, we can get better slide-level representation, $\mathbf{H}_{\text{final}} = [\mathbf{cls}_1, \mathbf{H}_{\text{P}}]$, and the final prediction of WSI is $\hat{Y} = \text{Pooling}(\mathbf{H}_{\text{final}})$. Specifically, the pooling here includes mean operation over $\mathbf{H}_{\text{final}}$ and a linear layer. To this end, the overall objective function for classification is:

$$\mathcal{L} = \mathcal{L}_{\text{cla}} + \alpha \mathcal{L}_{\text{reg}} = \text{CE}(\hat{Y}, Y) + \alpha \| \bar{\mathbf{P}}^T \bar{\mathbf{P}} - \mathbf{I} \|_2, \tag{5}$$

where $\hat{Y}$ is the prediction of WSI, α is the regularization loss term weight.

Survival Analysis. The survival analysis model is used to predict the survival hazard score, which can be formulated by $f_{\text{hazard}}\,(T = t) = \lim\limits_{\partial t \to 0} \frac{P(t \leq T \leq t + \partial t | T \geq t)}{\partial t}$ $= \lambda_0(t) e^{\beta \mathbf{H}_{\text{final}}}$, which measure the probability of patient death instantaneously at t, and β is the parameters of the last linear prediction layer. In training with WSIs, we follow [7] to construct weak supervision of transforming the continuous observation time to discrete time intervals: $T_j = r, \text{ if } T_{j, \text{ cont }} \in [t_r, t_{r+1}) \text{ for } r \in \{0, 1, 2, 3\}$, where j is the patient index, $T_{j, \text{ cont }}$ is the continuous event time. For a given patient with bag-level feature $\mathbf{H}_{\text{final}}$, the hazard function can be defined as: $f_{\text{hazard}}\ \left(r \mid \mathbf{H}_{\text{final}_j}\right) = P\left(T_j = r \mid T_j \geq r, \mathbf{H}_{\text{final}_j}\right)$ and the survival function is $f_{\text{surv}}\left(r \mid \mathbf{H}_{\text{final}_j}\right) = P\left(T_j > r \mid \mathbf{H}_{\text{final}_j}\right) = \prod_{u=1}^{r}\left(1 - f_{\text{hazard}}\ \left(u \mid \mathbf{h}_{\text{final }j}\right)\right)$. The loss is log likelihood function of survival is:

$$\begin{aligned} L_{\text{surv}} = & -c_j \log\left(f_{\text{surv}}(Y_j \mid \mathbf{H}_{\text{final},j})\right) - (1 - c_j)\log\left(f_{\text{surv}}(Y_j - 1 \mid \mathbf{H}_{\text{final},j})\right) \\ & - (1 - c_j)\log\left(f_{\text{hazard}}(Y_j \mid \mathbf{H}_{\text{final},j})\right), \end{aligned} \tag{6}$$

where c_j is the binary censorship status, $c_j = 1$ means the patient live longer than the follow-up period, $c_j = 0$ means the patient passed away within time T_j. Similar to Eq (5), we add $\mathcal{L}_{\text{reg}}$ to $\mathcal{L}_{\text{surv}}$ as the total loss for survival prediction.

3 Experiment

3.1 Dataset

Classification. We evaluate PTCMIL on Camelyon16 (2-class) [1], TCGA-Non-Small Cell Lung Cancer (NSCLC) (2-class) [3], Prostate cANcer graDe Assessment (PANDA) (6-class) [2] and an in-house prostate WSI dataset (1-class). Camelyon16 is for detecting metastases (abnormal) (129) or normal (270) in breast cancer, TCGA-NSCLC is for subtyping the subtypes LUAD (538) and LUSC (512) of lung cancer. and PANDA is for grading Prostate cancer diagnosis (10,616). For Camelyon16 and TCGA-NSCLC, we follow [14] to split the training, validation and testing sets and use five-fold validation to report the result. For PANDA, we use the data splits of the challenge. To evaluate the adaptability of our method, we use the in-house prostate WSI dataset (749 cancerous slides) for testing with the model trained on PANDA.
Survival Analysis. We evaluate the survival prediction performance on Breast Invasive Carcinoma (BRCA) (1,041), Colon and Rectum Adenocarcinoma (CRC) (575), Bladder Urothelial Carcinoma (BLCA) (437) and Lung adenocarcinoma (LUAD) (519) from [3]. We follow [19] to use 5-fold site-stratified cross-validation.

