Less Is More: Proportional Memory-Guided Differential-Attention MIL for Whole-Slide Image Classification

Hongpeng Yang[1], **Yingxin Chen**[2], **Shiqiang Ma**[3*] and **Fei Guo**[2*]
[1]Molinaroli College of Engineering and Computing, University of South Carolina
[2]School of Computer Science and Engineering, Central South University
[3]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
hongpeng@email.sc.edu, chenyingxin@csu.edu.cn, sq.ma@siat.ac.cn, guofei@csu.edu.cn

Abstract

Whole-slide images (WSIs) provide gigapixel-scale visual evidence for cancer diagnosis, yet diagnostically relevant regions are typically sparse and embedded within large amounts of weakly informative tissue. Existing multiple instance learning methods often aggregate all patches by using attention mechanisms or sequence modeling, leading to redundant computation and limited adaptability to slides with highly variable sizes and lesion burdens. We propose Proportional Memory-guided Differential-attention Multiple Instance Learning (PMDMIL) network, a scalable MIL framework that follows a "less is more" principle for WSI classification. PMDMIL retains a fixed ratio of patches, pre-filters noise with a class-wise memory bank, and fuses the sparse survivors through differential attention to amplify subtle yet decisive morphological differences. Experiments on five public WSI benchmarks demonstrate that PMDMIL consistently outperforms representative MIL and sequence-based baselines across multiple evaluation metrics. Notably, on the DHMC dataset, our model can achieve competitive performance, while processing only 10% of patches, indicating that effective instance selection is more important than exhaustive aggregation for WSI classification.

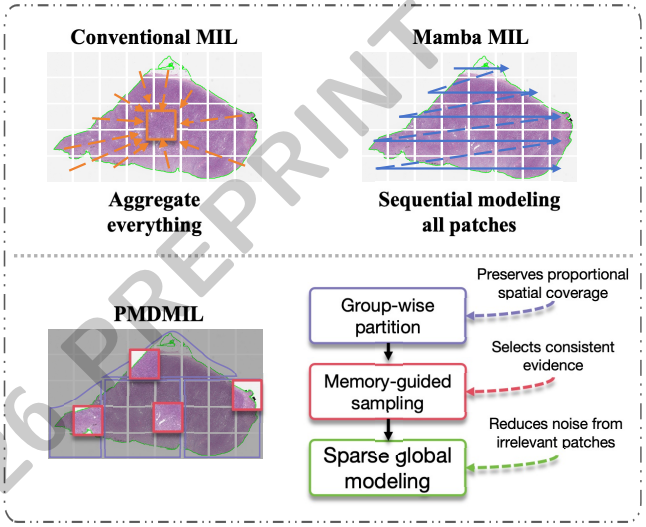

Figure 1: Conventional MIL and sequence modeling approaches process all patches, which not only incurs redundant computation but also dilutes sparse discriminative evidence with a large amount of weakly informative instances. PMDMIL performs group-wise proportional sampling, which preserves spatial organization while removing redundant patches, enabling efficient and discriminative slide-level representation learning.

1 Introduction

Histopathological examination of tissue specimens is the gold standard for cancer diagnosis, subtyping, and grading, providing critical insights into tissue morphology and cellular organization [Campanella *et al.*, 2019; Di *et al.*, 2022]. With the advent of digital pathology, conventional glass slides can now be scanned at high resolution to produce whole-slide images (WSIs), enabling large-scale computational analysis of histopathological data [He *et al.*, 2012]. However, learning effective representations from WSIs remains fundamentally challenging. WSIs are gigapixel-scale images that exhibit extreme variability in tissue area, pronounced intra-slide heterogeneity, and sparse diagnostically relevant regions embedded in large amounts of normal or weakly informative background tissue [Srinidhi *et al.*, 2021]. Moreover, acquiring fine-grained annotations at the pixel or patch level is prohibitively expensive and rarely available in routine clinical practice [Lu *et al.*, 2021]. These factors jointly make direct supervised learning on WSIs impractical.

