Content-Aware Information Compression and Selection for Whole Slide Image Analysis

Tingting Zheng[1,2], Hongxun Yao[1,2*], Sicheng Zhao[3], Yi Xiao[4]
[1]School of Computer Science and Technology, Harbin Institute of Technology
[2]China Mobile 5G Application Innovation Joint Research, Harbin Institute of Technology
[3]Department of Psychological and Cognitive Sciences, Tsinghua University
[4]School of Computer Science and Artificial Intelligence, Zhengzhou University
23b903051@stu.hit.edu.cn, h.yao@hit.edu.cn, schzhao@tsinghua, yixiao@zzu.edu.cn

Abstract

Recent advances in multi-instance learning (MIL) have demonstrated impressive performance in whole slide image (WSI) analysis. However, current methods search for cues and draw conclusions from all instances or regions, resulting in excessive redundant computation and suboptimal representation quality due to irrelevant and uninformative feature interference. To address these issues, we propose CICS, an efficient and general framework that performs compact information compression and selection for high-efficiency WSI analysis. In particular, CICS features two key components: (1) context-aware compression (CAC), which partitions the instance space into sub-regions and applies learnable compression to discard irrelevant components, reduce computational complexity while facilitating information selection, and (2) global-proximity selective attention (GPSA), which cherry-picks the most informative representation with a proximity-assisted global dynamic selection strategy. Building upon these innovations, CICS forms a plug-and-play module that reduces computational complexity through compact instance representations while improving feature quality by preserving the most informative cues. Extensive experiments on six WSI classification and survival prediction datasets show that CICS consistently improves the performance of multiple representative MIL methods. It achieves 2.5%, 7.7%, and 3.9% accuracy gain over the state-of-the-art Transformer-based TransMIL, Mamba-based MambaMIL, and graph-based WIKG methods on the ESCA dataset.

Introduction

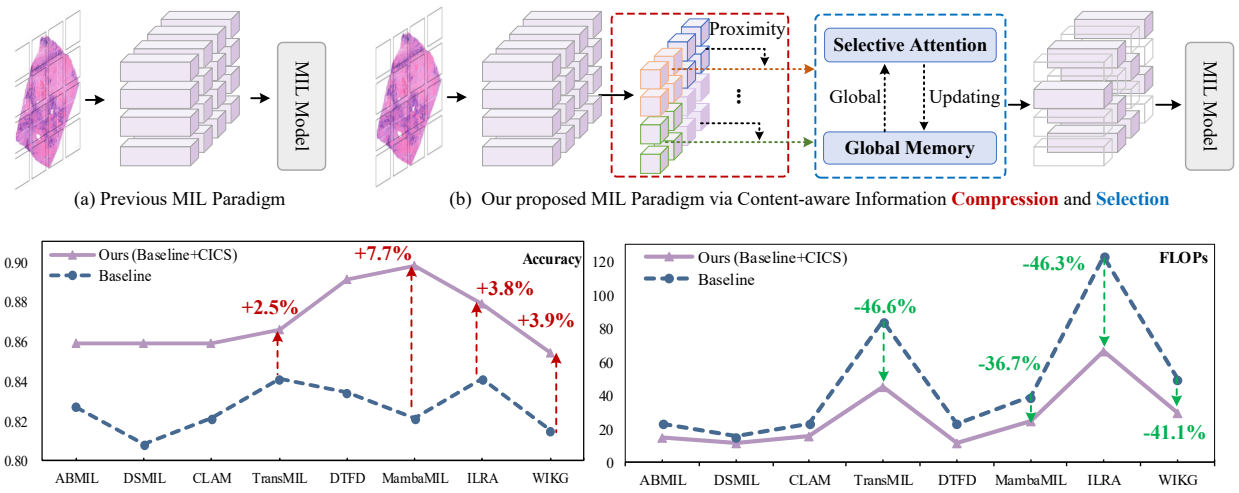
Histopathology slides remain the gold standard in cancer diagnosis and prognosis (Chen, Sun, and Zhao 2024; Quan et al. 2024). The advent of computational pathology has enabled their conversion into whole slide images (WSIs), facilitating automated detection of subtle pathological variations (Chen et al. 2024; Li et al. 2024a). However, the dispersed and heterogeneous nature of tumor regions in high-resolution WSIs makes pixel-level annotation labor-intensive and time-consuming (Tang, Zhang, and Li 2025; Zhuang et al. 2025; Wang et al. 2025). To address this challenge, multi-instance learning (MIL) has emerged as a widespread paradigm for WSI analysis, where WSIs are divided into unlabed patches (or instances), and treated as "bags" (Maron and Lozano-Pérez 1997; Zheng, Jiang, and Yao 2024), as shown in Figure 1 (a). Broadly, existing MIL methods fall into two categories: instance-level (Bin 2021; Fourkioti, De Vries, and Bakal 2024; Qu et al. 2022) and bag-level (Shao et al. 2021; Yang, Wang, and Chen 2024; Zhang et al. 2022; Zheng et al. 2025b) approaches.

Instance-level approaches predict labels for individual patches and aggregate them for the bag-level prediction (Campanella et al. 2019), but often overlook tissue structure and contextual relationships. This limitation becomes critical when analyzing WSIs with small tumor regions amidst abundant normal tissue, leading to unreliable predictions (Li et al. 2024b; Xiang and Zhang 2023). Bag-level methods address these issues using advanced techniques like attention mechanisms (Ilse, Tomczak, and Welling 2018; Gou et al. 2025), Transformers (Shao et al. 2021; Zheng et al. 2023), graph neural networks (Li et al. 2024b), and Mamba (Yang, Wang, and Chen 2024). While these approaches aggregate instances into unified bag representations, Transformer-based methods (*e.g.*, TransMIL (Shao et al. 2021) and ILRA (Xiang and Zhang 2023)) face significant computational challenges due to quadratic complexity relative to instance count (Xu et al. 2024). Moreover, their overemphasis on salient features while interference from semantically irrelevant or uninformative instances creates prediction biases, compromising model optimization and representation quality (Tang et al. 2023; Zhang et al. 2025).

Recent attempts to activate non-most salient features through teacher-student masking (Tang et al. 2023), post-intervention pre-training (Lin et al. 2023), and multi-branch attention policies (Zhang et al. 2025) struggle to balance efficiency with discriminative power. As shown in Figure 1 (c), our experiments demonstrate that eliminating redundant computations and irrelevant instances simultaneously improves classification accuracy while reducing GFLOPs. However, two key challenges persist: *(1)* the massive instance space (often $> 10^4$ instances/WSI) complicates efficient selection, and *(2)* diagnostically relevant instances are often insufficiently preserved.

To address the core challenge of excessive redundant computation and suboptimal representation in existing MIL

*Corresponding authors.
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.