3.2 Implementation and Evaluation

We use CTransPath [22] and UNI [6] to extract patch feature with CLAM's toolbox [14] to crop non-overlapping 256 × 256 (20×) patches. We use Adam for our model and keep the original optimizers for baselines. Cosine scheduler is used with a starting learning rate of 2e-4. The weight decay is set to 1e-5. The regularization loss weight α is 0.1 for Camelyon16 and PANDA, 0.2 for all TCGA datasets, decay factor θ is 0.9 for all. The numbers of clusters for Camelyon16, TCGA and PANDA are 7, 5 and 5 respectively. For the cancer normal/abnormal and subtype classification tasks, we report accuracy and AUC, presenting mean and standard deviation. For PANDA, we report Cohen's kappa. For the in-house dataset that contains only one class, we use accuracy. We report the concordance index (c-index) for survival prediction.

Table 1: Classification result on four datasets. (†: We use the reported result from the original paper.)

Feature extraction	Method	Camelyon16		TCGA-NSCLC		PANDA	PANDA → in-house prostate dataset
		AUC	Acc	AUC	Acc	Cohen's κ	Acc
CTransPath	ABMIL (NeurIPS'18) [10]	$92.40_{4.17}$	$90.31_{1.80}$	$95.61_{1.88}$	$89.81_{2.60}$	0.892	85.81
	DSMIL (CVPR '21) [13]	$93.26_{2.83}$	$87.03_{1.18}$	$96.79_{0.94}$	$90.87_{2.02}$	0.900	87.28
	CLAM (Nat. Biomed. Eng. '21) [14]	$95.89_{2.48}$	$92.19_{1.91}$	$97.13_{0.83}$	$91.60_{1.36}$	0.915	86.75
	DTFD-MIL (CVPR'22) [27]	$94.93_{1.32}$	$92.81_{3.09}$	$97.24_{0.43}$	$91.02_{1.72}$	0.913	87.00
	TransMIL (NeurIPS'22) [17]	$96.47_{1.12}$	$93.13_{2.56}$	$96.67_{0.87}$	$90.72_{0.74}$	0.897	84.34
	ILRA (ICLR'23) [23]	$94.29_{2.82}$	$90.78_{2.02}$	$96.33_{0.67}$	$90.19_{1.07}$	**0.928**	84.07
	PANTHER (CVPR'24) [19]	$67.01_{4.79}$	$64.19_{6.01}$	$93.39_{0.88}$	$91.64_{2.30}$	0.720	81.52
	MambaMIL(MICCAI'24) [25]	$92.31_{1.37}$	$91.09_{1.31}$	$96.85_{1.10}$	$91.85_{0.69}$	0.902	87.15
	DGR-MIL (ECCV'24) [28]	$91.25_{7.18}$	$90.06_{3.60}$	$96.13_{1.17}$	$89.51_{1.86}$	0.894	87.82
	PTCMIL (ours)	$\mathbf{98.06_{0.90}}$	$\mathbf{94.73_{1.33}}$	$\mathbf{97.31_{0.67}}$	$\mathbf{92.17_{1.80}}$	**0.928**	**89.96**
UNI	ABMIL (NeurIPS'18) [10]	$98.42_{0.67}$	$95.73_{2.92}$	$97.72_{0.55}$	$92.30_{1.55}$	0.935	84.74
	DSMIL (CVPR '21) [13]	$98.75_{1.08}$	$97.50_{1.02}$	$97.56_{0.69}$	$93.43_{1.30}$	0.857	84.61
	CLAM (Nat. Biomed. Eng. '21) [14]	$98.92_{0.82}$	$97.97_{0.43}$	$98.11_{0.46}$	$93.21_{0.75}$	0.933	86.75
	DTFD-MIL (CVPR'22) [27]	$98.38_{0.74}$	$96.56_{2.38}$	$97.89_{0.55}$	$92.23_{1.66}$	0.911	84.74
	TransMIL (NeurIPS'22) [17]	$99.08_{0.74}$	$95.31_{3.95}$	$98.20_{0.30}$	$93.58_{0.80}$	0.936	89.56
	ILRA (ICLR'23) [23]	$94.38_{5.48}$	$93.44_{4.51}$	$96.72_{0.71}$	$90.04_{1.64}$	0.924	89.29
	PANTHER (CVPR'24) [19]	$84.21_{6.02}$	$79.19_{6.02}$	$97.82_{0.67}$	$91.92_{1.66}$	$0.923^{\dagger}$	87.42
	MambaMIL (MICCAI'24) [25]	$99.06_{0.77}$	$97.97_{1.96}$	$97.96_{0.97}$	$92.68_{1.27}$	0.929	86.61
	DGR-MIL (ECCV'24) [28]	$98.54_{2.03}$	$97.35_{0.89}$	$97.54_{0.52}$	$92.30_{1.57}$	0.915	89.69
	PTCMIL (ours)	$\mathbf{99.60_{0.34}}$	$\mathbf{98.60_{0.35}}$	$\mathbf{98.44_{0.39}}$	$\mathbf{93.81_{1.02}}$	**0.937**	**92.64**