To address these limitations, multiple instance learning (MIL) has emerged as the dominant framework for WSI classification. In MIL, each WSI is modeled as a bag of patch-level instances, and learning is performed using only slide-level labels [Zheng *et al.*, 2023; Zheng *et al.*, 2024; He *et al.*, 2026]. Early MIL approaches relied on simple pooling strategies, while subsequent works introduced attention-based mechanisms and more sophisticated aggregation architectures to learn instance importance and improve slide-level predictions [Ilse *et al.*, 2018; Li *et al.*, 2021; Xiang and Zhang, 2023; Zheng *et al.*, 2025].

*Corresponding authors sq.ma@siat.ac.cn, guofei@csu.edu.cn

However, existing MIL approaches still suffer from fundamental limitations. Most methods either aggregate all available instances or retain a fixed number of top-ranked patches based on local classifier confidence. Such designs are poorly aligned with the intrinsic properties of WSIs: fixed instance budgets fail to adapt to the extreme variability in slide size and lesion burden, while confidence-based instance selection is often unstable and locally myopic. As a consequence, informative regions may be discarded on large slides, whereas redundant or irrelevant patches may dominate the representation on smaller or less pathological slides.

More importantly, recent studies suggest that the core challenge of WSI analysis is not merely how to aggregate instance representations, but which instances should be treated as reliable evidence for slide-level prediction [Wang *et al.*, 2025]. Not all patches contribute equally to the final decision: visually dominant or high-confidence regions are not necessarily the most informative, while small and spatially localized regions may carry decisive diagnostic cues.

In this work, we propose Proportional Memory-guided Differential-attention Multiple Instance Learning network (PMDMIL), a scalable MIL framework that reconsiders how evidence should be selected and organized for whole-slide image classification (Figure 1). Motivated by the sparsity and spatial structure of diagnostic evidence in WSIs, PMDMIL performs group-wise proportional instance selection to adaptively retain informative patches across slides with varying sizes and lesion extents. A class-wise prototype memory is introduced to guide instance selection, encouraging the retention of patches that are both informative and consistent at the class level. The selected instances are then integrated through a differential-attention mechanism [Ye *et al.*, 2024; Hammoud and Ghanem, 2025] with coordinate-based positional encoding [Vaswani *et al.*, 2017], enabling effective modeling of spatially distributed evidence using a compact token set. Experiments on five public WSI classification benchmarks demonstrate that PMDMIL consistently outperforms representative MIL baselines, while ablation and efficiency analyses confirm that its components play complementary roles in improving performance and scalability. Our main contributions are summarized as follows:

- We propose PMDMIL, a proportional memory-guided differential-attention MIL framework for scalable WSI classification.
- We introduce a memory-guided instance sampler that maintains a class-wise prototype memory bank and enables adaptive, proportional instance selection.
- We design a differential-attention refiner operating on spatially grouped tokens with coordinate-based rotary positional encoding to capture inter-region relations and global tissue context.

2 Related Work

2.1 MIL-based Whole-Slide Image Classification

MIL is widely adopted for WSI classification task [Campanella *et al.*, 2019; Chen *et al.*, 2022; Coudray *et al.*, 2018], where only slide-level labels are available. Most existing MIL approaches focus on designing aggregation mechanisms to combine patch-level embeddings into a slide-level representation. Representative methods include attention-based MIL models, hierarchical aggregation frameworks, and more recent transformer- or graph-based architectures for modeling instance relationships across WSIs [Yu *et al.*, 2025].

Several studies investigate attention mechanisms to estimate instance importance and emphasize discriminative patches for slide-level prediction. Beyond simple attention-based pooling, structured MIL architectures have been proposed to improve robustness and incorporate contextual information among instances, for example by modeling spatial or relational dependencies [Zhang *et al.*, 2024; Fang *et al.*, 2024]. While effective, these approaches generally operate on all instances or rely on fixed aggregation schemes, which can be inefficient for large-scale WSIs.

2.2 Memory-Based and Prototype-Guided Models

Memory mechanisms [Wu *et al.*, 2018; Caron *et al.*, 2020; Wei *et al.*, 2025; Zhou *et al.*, 2023] and prototype [Song *et al.*, 2024] representations have been extensively studied in metric learning, few-shot learning, and self-supervised representation learning, where representative features are maintained to provide stable semantic references and improve generalization [Snell *et al.*, 2017; Fang *et al.*, 2025; Ho *et al.*, 2023]. In computational pathology, prototype-based ideas have been explored to capture characteristic morphological patterns or to regularize feature learning across heterogeneous tissue regions [Adnan *et al.*, 2022; Yu *et al.*, 2023; Yao *et al.*, 2020].