(a) Previous MIL Paradigm

(b) Our proposed MIL Paradigm via Content-aware Information **Compression** and **Selection**

(c) Comparative results of accuracy and FLOPs for representative MIL methods on the TCGA-ESCA dataset.

Figure 1: Previous methods (a) that predict or aggregate across all instances suffer from redundant computation and suboptimal performance (c), especially quadratic complexity in Transformer-based methods (e.g., TransMIL (Shao et al. 2021) and ILRA (Xiang and Zhang 2023)) and multiple feature multiplications in Mamba-based MambaMIL (Yang, Wang, and Chen 2024) and graph-based WIKG (Li et al. 2024b). In contrast, integrating our CICS framework (b) into the MIL model improves accuracy while reducing giga floating-point operations (GFLOPs) via compact instance representations and selective attention.

methods, we propose a content-aware information compression and selection framework (CICS) to enable compact and high-quality representation of instances, thus upgrading conventional MIL-based WSI analysis, as depicted in Figure 1 (b). Specifically, a content-aware compression (CAC) strategy is developed to project high-dimensional instance features into compact representations while preserving discriminative information, reducing both sequence length and feature dimensionality. To effectively identify the most reliable instances, we propose the global-proximity selective attention (GPSA) module, which integrates local context with global memory embeddings to identify robust instances, dynamically updating representations through adaptive feature fusion.

Our main contributions are summarized as follows:

- We introduce a concise yet effective CICS framework that enhances discriminative power through optimized instance selection while reducing computational complexity for improved WSI analysis.
- A content-aware compression (CAC) is devised to project instance features into a compact space through a learnable compression function. Furthermore, we design a global-proximity selective attention (GPSA) mechanism to integrate local context and global patterns for reliable instance identification and robust representation.
- Extensive experiments on six public datasets validate the superiority of CICS, achieving accuracy gains of 2.1%, 2.5%, and 1.5% on BRCA, ESCA, and BRACS for classification, and improving the C-index by 5.1%, 3.4%, 0.8%, and 3.8% on BRCA, BLCA, LUSC, and LUAD for survival prediction, compared to TransMIL, while reducing computational cost.

Related Work

MIL predicts bag-level labels from unlabeled instances (patches). To address the computational burden of high-resolution WSIs, existing methods typically use offline feature encoders pre-trained on large-scale natural or pathology-specific datasets (Huang et al. 2023) to embed instances. These methods can be categorized into two main types: instance-level (Bin 2021) and bag-level prediction (Zheng, Jiang, and Yao 2024). Instance-level methods use bag labels as pseudo-labels to train the instance predictor, while bag-level methods integrate the attention mechanism (Ilse, Tomczak, and Welling 2018) with the global correlation modeling models (*e.g.*, Transformer (Shao et al. 2021), graph neural networks (Li et al. 2024b), Mamba (Zheng et al. 2025d,c,a)) to aggregate all instances for prediction. While effective, both approaches face critical limitations: offline features often contain irrelevant components that obscure diagnostically relevant instances (Song et al. 2024), and self-attention-based aggregation introduces extensive computation while risking prediction bias through overemphasis on globally salient features (Lin et al. 2023). Recent work has explored selective attention mechanisms (Xiao et al. 2024; Zhu et al. 2024), which aim to prioritize informative inputs for efficient learning. Popular techniques include gated filters (e.g., LSTM, RNN, and Mamba (Campanella et al. 2019; Liu et al. 2024b)) and top-K attention strategies (Lu et al. 2021). While top-K selection captures globally important features, it often struggles to capture locally diverse features, particularly in WSIs with complex tissue structures. In this work, our proposed model-agnostic CICS framework addresses these issues through synergistic feature compression and instance selection.

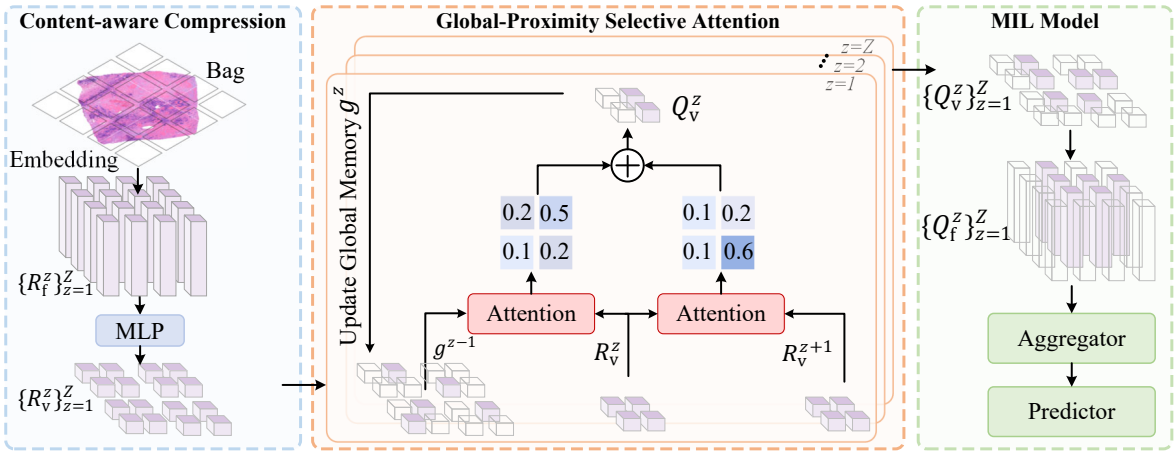


Figure 2: The architecture of our proposed CICS framework features a content-aware compression (CAC) module and a global-proximity selective attention (GPSA) mechanism.

Method

Preliminaries

Given a WSI $\{X, Y\}$, we divide the non-background tissue regions X into N non-overlapping instances $\{(x_j \in \mathbb{R}^{W\times H\times 3}, y_j) \mid 1 \le j \le N\}$ as a "bag", where Y and y_j denote the bag label and unknown label of the instance x_j, H and W are the height and width, respectively. The MIL paradigm for WSI analysis can be defined as:

$$Y = \begin{cases} 0, & \text{if } \forall\, x_j \in X,\ y_j = 0, \\ 1, & \text{if } \exists\, x_j \in X,\ y_j = 1. \end{cases} \tag{1}$$

To facilitate an efficient training process, instance-level methods often embed an instance x_j into a 1D-dimensional feature vector $f^j_{\text{ins}} \in \mathbb{R}^D$ using pre-trained online encoders (Huang et al. 2023; Lu et al. 2021; Zheng, Jiang, and Yao 2024). To better capture the complex tumor features within WSIs, bag-level methods (Bin 2021; Shao et al. 2021; Xiang and Zhang 2023) aggregate instance features $\{f^j_{\text{ins}}\}^N_{j=1}$ into a bag representation to predict the bag-wise label $\hat{Y}$ via a predictor $\mathcal{P}(\cdot)$. The procedures are formally expressed as:

$$\hat{Y} = \mathcal{P}\left(\mathcal{A}\left(\{f^j_{\text{ins}}\}^N_{j=1}\right)\right), \tag{2}$$

where $\mathcal{A}(\cdot)$ is typically derived from an attention-based aggregation operation (*e.g.*, attention (Ilse, Tomczak, and Welling 2018) and Transformer (Shao et al. 2021)). However, redundant computation and interference from massive irrelevant or low-contribution instances lead to high computational overhead and suboptimal performance. There still exists room for improved MIL methods.