3.3 Comparison with Baselines

Classification and Survival Analysis. Tables 1 shows the result ofclassification on Camelyon16 (abnormal detection), TCGA-NSCLC (subtyping), PANDA (grading), and in-house prostate (adaptation) datasets. We also conduct survival prediction analysis on four TCGA datasets in Table 2. PTCMIL consistently demonstrates high performance, highlighting the end-to-end integration of clustering enables optimal WSI representation learning for various downstream task.

Adaptability. We explore few-shot (20 random WSIs with balanced labels) domain adaptation using limited WSIs to transfer learned prompt tokens to new tasks. We only update the classifier and prototypes (if have). Table 3 shows that a small number of prompt tokens enable effective cross-domain adaptation. This is promising for resource-constrained scenarios with limited training data, highlighting PTCMIL's adaptability and robustness across varied domains.

3.4 Visualization and Interpretation

Fig. 2 shows (a) WSI clustering maps, cluster assignment bar plots and example patches in each cluster (b) comparison to PANTHER [18]. In details, most patches in c0 are related to tumor cells that are irregular in shape, with enlarged, darkly stained nuclei and often disordered arrangement, showing high mitotic activity. c1 are mainly tumor cells, which are characterized by irregular shapes, enlarged nuclei, disordered arrangements. Lung alveoli (c2) are thin-walled sacs primarily lined by flattened type I cells and cuboidal type II cells. The stroma

Table 2: Survival analysis (c-index).

Method	LUAD	BLCA	BRCA	CRC
DSMIL [13]	$0.659_{0.07}$	$0.586_{0.06}$	$0.720_{0.06}$	$0.696_{0.11}$
CLAM [14]	$0.625_{0.12}$	$0.603_{0.06}$	$0.698_{0.03}$	$0.678_{0.09}$
DTFD-MIL [27]	$0.637_{0.08}$	$0.609_{0.08}$	$0.693_{0.05}$	$0.697_{0.09}$
TransMIL [17]	$0.660_{0.12}$	$0.616_{0.08}$	$0.708_{0.05}$	$0.686_{0.06}$
ILRA [23]	$\mathbf{0.688_{0.06}}$	$0.603_{0.04}$	$0.726_{0.08}$	$0.704_{0.09}$
PANTHER [19]	$0.632_{0.07}$	$0.612_{0.07}$	$0.729_{0.08}$	$0.632_{0.14}$
MambaMIL [25]	$0.670_{0.08}$	$0.606_{0.04}$	$0.668_{0.05}$	$0.680_{0.06}$
DGR-MIL [28]	$0.674_{0.05}$	$0.608_{0.04}$	$0.658_{0.05}$	$0.700_{0.09}$
PTCMIL (ours)	$\mathbf{0.688_{0.09}}$	$\mathbf{0.630_{0.05}}$	$\mathbf{0.745_{0.04}}$	$\mathbf{0.738_{0.09}}$

Table 3: Few-shot adaptation (%).

Pretrained on Fewshot on	TCGA-NSCLC Camelyon16		Camelyon16 TCGA-NSCLC	
	AUC	Acc	AUC	Acc
DSMIL [13]	$65.57_{8.81}$	$62.97_{6.76}$	$83.88_{6.74}$	$75.47_{6.15}$
CLAM [14]	$62.04_{10.63}$	$61.41_{4.12}$	$84.72_{5.36}$	$76.08_{5.15}$
DTFD-MIL [27]	$66.47_{5.89}$	$62.50_{5.21}$	$83.51_{6.61}$	$66.87_{8.70}$
TransMIL [17]	$57.28_{6.14}$	$53.44_{5.78}$	$66.00_{10.59}$	$61.36_{9.54}$
ILRA [20]	$50.15_{12.59}$	$53.91_{7.35}$	$53.73_{10.79}$	$53.46_{6.96}$
MambaMIL [25]	$65.45_{10.14}$	$65.31_{7.17}$	$83.93_{5.71}$	$75.62_{4.45}$
DGR-MIL [28]	$55.84_{4.20}$	$62.03_{1.05}$	$54.00_{3.41}$	$51.09_{2.77}$
PTCMIL (ours)	$\mathbf{69.49_{10.27}}$	$\mathbf{67.03_{10.53}}$	$\mathbf{85.73_{3.42}}$	$\mathbf{77.36_{3.89}}$