Despite these efforts, most existing MIL methods for WSIs do not explicitly maintain a class-level memory across slides to guide instance selection or aggregation. As a result, instance importance is often estimated independently for each slide, without leveraging accumulated class-consistent evidence from the dataset.

3 Methodology

We propose PMDMIL for supervised WSI classification. It adopts a group-first processing strategy with memory-guided proportional instance selection and differential-attention refinement, enabling robust and scalable slide-level representation learning. Figure 2 illustrates the process setting and contrasts PMDMIL with conventional MIL pipelines, while Figure 3 details the overall architecture of PMDMIL.

3.1 Problem Formulation

A WSI is represented as a bag of patch features and their spatial coordinates,

$$\mathcal{X} = \{(x_i, p_i)\}_{i=1}^{N}, \tag{1}$$

where $x_i \in \mathbb{R}^{d_{\text{in}}}$ denotes a patch feature extracted by a pretrained patch encoder and $p_i \in \mathbb{R}^2$ denotes its spatial location on the slide. Each slide is associated with a slide-level label $Y \in \{1, \ldots, C\}$, while instance-level labels are unavailable. The objective is to predict the slide label by identifying and integrating informative regions within the bag.

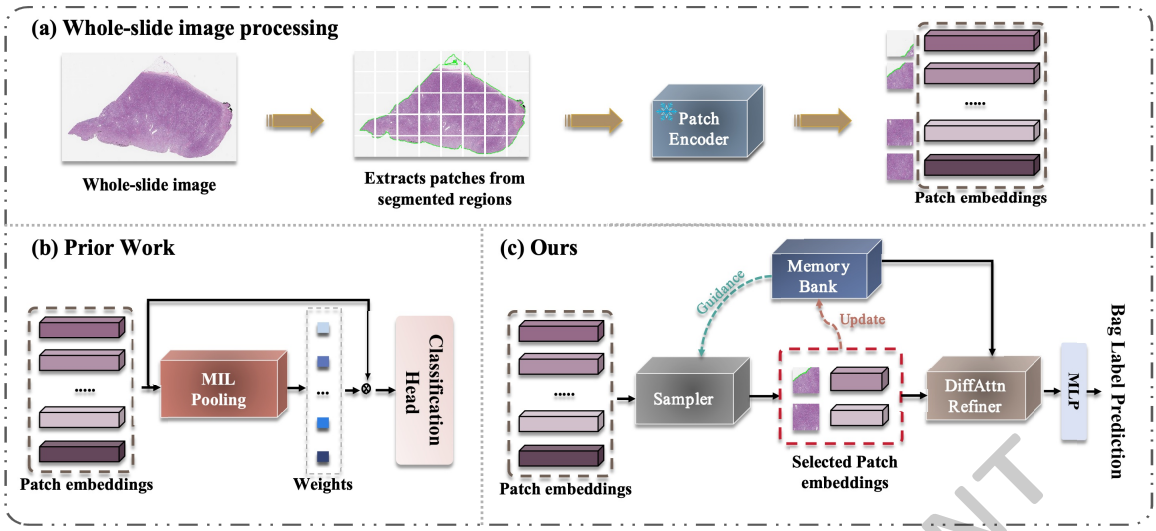

Figure 2: Conceptual comparison between conventional MIL pipelines and PMDMIL. (a) WSI preprocessing: tissue regions are segmented, tiled into patches, and encoded into patch embeddings by a frozen patch encoder. (b) Conventional MIL model directly aggregates all patch embeddings to obtain a slide representation. (c) PMDMIL introduces a memory-guided instance sampler and a differential-attention refiner operating on selected patch embeddings, producing compact and informative slide representations for classification.

3.2 Group-Wise Feature Adaptation

Directly operating on all patches of a WSI is computationally expensive and prone to noise accumulation. PMDMIL therefore follows a group-first strategy. Patch features are first adapted to a shared embedding space using a linear adapter,

$$h = \phi(x) \in \mathbb{R}^D, \tag{2}$$

where $\phi(\cdot)$ consists of normalization, linear projection, and non-linear activation. This adapter serves to harmonize features extracted from a frozen patch encoder and prepares them for subsequent aggregation.