Overview of CICS

Figure 2 illustrates our proposed CICS network, which aims to efficiently identify and retain informative instances, improving performance while reducing computational cost. Unlike prior methods that process all instances indiscriminately, CICS achieves highly efficient and reliable feature representation by abandoning irrelevant content (compression) while capturing the most informative instances (selection).

Specifically, the CAC module employs learnable compression functions to project each instance $f^j_{\text{ins}} \in \mathbb{R}^{1\times D}$ to a compact representation $v^j_{\text{ins}} \in \mathbb{R}^{1\times E}$ ($E = D/\tau$), where E and τ denotes the compression dimension and ratio, respectively. Spatial partitioning divides each bag into Z sub-regions $\{R^z_{\text{f}} \in \mathbb{R}^{L\times D}\}^Z_{z=1}$ (uncompressed) and $\{R^z_{\text{v}} \in \mathbb{R}^{L\times E}\}^Z_{z=1}$ (compressed), where $L = (N/Z)$ is number of features within each sub-region. To facilitate instance selection, global memory embedding $g^0 \in \mathbb{R}^{M\times E}$ is initialized from top-M high-attention instances of $\{R^z_{\text{v}}\}^Z_{z=1}$. Next, the proposed GPSA module selects the most informative instance Q^z_{v} within each region R^z_{v} by jointly leveraging global memory cues and proximate local context R^{z+1}_{v}. Moreover, global memory embedding is iteratively updated based on comparisons with Q^z_{v}. After identifying the most informative instance indices through compression and selection, the corresponding uncompressed instances $\{Q^z_{\text{f}}\}^Z_{z=1}$ are used for final bag-level prediction via standard MIL, achieving improved accuracy with minimal overhead.

Content-aware Compression

The core objective of CAC is to reduce redundant information while preserving discriminative properties within instances, which benefits the efficient selection process. Each instance embedding f^j_{ins} is compressed into v^j_{ins} via a multi-layer perceptron (MLP). To prevent compression from harming the discriminative power of uncompressed features, we minimize the discrepancy between attention scores before and after dimensionality reduction. The procedures are

formally expressed as:

$$\mathcal{L}_{\text{cac}} = \frac{1}{N}\sum_{j=1}^{N}\left(\text{SoftMax}\left(W_{\text{f}}^{j} f_{\text{ins}}^{j}\right) - \text{SoftMax}\left(W_{\text{v}}^{j} v_{\text{ins}}^{j}\right)\right)^2, \tag{3}$$

where $v_{\text{ins}}^{j} = \text{MLP}\left(f_{\text{ins}}^{j}\right)$, W_{f}^{j} and W_{v}^{j} are learnable parameters. To better improve prediction-related components in the compressed features, we introduce a task-specific loss $\mathcal{L}_{\text{task}}$ that filters irrelevant information.

Global-proximity Selective Attention

To efficiently identify the most informative instances, GPSA integrates both global discriminative cues and proximity correlations. Take the R_{v}^{z} as an example, the selection operation is formulated as:

$$\sigma_{\text{g}}^{z} = \frac{1}{h}\sum_{i=1}^{h}\text{SoftMax}\left(\frac{W_{\text{gq}}^{i} R_{\text{v}}^{z} \cdot (W_{\text{gk}}^{i} g^{z-1})^{\top}}{\sqrt{d_{gk}}}\right), \tag{4}$$

$$\sigma_{\text{r}}^{z} = \frac{1}{h}\sum_{i=1}^{h}\text{Softmax}\left(\frac{W_{\text{rq}}^{i} R_{\text{v}}^{z} \cdot (W_{\text{rk}}^{i} R_{\text{v}}^{z+1})^{\top}}{\sqrt{d_{rk}}}\right). \tag{5}$$

The $\sigma_{\text{g}}^{z} \in \mathbb{R}^{L}$ and $\sigma_{\text{r}}^{z} \in \mathbb{R}^{L}$ represent the attention scores of R_{v}^{z} to the global memory embedding g^{z-1} and the proximity region instances R_{v}^{z+1}, respectively, computed via multi-head attention mechanism (MHA), where h is the number of heads. The d_{gk} and d_{rk} represent the dimensions per head, while W_{gq}^{i}, W_{rq}^{i}, W_{gk}^{i}, and W_{rk}^{i} are the learnable parameters of the i-th head. Considering the limitation that selecting only salient instances tends to lose rich contextual information, we fuse σ_{g}^{z} and σ_{r}^{z} to adaptively retain instances in R_{v}^{z} whose scores exceed the average threshold. This selection procedure is expressed as:

$$Q_{\text{v}}^{z} = \left\{ v_{\text{ins}}^{z,j} \mid \left(\sigma_{\text{g}}^{z}(j) + \sigma_{\text{r}}^{z}(j)\right) > \sigma_{\text{average}} \right\}, \tag{6}$$

where $\sigma_{\text{average}} = \frac{1}{L}\sum_{l=1}^{L}\left(\sigma_{\text{g}}^{z}(j) + \sigma_{\text{r}}^{z}(j)\right)$ and $Q_{\text{v}}^{z} = \{v_{\text{ins}}^{z,s}\}_{s=1}^{S}$, and $S < L$ is the number of selected features in the z-th sub-region. Unlike the fixed top-K method, the adopted dynamic selection can better handle the variability and complexity of each region. To enhance the discriminative and contextual relevance of g^{z-1}, we update it by comparing the importance scores between Q_{v}^{z} and g^{z-1} using a linear and sigmoid function. Furthermore, the selected instances $\{Q_{\text{v}}^{z}\}_{z=1}^{Z}$ are processed through an MLP for bag prediction, and the loss $\mathcal{L}_{\text{task}}$ is incorporated to collaboratively optimize both instance selection and feature compression. These operations are formally represented as:

$$\mathcal{L}_{\text{CICS}} = (1-\alpha)\cdot\mathcal{L}_{\text{task}}\left(Y,\ \text{MLP}\left(\{Q_{\text{v}}^{z}\}_{z=1}^{Z}\right)\right) + \alpha\cdot\mathcal{L}_{\text{cac}}, \tag{7}$$

where α is a hyperparameter in the range $[0, 1]$, controlling the trade-off between discriminative and diverse components in the compressed space. Following the above process, the most informative instance indices are obtained. The uncompressed features $\{Q_{\text{f}}^{z}\}_{z=1}^{Z}$ are retrieved from $\{R_{\text{f}}^{z}\}_{z=1}^{Z}$ for MIL-based training and inference, preserving full feature fidelity with minimal computational cost.