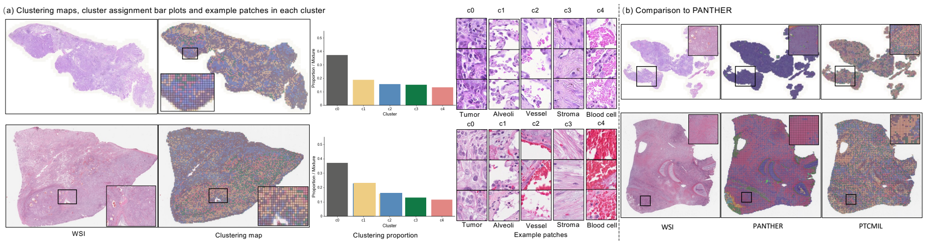
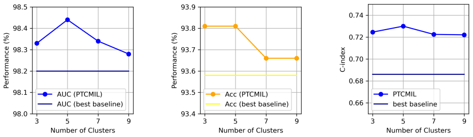


Fig. 2: Visualization and interpretation of PTCMIL. (a) Clustering maps, cluster assignment bar plots and example patches in each cluster; (b) Comparison to PANTHER.

(c3) consists of spindle-shaped cells within a collagen-rich extracellular matrix. c4 are mainly pools of red blood cells that appear as tightly packed, uniform red cells. Besides, in Fig. 2(b), the two-stage clustering MIL model PANTHER [18] shows clustering collapse (homogeneous colors, poor tissue separation), while PTCMIL produces more structured maps, better reflecting local heterogeneity.

3.5 Ablation Studies and Hyperparameter Analysis

We present ablation studies on key modules in Table 4. **Clustering (sec 1):** We evaluate the effectiveness of clustering in ViT-based MIL aggregation and its sensitivity to the number of clusters. PTCMIL (last line) achieves higher AUC and accuracy for classification on TCGA-NSCLC and improved survival prediction on CRC. **Merging (sec 2):** We assess the impact of merging tokens to create prototypes versus directly using prompt tokens. Merging reduces token redundancy and enhances WSI representation, leading to better performance. In contrast, using prompt tokens alone results in lower performance, as further illustrated in Fig. 1(c). **Pooling (sec 3):** While the cls token provides global image representation in ViT, integrating prototype tokens alongside the cls token enhances performance, showing the advantage of prototype-guided pooling. In Fig. 3, PTCMIL consistently outperforms the best baseline within a cluster range of 3 to 9, achieving peaks at number of cluster equals 5 on TCGA-NSCLC (classification) and CRC (survival).



(a) NSCLC (b) NSCLC (c) CRC

Fig. 3: Variation in the number of clusters.

Table 4: Ablation of clustering, merging, and pooling.

Clustering		Merging		Pooling			TCGA-NSCLC		TCGA-CRC
w/o	w/	w/o	w/	pro	cls	pro + cls	AUC	Acc	c-index
✓		–	–	–	–	–	$96.58_{0.45}$	$90.87_{0.68}$	$0.705_{0.07}$
	✓	✓				✓	$96.77_{0.79}$	$94_{1.87}$	$0.685_{0.08}$
	✓		✓	✓			$96.92_{0.72}$	$91.77_{1.01}$	$0.726_{0.07}$
	✓		✓		✓		$96.44_{0.54}$	$95_{0.87}$	$0.706_{0.07}$
	✓		✓			✓	$\mathbf{97.31_{0.67}}$	$\mathbf{92.17_{1.80}}$	$\mathbf{0.738_{0.09}}$