The embedded patches are then partitioned into a set of spatial groups based on their coordinates.

$$\mathcal{G} = \{\mathcal{G}_1, \ldots, \mathcal{G}_G\}, \tag{3}$$

Grouping localizes instance selection and aggregation, ensuring that spatially dense regions do not dominate sparse but informative areas.

3.3 Memory-Guided Instance Sampler

A core component of PMDMIL is memory-guided instance sampler, which aims to retain informative patches while suppressing redundant or weakly informative regions.

Within each spatial group, the number of retained instances is determined proportionally to the group size rather than being fixed across slides. This proportional strategy naturally adapts to whole-slide images with highly variable tissue coverage and lesion burden.

To stabilize instance selection, PMDMIL maintains a class-wise prototype memory bank,

$$\mathcal{M} = \{M_c \in \mathbb{R}^{L\times D} \mid c = 1, \ldots, C\} \tag{4}$$

where L denotes the memory length per class. For efficient similarity computation, we use the mean prototype for each class $\bar{M}_c = \frac{1}{L}\sum_{s=1}^{L} M_{c,s}$.

Within each spatial group $\mathcal{G}_g$, patches are first pre-scored based on their cosine similarity to the mean class-wise prototypes. Given a patch embedding $h_i \in \mathbb{R}^D$, its memory-guided pre-score is computed as

$$s_i = \max_{c\in\{1,\ldots,C\}} \left\langle \frac{h_i}{\|h_i\|}, \frac{\bar{M}_c}{\|\bar{M}_c\|} \right\rangle, \quad i \in \mathcal{G}_g. \tag{5}$$

Based on these pre-scores, a proportion of patches is retained from each group,

$$q_g = \lceil q \cdot |\mathcal{G}_g| \rceil, \tag{6}$$

$$\mathcal{S}_g = \mathrm{Sample}\left(\{s_i\}_{i\in\mathcal{G}_g},\ q_g\right), \tag{7}$$

where $q \in (0, 1]$ is a predefined sampling ratio and $\mathcal{S}_g \subset \mathcal{G}_g$ denotes the selected instance set.

Instance-level class evidence is then computed only for the selected patches. Specifically, for each patch $j \in \mathcal{S}_g$, we compute instance-level logits as

$$\ell_j = f_{\mathrm{inst}}\left(\frac{h_j}{\|h_j\|}\right) \in \mathbb{R}^C, \tag{8}$$

where $f_{\mathrm{inst}}(\cdot)$ denotes an instance classification head. The bag-level class probability for the group is estimated using a LogSumExp [Blanchard *et al.*, 2021] aggregation over the selected instances,

$$P(c) = \mathrm{softmax}_c\left(\log \sum_{j\in\mathcal{S}_g} \exp\left(\ell_j^{(c)}\right)\right). \tag{9}$$

Based on the resulting class distribution, a group-level memory token is constructed as a weighted combination of mean class prototypes,

$$t_g = \sum_{c=1}^{C} P(c)\,\frac{\bar{M}_c}{\|\bar{M}_c\|} \in \mathbb{R}^D, \tag{10}$$