CICS-based MIL Methods

CICS enables end-to-end optimization with any aggregator $\mathcal{A}(\cdot)$ and predictor $\mathcal{P}(\cdot)$ for specific downstream tasks. Unlike methods that require pre-training operation, CICS cooperatively optimizes both $\mathcal{A}(\cdot)$ and $\mathcal{P}(\cdot)$ models. The procedure is expressed as,

$$\{\hat{\theta}_{\text{CICS}}, \hat{\theta}_{\mathcal{A}}, \hat{\theta}_{\mathcal{P}}\} \leftarrow \mathcal{L}_{\text{task}}\left(Y,\ \mathcal{P}\left(\mathcal{A}\left(\{Q_{\text{f}}^{z}\}_{z=1}^{Z}\right)\right)\right) + \mathcal{L}_{\text{CICS}}, \tag{8}$$

where $\hat{\theta}_{\text{CICS}}$, $\hat{\theta}\mathcal{A}$, and $\hat{\theta}_{\mathcal{P}}$ represent the learnable parameters of the CAC and GPSA modules, the aggregator, and the predictor, respectively. After applying our CICS, the computational complexity of the Transformer-based aggregator reduces from $O(N^2D)$ to $O(\bar{N}^2D)$, where $\bar{N} < N$ denotes the number of selected instances from Z sub-regions in a bag. The additional computational costs of the CAC and GPSA modules are $O(ND'(D+E))$ and $O(\frac{N}{Z}^2 E)$, where D' is the hidden dimension. Usually, $N \gg D$ in most WSIs, which means that our CICS reduces computational complexity by a large margin.

Experiments

Datasets

Cancer Prediction. (1) TCGA Breast Cancer Dataset (BRCA). This dataset comprises 952 WSIs from the BRCA project, including 749 invasive ductal carcinoma (IDC) and 203 invasive lobular carcinoma (ILC) cases. Following the protocol in (Liu et al. 2024a), the data are split into training, validation, and test sets with a 65:10:25 ratio. Preprocessing involves extracting non-overlapping 256×256 patches at 10× magnification using CLAM. **(2) TCGA Esophageal Cancer Dataset (ESCA).** This cohort contains 156 diagnostic WSIs: 90 squamous cell carcinoma and 66 adenocarcinoma cases (Tomczak, Czerwińska, and Wiznerowicz 2015). Data are partitioned into training, validation, and test sets (3:1:1 ratio) as in (Zhu et al. 2022). Non-overlapping 256×256 patches at 20× magnification are generated via CLAM. **(3) Breast Carcinoma Subtyping Dataset (BRACS).** BRACS comprises 265 benign, 89 atypical, and 193 malignant breast tumors (Brancati et al. 2022). Aligned with the official data split (Brancati et al. 2022; Zhang et al. 2025), 395 WSIs are used for training, 65 WSIs for validation, and 87 WSIs for testing. Patch extraction follows the same CLAM-based (256×256 patches, 10× magnification).

Survival Prediction. Besides the typical BRCA dataset (Shao et al. 2023; Yang, Wang, and Chen 2024; Yang et al. 2025) for survival prediction, we incorporate three additional public datasets. **(1) TCGA Bladder Carcinoma Dataset (BLCA).** The BLCA dataset contains 376 cases of bladder urothelial carcinoma. **(2) TCGA Lung Adenocarcinoma Dataset (LUAD).** It comprises 541 LUAD WSIs from 478 cases. **(3) TCGA Squamous Cell Carcinoma Dataset (LUSC).** This dataset includes 512 LUSC WSIs from 478 cases. In accordance with (Tang et al. 2024), CLAM (Lu et al. 2021) generates non-overlapping 256×256 patches at 20× magnification.

Methods	BRCA			ESCA		
	Accuracy	AUC	F1	Accuracy	AUC	F1
ABMIL	0.852±0.024	0.885±0.041	0.908±0.017	0.827±0.092	0.914±0.066	0.859±0.079
+CICS	0.877±0.021 (+2.5)	0.894±0.026 (+0.9)	0.925±0.013 (+1.7)	0.859±0.036 (+3.2)	0.929±0.046 (+1.5)	0.883±0.028 (+2.4)
DSMIL	0.823±0.021	0.820±0.033	0.892±0.014	0.808±0.065	0.882±0.084	0.833±0.062
+CICS	0.843±0.022 (+2.0)	0.859±0.027 (+3.9)	0.904±0.012 (+1.2)	0.859±0.091 (+5.1)	0.913±0.096 (+3.1)	0.880±0.083 (+4.7)
CLAM	0.865±0.020	0.890±0.029	0.917±0.014	0.821±0.078	0.902±0.088	0.843±0.075
+CICS	0.872±0.013 (+0.7)	0.894±0.024 (+0.4)	0.920±0.009 (+0.3)	0.859±0.074 (+3.8)	0.922±0.072 (+2.0)	0.883±0.064 (+4.0)
TransMIL	0.847±0.021	0.846±0.036	0.905±0.013	0.841±0.101	0.910±0.083	0.864±0.084
+CICS	0.868±0.029 (+2.1)	0.876±0.017 (+3.0)	0.919±0.017 (+1.4)	0.866±0.098 (+2.5)	0.934±0.057 (+2.4)	0.894±0.071 (+3.0)
DTFD-MaxMin	0.816±0.023	0.810±0.033	0.885±0.013	0.834±0.110	0.881±0.145	0.875±0.074
+CICS	0.834±0.032 (+1.8)	0.823±0.042 (+1.3)	0.900±0.020 (+1.5)	0.859±0.117 (+2.5)	0.924±0.086 (+4.3)	0.885±0.097 (+1.0)
DTFD-AFS	0.823±0.028	0.824±0.034	0.892±0.017	0.834±0.091	0.927±0.053	0.862±0.069
+CICS	0.831±0.035 (+0.8)	0.842±0.056 (+1.8)	0.901±0.019 (+0.9)	0.891±0.048 (+5.7)	0.948±0.048 (+2.1)	0.907±0.043 (+4.5)
ILRA	0.857±0.035	0.886±0.026	0.908±0.027	0.841±0.098	0.901±0.091	0.857±0.089
+CICS	0.878±0.010 (+2.1)	0.879±0.040 (-0.7)	0.924±0.006 (+1.6)	0.879±0.077 (+3.8)	0.925±0.076 (+2.4)	0.891±0.075 (+3.4)
MambaMIL	0.868±0.017	0.878±0.032	0.917±0.009	0.821±0.098	0.908±0.074	0.838±0.092
+CICS	0.872±0.009 (+0.4)	0.885±0.017 (+0.7)	0.922±0.004 (+0.5)	0.898±0.097 (+7.7)	0.944±0.069 (+3.6)	0.915±0.081 (+7.7)
WIKG	0.863±0.018	0.887±0.030	0.914±0.013	0.815±0.107	0.876±0.100	0.845±0.094
+CICS	0.875±0.026 (+1.2)	0.895±0.039 (+0.8)	0.922±0.015 (+0.8)	0.854±0.122 (+3.9)	0.896±0.103 (+2.0)	0.864±0.130 (+1.9)

Table 1: Comparison of our results with MIL methods on BRCA and ESCA datasets. (%) shows CICS improvement.