4 Conclusion

In this paper, we propose PTCMIL, an end-to-end clustering ViT-based MIL for WSI feature aggregation, addressing WSI's giga-pixel scale and heterogeneity. By introducing learnable prompt tokens and integrating clustering with prediction, PTCMIL handles WSI heterogeneity effectively. Experiments show superior performance across tasks and improved slide clustering where prior methods struggled. Our approach enables simultaneous prototype learning and task performance enhancement while identifying interpretable biomarkers. Future work will explore automatic cluster number selection for cancer types and the integration of vision-language models with clinical knowledge for guided clustering.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bejnordi, B.E., Veta, M., Van Diest, P.J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J.A., Hermsen, M., Manson, Q.F., Balkenhol, M., et al.: Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. Jama **318**(22), 2199–2210 (2017)
2. Bulten, W., Kartasalo, K., Chen, P.H.C., Ström, P., Pinckaers, H., Nagpal, K., Cai, Y., Steiner, D.F., Van Boven, H., Vink, R., et al.: Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. Nature medicine **28**(1), 154–163 (2022)
3. Cancer Genome Atlas Research Network, J., et al.: The cancer genome atlas pan-cancer analysis project. Nat. Genet **45**(10), 1113–1120 (2013)
4. Chan, T.H., Cendra, F.J., Ma, L., Yin, G., Yu, L.: Histopathology whole slide image analysis with heterogeneous graph representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15661–15670 (2023)
5. Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F.: Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16144–16155 (2022)

6. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine **30**(3), 850–862 (2024)
7. Chen, R.J., Lu, M.Y., Weng, W.H., Chen, T.Y., Williamson, D.F., Manz, T., Shady, M., Mahmood, F.: Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4015–4025 (October 2021)
8. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics. pp. 249–256. JMLR Workshop and Conference Proceedings (2010)
9. Hou, W., Yu, L., Lin, C., Huang, H., Yu, R., Qin, J., Wang, L.: Hˆ 2-mil: exploring hierarchical representation with heterogeneous multiple instance learning for whole slide image analysis. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 933–941 (2022)
10. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136. PMLR (2018)
11. Jain, A.K., Dubes, R.C.: Algorithms for clustering data. Prentice-Hall, Inc. (1988)
12. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: European Conference on Computer Vision. pp. 709–727. Springer (2022)
13. Li, B., Li, Y., Eliceiri, K.W.: 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. pp. 14318–14328 (2021)
14. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering **5**(6), 555–570 (2021)
15. Marin, D., Chang, J.H.R., Ranjan, A., Prabhu, A., Rastegari, M., Tuzel, O.: Token pooling in vision transformers for image classification. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 12–21 (2023)
16. Rao, Y., Zhao, W., Liu, B., Lu, J., Zhou, J., Hsieh, C.J.: Dynamicvit: Efficient vision transformers with dynamic token sparsification. Advances in neural information processing systems **34**, 13937–13949 (2021)
17. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in neural information processing systems **34**, 2136–2147 (2021)
18. Song, A.H., Chen, R.J., Ding, T., Williamson, D.F., Jaume, G., Mahmood, F.: Morphological prototyping for unsupervised slide representation learning in computational pathology. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11566–11578 (2024)
19. Song, A.H., Chen, R.J., Ding, T., Williamson, D.F., Jaume, G., Mahmood, F.: Morphological prototyping for unsupervised slide representation learning in computational pathology. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11566–11578 (June 2024)
20. Wang, H., Luo, L., Wang, F., Tong, R., Chen, Y.W., Hu, H., Lin, L., Chen, H.: Iteratively coupled multiple instance learning from instance to bag classifier for whole slide image classification. arXiv preprint arXiv:2303.15749 (2023)

21. Wang, J., Liu, Z., Zhao, L., Wu, Z., Ma, C., Yu, S., Dai, H., Yang, Q., Liu, Y., Zhang, S., et al.: Review of large vision models and visual prompt engineering. Meta-Radiology p. 100047 (2023)
22. Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., Han, X.: Transformer-based unsupervised contrastive learning for histopathological image classification. Medical image analysis **81**, 102559 (2022)
23. Xiang, J., Zhang, J.: Exploring low-rank property in multiple instance learning for whole slide image classification. In: The Eleventh International Conference on Learning Representations (2023)
24. Yan, J., Chen, H., Li, X., Yao, J.: Deep contrastive learning based tissue clustering for annotation-free histopathology image analysis. Computerized Medical Imaging and Graphics **97**, 102053 (2022)
25. Yang, S., Wang, Y., Chen, H.: Mambamil: Enhancing long sequence modeling with sequence reordering in computational pathology. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 296–306. Springer (2024)
26. Zeng, W., Jin, S., Liu, W., Qian, C., Luo, P., Ouyang, W., Wang, X.: Not all tokens are equal: Human-centric visual analysis via token clustering transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11101–11111 (2022)
27. Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., Yang, X., Coupland, S.E., Zheng, Y.: 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. pp. 18802–18812 (2022)
28. Zhu, W., Chen, X., Qiu, P., Sotiras, A., Razi, A., Wang, Y.: Dgr-mil: Exploring diverse global representation in multiple instance learning for whole slide image classification. In: European Conference on Computer Vision. pp. 333–351. Springer (2024)