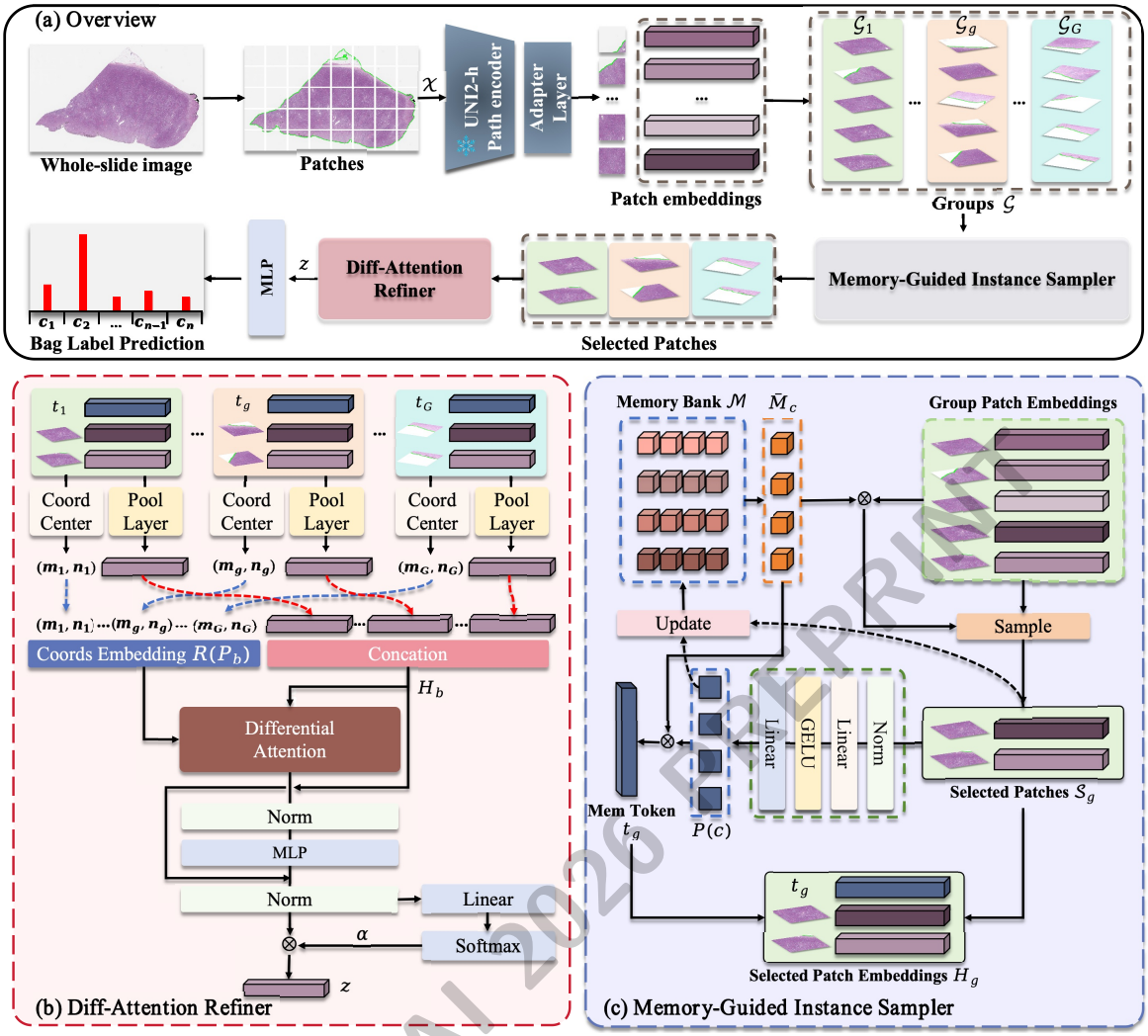

Figure 3: The architecture of our proposed PMDMIL. (a) Overall framework: given a WSI, patches are extracted and encoded into patch embeddings, which are further partitioned into spatial groups and fed into sampler and refiner to predict. (b) Memory-guided proportional instance sampler: within each group, patches are pre-selected based on similarity to class-wise memory prototypes, followed by instance-level evidence estimation and construction of a group memory token. (c) Differential-attention refiner: aggregated group tokens are refined with coordinate-based positional encoding and pooled to produce a slide-level representation for classification.

which represents the dominant semantic content of the group. The memory token is prepended to the selected instance tokens and assigned the mean spatial coordinate of the group for subsequent processing.