Implementation Details

(1) Cancer Prediction. Following prior works (Lin et al. 2023; Liu et al. 2024a; Tang et al. 2024; Zhang et al. 2025; Zhu et al. 2022), we employ a ResNet18 (He et al. 2016) encoder pre-trained on ImageNet (Deng et al. 2009) (ResNet18-ImageNet) to extract 512-dimensional features for the BRCA, ESCA, and BRACS datasets. To further evaluate CICS robustness, we use 384-dimensional ViT-S/16-SSL features pretrained by DINO (Caron et al. 2021) on the BRACS dataset across three classification tasks. To optimize the cancer prediction networks, we use a bag-level cross-entropy loss. **(2) Survival Prediction.** For the BLCA, LUSC, and LUAD datasets, we use 512-dimensional feature vectors, obtained via the PLIP (Huang et al. 2023; Tang et al. 2024) foundation model pre-trained for pathology. Survival prediction models are trained with a Negative Log-Likelihood Survival Loss. All models are trained with the AdamW optimizer (Loshchilov and Hutter 2018), using a weight decay of $1e{-}5$, an initial learning rate of $1e{-}4$, and a batch size of 1 (*i.e.*, one bag per batch). Experiments are conducted using PyTorch (Paszke 2019) on a single NVIDIA RTX 3090 GPU.

Baseline and Evaluation Metrics

Baseline. We integrate our proposed CICS framework into 13 MIL-based methods for comparison, including attention-based approaches (ABMIL (Ilse, Tomczak, and Welling 2018), CLAM (Lu et al. 2021), DSMIL (Bin 2021), MHIM-ABMIL (Tang et al. 2023), IBMIL-ABMIL (Lin et al. 2023), ACMIL (Zhang et al. 2025)), Transformer-based methods (TransMIL (Shao et al. 2021), DTFD (Zhang et al. 2022), MHIM-TransMIL (Tang et al. 2023), IBMIL-TransMIL (Lin et al. 2023), ILRA (Xiang and Zhang 2023)), graph-based approaches like WIKG (Li et al. 2024b), and Mamba-based MambaMIL (Yang, Wang, and Chen 2024). All baselines are implemented using their official repositories.

Evaluation Metrics. (1) Cancer Prediction. For model evaluation, we use widely adopted performance metrics from (Bin 2021; Lin et al. 2023; Zheng, Jiang, and Yao 2024), including the area under the receiver operating characteristic curve (AUC), accuracy, and F1 score (F1) with a threshold set at 0.5. Following established protocols (Tang et al. 2023; Yang, Wang, and Chen 2024; Yu et al. 2023), we use 5-fold cross-validation to mitigate class imbalance and ensure reliable evaluation.For each fold, training and validation subsets are partitioned according to predefined dataset ratios for hyperparameter tuning. Final performance is reported as the mean classification performance and standard deviation across all test folds. **(2) Survival Prediction.** We report the C-index in all datasets. To reduce the impact of data split on model evaluation, we follow (Tang et al. 2024; Yang, Wang, and Chen 2024) and implement a 5-fold cross-validation approach, partitioning the data into training and validation subsets in a 4:1 ratio. We report the mean and standard deviation of the metrics over 5-folds.

Carcer Prediction

We evaluate CICS on two benchmark WSI datasets for two classification tasks, with results summarized in Table 1. As shown, our CICS consistently outperforms all baselines across all metrics. When integrated with CICS, the attention-based ABMIL (Ilse, Tomczak, and Welling 2018), DSMIL (Bin 2021), and CLAM (Lu et al. 2021) achieve average accuracy gains of 2.85%, 3.55%, and 2.25% on BRCA and ESCA datasets. Besides, by effectively reducing interference caused by semantically irrelevant instances, CICS boosts Transformer-based TransMIL (Shao et al. 2021), DTFD (Zhang et al. 2022), and ILRA (Xiang and Zhang 2023) by 2.30%, 3.25%, and 2.95% in average accuracy. Furthermore, building upon MambaMIL, CICS introduces instance selection to grasp the most informative representations, achieving notable gains of 4.05% in accuracy and 2.15% in AUC. To further validate the effectiveness and reliability of our CICS for three-classification tasks, we evaluate

Methods	Resnet18-ImageNet			ViT-S/16-SSL		
	Accuracy	AUC	F1	Accuracy	AUC	F1
ABMIL	0.691±0.041	0.816±0.020	0.604±0.055	0.750±0.047	0.883±0.028	0.662±0.050
+CICS	0.741±0.050 (+5.0)	0.838±0.042 (+2.2)	0.664±0.074 (+6.0)	0.775±0.039 (+2.5)	0.886±0.024 (+0.3)	0.678±0.042 (+1.6)
DSMIL	0.657±0.026	0.795±0.021	0.555±0.016	0.736±0.044	0.876±0.029	0.644±0.051
+CICS	0.676±0.046 (+1.9)	0.814±0.037 (+1.9)	0.585±0.030 (+2.3)	0.769±0.050 (+3.3)	0.881±0.035 (+0.5)	0.695±0.050 (+5.1)
CLAM	0.689±0.036	0.830±0.028	0.601±0.024	0.747±0.038	0.877±0.032	0.684±0.045
+CICS	0.719±0.029 (+3.0)	0.842±0.026 (+1.2)	0.637±0.065 (+3.6)	0.775±0.040 (+2.8)	0.890±0.030 (+1.3)	0.698±0.046 (+1.4)
TransMIL	0.706±0.044	0.799±0.035	0.596±0.036	0.767±0.029	0.886±0.036	0.671±0.042
+CICS	0.717±0.041 (+1.1)	0.832±0.046 (+3.3)	0.613±0.047 (+1.7)	0.786±0.029 (+1.9)	0.891±0.027 (+0.5)	0.715±0.042 (+4.4)
DTFD-MaxMin	0.698±0.030	0.815±0.062	0.610±0.044	0.741±0.037	0.870±0.038	0.662±0.051
+CICS	0.747±0.049 (+4.9)	0.844±0.037 (+2.9)	0.678±0.069 (+6.8)	0.763±0.029 (+2.2)	0.888±0.032 (+1.8)	0.666±0.043 (+0.4)
DTFD-AFS	0.676±0.056	0.823±0.039	0.614±0.054	0.760±0.043	0.881±0.027	0.665±0.057
+CICS	0.728±0.050 (+5.2)	0.858±0.030 (+3.5)	0.659±0.066 (+4.5)	0.769±0.034 (+0.9)	0.891±0.035 (+1.0)	0.691±0.044 (+2.6)
ILRA	0.732±0.076	0.847±0.050	0.650±0.094	0.767±0.034	0.874±0.025	0.678±0.021
+CICS	0.737±0.033 (+0.5)	0.853±0.031 (+0.6)	0.652±0.045 (+0.2)	0.771±0.041 (+0.4)	0.877±0.033 (+0.3)	0.707±0.067 (+2.9)
MambaMIL	0.706±0.066	0.834±0.039	0.636±0.071	0.748±0.042	0.865±0.017	0.646±0.064
+CICS	0.728±0.062 (+2.2)	0.841±0.041 (+0.7)	0.658±0.080 (+2.2)	0.767±0.007 (+1.9)	0.867±0.026 (+0.2)	0.687±0.022 (+4.1)
WIKG	0.706±0.053	0.837±0.030	0.637±0.055	0.762±0.049	0.881±0.031	0.674±0.068
+CICS	0.726±0.042 (+2.0)	0.847±0.040 (+1.0)	0.657±0.061 (+2.0)	0.773±0.029 (+1.1)	0.883±0.043 (+0.2)	0.703±0.042 (+2.9)