During training, the class-wise prototype memory is updated using an exponential moving average. Let $\bar{h}_{\mathcal{S}_g}$ denote the normalized mean feature of the selected instances. The mean prototype of the most probable class is updated as

$$\bar{h}_{\mathcal{S}_g} = \frac{1}{|\mathcal{S}_g|} \sum_{i \in \mathcal{S}_g} \frac{h_i}{\|h_i\|} \tag{11}$$

$$c = \arg\max_{c'} P(c') \tag{12}$$

$$\bar{M}_c = \text{Normalize}\left(\alpha\, \bar{M}_c + (1-\alpha)\, \bar{h}_{\mathcal{S}_g}\right) \tag{13}$$

Finally, for each spatial group, the sampler outputs a compact set of tokens composed of the group-level memory token and the selected patch tokens. Specifically, the output feature set is formed by concatenating the memory token with the selected instance features,

$$H_g = \left[t_g,\ \{h_i \mid i \in \mathcal{S}_g\}\right] \in \mathbb{R}^{(q_g+1)\times D}. \tag{14}$$

To preserve spatial information, the corresponding coordinates are constructed by assigning the mean coordinate of the group to the memory token and retaining the original coordinates of the selected patches,

$$P_g = \left[\bar{p}_g,\ \{p_i \mid i \in \mathcal{S}_g\}\right] \in \mathbb{R}^{(q_g+1)\times 2}, \tag{15}$$

where $\bar{p}_g = \frac{1}{|\mathcal{G}_g|} \sum_{i \in \mathcal{G}_g} p_i$. These group-level outputs are then passed to the subsequent aggregation and refinement modules.

3.4 Differential-Attention Refiner

After the sampler, each spatial group $\mathcal{G}_g$ outputs a compact set of tokens composed of a group-level memory token and the selected patch tokens and coordinates (H_g, P_g). To summarize local evidence within each group, PMDMIL applies a simple mean pooling layer over the sampler outputs, yielding a single group-level representation for $\mathcal{G}_g$.

The aggregated representations from all spatial groups are concatenated to form a bag-level token sequence $H_b \in \mathbb{R}^{G\times D}$. Let $P_b \in \mathbb{R}^{G\times 2}$ denote the corresponding group-level coordinates, obtained by averaging the coordinates of patches within each group.

Given the token sequence and their coordinates, 2D rotary embeddings are generated from the spatial coordinates and injected into the attention mechanism. The refiner first applies differential multi-head attention with a residual connection, and the attention output is further refined using a feed-forward network with normalization and a residual connection,

$$H^{'} = H_b + \mathrm{DiffAttn}(H_b,\, R(P_b)) \tag{16}$$

$$H^{''} = \mathrm{LN}\Big(H^{'} + \mathrm{MLP}\Big(\mathrm{LN}(H^{'})\Big)\Big)\,. \tag{17}$$

where $R(\cdot)$ denotes the rotary positional encoding derived from the 2D coordinates.

Finally, a learnable attention-based pooling operation is applied to aggregate the refined group representations into a single slide-level embedding. Specifically, The pooling weights and slide-level representation are computed as,

$$\alpha_g = \frac{\exp\Big(w^{\top} H_g^{''}\Big)}{\sum_{g'=1}^{G} \exp\Big(w^{\top} H_{g'}^{''}\Big)} \tag{18}$$

$$z = \sum_{g=1}^{G} \alpha_g\, H_g^{''} \in \mathbb{R}^{D}. \tag{19}$$

where $w \in \mathbb{R}^D$ is a learnable projection vector.

The resulting slide-level embedding z is fed into a classification head to obtain the final prediction. Specifically, z is first normalized and then passed through a two-layer multilayer perceptron with GELU activation and dropout to produce slide-level logits $\hat{y} \in \mathbb{R}^C$. The predicted class probabilities are computed using a softmax function, and the final slide label is obtained by selecting the class with the highest logit value.

3.5 Training Objective

Our model is trained primarily with cross-entropy loss at the slide level. Let $\hat{y} \in \mathbb{R}^C$ denote the slide-level logits predicted from the final slide representation z. The primary objective of PMDMIL is the slide-level cross-entropy loss,

$$\mathcal{L}_{\mathrm{bag}} = \mathrm{CE}(\mathrm{softmax}(\hat{y}),\, Y)\,. \tag{20}$$

In addition, the sampler produces instance-level logits for the selected patches within each group. We optionally use them as an auxiliary supervision signal by propagating the slide label to instances. Let $\{\ell_k\}_{k=1}^{K}$ be the instance logits (Eq. 8) returned by the sampler for a slide. The auxiliary instance loss is defined as

$$\mathcal{L}_{\mathrm{inst}} = \frac{1}{K}\sum_{k=1}^{K} \mathrm{CE}(\mathrm{softmax}(\ell_k),\, Y)\,. \tag{21}$$

The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{bag}} + \lambda\, \mathcal{L}_{\mathrm{inst}}, \tag{22}$$

where $\lambda \geq 0$ controls the contribution of the auxiliary instance-level supervision.

4 Experiments and Results

4.1 Dataset Description

We evaluate PMDMIL on five publicly available WSI classification datasets: TCGA-BRCA, BRACS [Brancati *et al.*, 2022], PANDA [Bulten *et al.*, 2022], DHMC [Zhu *et al.*, 2021] and TCGA-NSCLC. These H&E-stained cohorts span multiple organs, including breast, prostate, lung and kidney, and cover a wide range of dataset scales, with the number of WSIs per cohort varying from 261 to 10,614. Figure 4 illustrates the distribution of the number of patches per WSI across the five datasets. For fair comparison, we follow the same dataset splits as used in previous study [Zhang *et al.*, 2025b].

4.2 Implementation Details

Whole-slide images are processed using the TRIDENT framework [Zhang *et al.*, 2025a; Vaidya *et al.*, 2025]. Tissue regions are first segmented to exclude background areas, after which non-overlapping patches are extracted at $20\times$ magnification. The patch size is fixed to 512×512 pixels. This preprocessing pipeline follows standard practice in computational pathology and is consistent across all datasets. Patch-level features are extracted using a frozen foundation model (UNI2-h [Chen *et al.*, 2024]) from TRIDENT framework.

All models are optimized using AdamW with a batch size of one slide per iteration. Training is performed for 20 epochs. The initial learning rate is set to 1×10^{-4} and is scheduled using cosine annealing. All experiments are conducted following the general experimental protocol of 2DMamba [Zhang *et al.*, 2025b], including data split, optimization strategy, and training schedule, to ensure fair and reproducible comparison. Unless otherwise specified, all hyperparameters are kept consistent across datasets. All experiments are conducted using the RTX 6000 Ada GPU. Model selection is based on validation performance.

4.3 Results

Table 1 summarizes the quantitative comparison between PMDMIL and representative multiple instance learning methods on five public WSI classification benchmarks. The compared baselines include classical attention-based MIL approaches as well as recent sequence modeling methods based on transformers and Mamba architectures. Performance is evaluated using accuracy (Acc), F1-score, and AUC, with all results averaged over five runs.

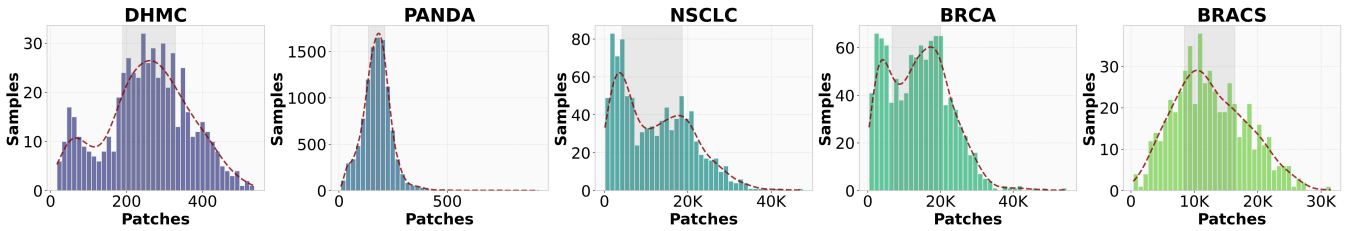

Figure 4: Histograms of patch counts per whole-slide image for five datasets. The red dashed curves denote smoothed density estimates, with μ and σ indicating the mean and standard deviation.