Table 2: Comparison of our results with MIL methods on the BRACS dataset. (%) shows CICS improvement.

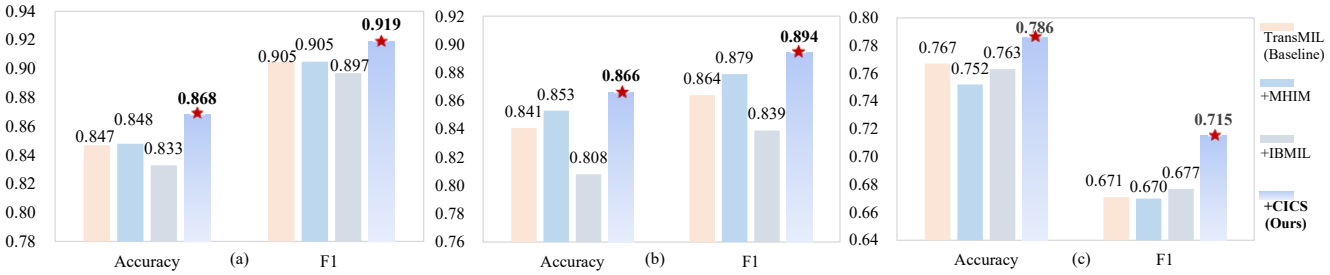


Figure 3: Comparison of our CICS and plug-and-play MIL methods on (a) BRCA, (b) ESCA, and (c) BRACS datasets.

it on the BRACS dataset using different encoders. As presented in Table 2, CICS achieves 3.80%, 3.55%, and 3.15% average F1 improvements over ABMIL (Ilse, Tomczak, and Welling 2018), DTFD (Zhang et al. 2022), and MambaMIL (Yang, Wang, and Chen 2024), respectively. These improvements mainly stem from the effective compression of instance interference in the challenging BRACS dataset, where advanced MIL paradigms such as MHIM (Tang et al. 2023) and IBMIL (Lin et al. 2023) fail to extract compact and reliable representations, leading to sub-optimal performance (as shown in Figure 3).

Survival Prediction

Table 3 demonstrates that our CICS framework achieves considerable improvements in C-Index over typical MIL methods. For the four survival analysis datasets, CICS boosts the classical MaxMIL and ABMIL (Ilse, Tomczak, and Welling 2018) by an average of 1.73% and 1.08%, respectively. Notably, by mitigating the interference computation and selecting the informative instances, CICS facilitates a compact yet reliable bag representation, significantly improving Transformer-based TransMIL (Shao et al. 2021), DTFD (Zhang et al. 2022), and ILRA (Xiang and Zhang 2023) by 3.28%, 1.25%, and 1.40% on average. These results highlight the robustness and superiority of our approach, validating its effectiveness.

Ablation Studies

We perform ablation studies to evaluate the effectiveness of GPSA and CAC components, using ABMIL (Ilse, Tomczak, and Welling 2018) as the baseline. To assess the contribution of both global and proximity cues in GPSA, we develop two variants: global memory-based self-attention (*w* GSA) and proximity-based self-attention (*w* PSA). To examine the effect of content-aware design in CAC for effective feature compression, we create three additional variants: *w/o* $\mathcal{L}_{\text{cac}}$ (without the $\mathcal{L}_{\text{cac}}$ loss in Eq. (7)), *w* RMask (compression via random masking), and *w* Top-K. (compression via top-K selection)

Validation on Basic Components. As shown in Table 4, our complete CICS model significantly outperforms its incomplete variants. For example, on the ESCA dataset, integrating *w* GSA and *w* PSA improves accuracy by 1.9% and 2.0%, respectively, compared to the baseline. These results highlight the limitations of global-attention-based MIL methods, which struggle to model both long-range and local instance relationships in large-scale WSIs with complex tissue structures. By synergistically combining GSA

Methods	BRCA		BLCA		LUSC		LUAD	
	Baseline	+CICS	Baseline	+CICS	Baseline	+CICS	Baseline	+CICS
MaxMIL	0.610±0.075	0.622±0.079	0.566±0.047	0.578±0.020	0.560±0.050	0.572±0.040	0.597±0.042	0.630±0.038
MeanMIL	0.616±0.070	0.631±0.054	0.592±0.036	0.595±0.047	0.538±0.033	0.540±0.037	0.634±0.072	0.638±0.067
ABMIL	0.620±0.042	0.630±0.050	0.581±0.035	0.591±0.032	0.581±0.044	0.585±0.044	0.625±0.047	0.644±0.042
DSMIL	0.647±0.077	0.654±0.033	0.604±0.034	0.607±0.047	0.578±0.055	0.590±0.037	0.640±0.068	0.645±0.041
CLAM	0.654±0.036	0.660±0.054	0.600±0.038	0.609±0.065	0.578±0.045	0.584±0.009	0.630±0.032	0.639±0.058
TransMIL	0.618±0.069	**0.669±0.075**	0.592±0.037	**0.626±0.013**	0.597±0.048	0.605±0.037	0.626±0.047	**0.664±0.038**
DTFD	0.653±0.056	0.659±0.051	0.611±0.037	0.625±0.025	0.587±0.062	**0.613±0.038**	0.646±0.044	0.650±0.052
ILRA	0.681±0.056	0.683±0.055	0.595±0.025	0.625±0.027	0.585±0.042	0.598±0.043	0.652±0.041	0.663±0.026
MambaMIL	0.668±0.047	0.670±0.041	0.606±0.002	0.608±0.027	0.598±0.058	0.604±0.065	0.674±0.040	0.675±0.053
WIKG	0.669±0.055	0.678±0.062	0.612±0.033	0.617±0.019	0.572±0.052	0.582±0.027	0.677±0.037	0.679±0.035

Table 3: Comparison of survival prediction results with MIL methods on four datasets. Bold indicates the best improvement.

Models	$\mathcal{L}_{\text{cac}}$	GSA	PSA	ESCA			BRACS		
				Accuracy	AUC	F1	Accuracy	AUC	F1
Baseline(ABMIL)	✗	✗	✗	0.827±0.092	0.914±0.066	0.859±0.079	0.691±0.041	0.816±0.020	0.604±0.055
w GSA	✓	✓	✗	0.846±0.086	0.922±0.072	0.864±0.084	0.710±0.012	0.804±0.011	0.618±0.046
w PSA	✓	✗	✓	0.847±0.064	0.916±0.074	0.868±0.057	0.728±0.046	0.838±0.034	0.629±0.095
w/o $\mathcal{L}_{\text{cac}}$	✗	✓	✓	0.853±0.054	0.913±0.053	0.870±0.054	0.728±0.046	0.838±0.034	0.629±0.095
w RMask	✗	✓	✓	0.801±0.041	0.897±0.078	0.832±0.046	0.708±0.053	0.826±0.043	0.595±0.074
w Top-K	✗	✓	✓	0.821±0.092	0.908±0.067	0.812±0.095	0.693±0.036	0.813±0.019	0.620±0.043
Baseline+CICS	✓	✓	✓	0.859±0.074	0.926±0.064	0.885±0.054	0.741±0.050	0.838±0.042	0.664±0.074

Table 4: Validation of basic components for ABMIL+CICS on ECSA and BRACS datasets.

Methods	$Z = 2 \times 2$		$Z = 3 \times 3$		$Z = 4 \times 4$		$Z = 5 \times 5$	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
$M = 2$	0.802	0.825	0.833	0.857	0.802	0.837	0.834	0.855
$M = 3$	0.833	0.864	0.859	0.875	0.853	0.883	0.840	0.860
$M = 5$	0.827	0.849	0.808	0.825	0.872	0.889	0.879	0.896
$M = 10$	0.802	0.840	0.827	0.840	0.840	0.862	0.808	0.830
$M = 20$	0.796	0.821	0.834	0.862	0.808	0.834	0.802	0.832

Table 5: Impact of global memory size (M) and region count (Z) on ABMIL+CICS performance (ESCA dataset).

and PSA (Baseline+CICS), our framework generates comprehensive representations that surpass the baseline, achieving improvements of 3.2% (accuracy) and 2.6% (F1) on the ESCA dataset, and 5.0% (accuracy) and 6.0% (F1) on the BRACS dataset. Furthermore, the proposed CAC outperforms naive strategies such as random masking (*w* RMask) and top-K selection (*w* Top-K), underscoring its ability to retain discriminative features while ensuring compactness. Removing the $\mathcal{L}_{\text{cac}}$ loss (*w/o* $\mathcal{L}_{\text{cac}}$) leads to a significant performance drop in F1 of 1.5%, respectively, emphasizing its critical role in preserving diagnostically relevant content during compression.

Hyperparameter Analysis. We investigate the effects of two key hyperparameters: the number of region Z and the size of global memory embedding M. By integrating our CICS with ABMIL (Ilse, Tomczak, and Welling 2018) method, we observe substantial performance improvements with larger Z and M (Table 5). Specifically, increasing Z allows the GPSA module to capture both globally dependent and locally correlated instances, enabling richer bag representations. For example, partitioning the bag into finer subregions improves feature discrimination in complex tissue structures. Besides, expanding M yields accuracy gains of 3.1%, 2.6%, 7.0%, and 4.5% compared to $M = 2$.

Conclusion

We propose CICS, a concise and general framework for multi-instance learning (MIL) that enhances computational efficiency and representation quality. Our first innovation, context-aware compression (CAC), eliminates redundant features in large-scale instance bags while preserving discriminative patterns, enabling efficient instance selection. Second, the global-proximity selective attention (GPSA) module integrates both long-range dependencies and local structural relationships, generating comprehensive bag representations for robust predictions. Experiments across cancer sub-typing and survival prediction benchmarks demonstrate that CICS achieves state-of-the-art accuracy. The generalizability of the framework is further evidenced by its applicability to domains requiring efficient long-sequence modeling, such as computational pathology and genomics. A current limitation of CICS lies in its sensitivity to fixed compression dimensions, potentially affecting adaptability across diverse datasets. Future work will investigate dynamic compression strategies to automatically optimize feature retention, enhancing robustness and scalability.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (No. 62476069) and in part by the Natural Science Foundation of Heilongjiang Province of China for Doctoral Students under Grant BS2025F001.

References

Bin, K. W. E., Li. Yin Li. 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.

Brancati, N.; Anniciello, A. M.; Pati, P.; Riccio, D.; Scognamiglio, G.; Jaume, G.; De Pietro, G.; Di Bonito, M.; Foncubierta, A.; Botti, G.; et al. 2022. Bracs: A dataset for breast carcinoma subtyping in h&e histology images. *Database*, 2022: baac093.

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.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9650–9660.

Chen, K.; Sun, S.; and Zhao, J. 2024. CaMIL: Causal Multiple Instance Learning for Whole Slide Image Classification. In *Proceedings of the AAAI conference on artificial intelligence*, 1120–1128.

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. 2024. Towards a general-purpose foundation model for computational pathology. *Nature medicine*, 30(3): 850–862.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 248–255.

Fourkioti, O.; De Vries, M.; and Bakal, C. 2024. CAMIL: Context-Aware Multiple Instance Learning for Cancer Detection and Subtyping in Whole Slide Images. In *International conference on learning representations*.

Gou, J.; Ji, L.; Liu, P.; and Ye, M. 2025. Queryable Prototype Multiple Instance Learning with Vision-Language Models for Incremental Whole Slide Image Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 3158–3166.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 770–778.

Huang, Z.; Bianchi, F.; Yuksekgonul, M.; Montine, T. J.; and Zou, J. 2023. 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. ISBN 2640-3498.

Li, H.; Zhang, Y.; Chen, P.; Shui, Z.; Zhu, C.; and Yang, L. 2024a. Rethinking Transformer for Long Contextual Histopathology Whole Slide Image Analysis. *Advances in neural information processing systems*.

Li, J.; Chen, Y.; Chu, H.; Sun, Q.; Guan, T.; Han, A.; and He, Y. 2024b. Dynamic graph representation with knowledge-aware attention for histopathology whole slide image analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11323–11332.

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.; Zhang, X.; and Ye, F. 2024a. Pseudo-Bag Mixup Augmentation for Multiple Instance Learning-Based Whole Slide Image Classification. *IEEE transactions on medical imaging*.

Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; and Liu, Y. 2024b. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37: 103031–103063.

Loshchilov, I.; and Hutter, F. 2018. Decoupled Weight Decay Regularization. In *International conference on learning representations*.

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.

Maron, O.; and Lozano-Pérez, T. 1997. A framework for multiple-instance learning. *Advances in neural information processing systems*, 10.

Paszke, A. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Qu, L.; Wang, M.; Song, Z.; et al. 2022. Bi-directional weakly supervised knowledge distillation for whole slide image classification. *Advances in neural information processing systems*, 35: 15368–15381.

Quan, H.; Li, X.; Chen, W.; Bai, Q.; Zou, M.; Yang, R.; Zheng, T.; Qi, R.; Gao, X.; and Cui, X. 2024. Global contrast-masked autoencoders are powerful pathological representation learners. *Pattern recognition*, 156: 110745.

Shao, Z.; Bian, H.; Chen, Y.; Wang, Y.; Zhang, J.; and Ji, X. 2021. Transmil: Transformer Based Correlated Multiple Instance Learning for Whole Slide Image Classification. *Advances in neural information processing systems*, 34: 2136–2147.

Shao, Z.; Chen, Y.; Bian, H.; Zhang, J.; Liu, G.; and Zhang, Y. 2023. Hvtsurv: Hierarchical vision transformer for patient-level survival prediction from whole slide image. In *Proceedings of the AAAI conference on artificial intelligence*, 2209–2217.

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.

Tang, W.; Huang, S.; Zhang, X.; Zhou, F.; Zhang, Y.; and Liu, B. 2023. 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*, 4078–4087.

Tang, W.; Zhou, F.; Huang, S.; Zhu, X.; Zhang, Y.; and Liu, B. 2024. Feature re-embedding: Towards foundation model-level performance in computational pathology. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11343–11352.

Tang, Z.; Zhang, X.; and Li, C. 2025. From Representation Space to Prognostic Insights: Whole Slide Image Generation with Hierarchical Diffusion Model for Survival Prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 7329–7337.

Tomczak, K.; Czerwińska, P.; and Wiznerowicz, M. 2015. Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary oncology onkologia*, 2015(1): 68–77.

Wang, H.; Du, X.; Liu, J.; Ouyang, S.; Chen, Y.-W.; and Lin, L. 2025. M2OST: Many-to-one Regression for Predicting Spatial Transcriptomics from Digital Pathology Images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 7709–7717.

Xiang, J.; and Zhang, J. 2023. Exploring low-rank property in multiple instance learning for whole slide image classification. In *International conference on learning representations*.

Xiao, Y.; Yuan, Q.; Jiang, K.; He, J.; Lin, C.-W.; and Zhang, L. 2024. TTST: A Top-k Token Selective Transformer for Remote Sensing Image Super-Resolution. *IEEE transactions on image processing*, 33: 738–752.

Xu, R.; Yang, S.; Wang, Y.; Cai, Y.; Du, B.; and Chen, H. 2024. Visual mamba: A survey and new outlooks. *arXiv:2404.18861*.

Yang, S.; Wang, Y.; and Chen, H. 2024. Mambamil: Enhancing long sequence modeling with sequence reordering in computational pathology. In *International conference on medical image cmputing and computer assisted intervention*, 296–306.

Yang, Z.; Wei, T.; Liang, Y.; Yuan, X.; Gao, R.; Xia, Y.; Zhou, J.; Zhang, Y.; and Yu, Z. 2025. A foundation model for generalizable cancer diagnosis and survival prediction from histopathological images. *Nature communications*, 16(1): 2366.

Yu, J.-G.; Wu, Z.; Ming, Y.; Deng, S.; Wu, Q.; Xiong, Z.; Yu, T.; Xia, G.-S.; Jiang, Q.; and Li, Y. 2023. Bayesian collaborative learning for whole-slide image classification. *IEEE transactions on medical imaging*.

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*, 18780–18790.

Zhang, Y.; Li, H.; Sun, Y.; Zheng, S.; Zhu, C.; and Yang, L. 2025. Attention-challenging multiple instance learning for whole slide image classification. In *European conference on computer vision*, 125–143.

Zheng, T.; Chen, W.; Li, S.; Quan, H.; Zou, M.; Zheng, S.; Zhao, Y.; Gao, X.; and Cui, X. 2023. Learning how to detect: A deep reinforcement learning method for whole-slide melanoma histopathology images. *CMIG*, 108: 102275.

Zheng, T.; Jiang, K.; Xiao, Y.; Zhao, S.; and Yao, H. 2025a. M3amba: Memory Mamba is All You Need for Whole Slide Image Classification. In *CVPR*, 15601–15610.

Zheng, T.; Jiang, K.; and Yao, H. 2024. Dynamic Policy-Driven Adaptive Multi-Instance Learning for Whole Slide Image Classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8028–8037.

Zheng, T.; Jiang, K.; Yao, H.; Xiao, Y.; and Wang, Z. 2025b. OODML: Whole Slide Image Classification Meets Online Pseudo-Supervision and Dynamic Mutual Learning. In *Proceedings of the AAAI conference on artificial intelligence*, 10626–10634.

Zheng, T.; Yao, H.; Jiang, K.; Xiao, Y.; and Zhao, S. 2025c. GMMamba: Group Masking Mamba for Whole Slide Image Classification. In *ICCV*, 9935–9944.

Zheng, T.; Yao, H.; Zhao, S.; Jiang, K.; and Xiao, Y. 2025d. GraphMamba: Whole slide image classification meets graph-driven selective state space model. *PR*, 111768.

Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; and Wang, X. 2024. Vision mamba: Efficient visual representation learning with bidirectional state space model. *Proceedings of the 41st international conference on machine learning*, 14.

Zhu, Z.; Yu, L.; Wu, W.; Yu, R.; Zhang, D.; and Wang, L. 2022. Murcl: Multi-instance reinforcement contrastive learning for whole slide image classification. *IEEE transactions on medical imaging*.

Zhuang, Z.; Cen, M.; Li, Y.; Zhou, F.; Yu, L.; Magnier, B.; and Wang, L. 2025. Dynamic Entity-Masked Graph Diffusion Model for Histopathology Image Representation Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 11058–11066.