Methods	DHMC			PANDA			NSCLC			BRCA			BRACS		
	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC
AB-MIL	0.8684	0.7774	0.9695	0.4883	0.4269	0.7797	0.8758	0.8756	0.9572	0.9292	0.8893	0.9747	0.7057	0.6015	0.8939
DSMIL	0.8711	0.7934	0.9583	0.4633	0.3847	0.7660	0.8782	0.8780	0.9567	0.9375	0.8961	0.9770	0.6759	0.5618	0.8618
CLAM	0.8711	0.7909	0.9727	0.4802	0.4224	0.7820	0.8804	0.8803	0.9536	0.9333	0.8960	0.9753	0.7103	0.6014	**0.9016**
DTFD-MIL	0.8711	0.7704	0.9521	0.4704	0.3853	0.7665	0.8736	0.8732	0.9559	0.9271	0.8809	0.9633	0.7012	0.6131	0.8787
TransMIL	0.8067	0.7136	0.9466	0.4636	0.3970	0.7728	0.8850	0.8845	0.9626	0.9375	0.9028	0.9763	0.6919	0.6063	0.8759
S4-MIL	0.8644	0.7847	0.9284	0.5047	0.4486	0.7986	0.8851	0.8849	0.9571	0.9458	0.9154	0.9770	0.6621	0.5904	0.8457
MambaMIL	0.8550	0.7789	0.9661	0.4679	0.4216	0.7781	0.8758	0.8756	0.9582	0.9333	0.8939	0.9657	0.7379	0.6832	0.8883
SRMambaMIL	0.8590	0.7735	0.9639	0.4711	0.4209	0.7776	0.8850	0.8849	0.9592	0.9313	0.8900	0.9657	0.7379	0.6789	0.8915
2DMambaMIL	0.8926	0.8027	0.9468	0.5075	0.4562	0.8184	0.8851	0.8850	0.9618	0.9458	0.9156	**0.9782**	0.7517	0.7045	0.8964
PMDMIL	**0.8940**	**0.8572**	**0.9815**	**0.7840**	**0.7522**	**0.9454**	**0.9494**	**0.9493**	**0.9691**	**0.9479**	**0.9207**	0.9411	**0.7563**	**0.7138**	0.8587

Table 1: The comparison of accuracy (Acc), F1 and AUC on five WSI classification datasets. We conducted each experiment five times using five different random seeds and reported their mean. The highest metrics are marked as bold.

PMDMIL achieves strong and consistent performance across all datasets and outperforms most competing methods on the majority of evaluation metrics. Compared with classical MIL baselines such as AB-MIL, CLAM, and DSMIL, PMDMIL shows clear improvements, particularly on datasets with high intra-slide heterogeneity. Moreover, while recent sequence modeling approaches (e.g., TransMIL and Mamba-based methods) achieve competitive results by modeling long-range dependencies, PMDMIL further improves performance on challenging datasets such as PANDA and NSCLC, especially in terms of accuracy and F1-score, indicating that explicit instance selection remains beneficial even with strong sequence models.

On DHMC, PMDMIL achieves the best performance across all three metrics, demonstrating its ability to capture subtle morphological differences. On BRCA and BRACS, PMDMIL consistently improves F1-score, reflecting enhanced class balance in multi-class settings. Notably, these gains are obtained without processing all available patches, providing quantitative evidence that effective instance selection can be more important than exhaustive aggregation for WSI classification.

4.4 Analysis of Sampling Ratio

To analyze whether retaining more patches consistently leads to better performance, or whether a compact subset of informative patches is sufficient for accurate WSI classification, we investigate the behavior of PMDMIL under different sampling ratios q, which control the proportion of patches retained for each whole-slide image. Specifically, we vary q from 0.1 to 1.0 and evaluate the resulting classification performance, computational cost, and learned slide-level representations on the DHMC dataset.

Figure 5 summarizes the experimental results. As shown in the performance curves, the classification metrics do not exhibit a monotonic improvement as the sampling ratio increases. Instead, PMDMIL achieves strong and stable performance at moderate sampling ratios, with performance around $q = 0.1$ being comparable to or better than that obtained using significantly larger ratios. Further increasing the sampling ratio yields diminishing or inconsistent performance gains, and in some cases slightly degrades performance. These results indicate that exhaustively processing all available patches is not necessary to achieve high classification accuracy.

In contrast to the relatively stable performance trends, the computational cost measured by FLOPs increases approximately linearly with the sampling ratio. This behavior reflects the direct relationship between the number of retained instances and the computational burden of downstream processing. Consequently, moderate sampling ratios provide a substantially more favorable trade-off between accuracy and efficiency compared to full sampling, highlighting the scalability advantage of proportional instance selection in large-scale WSI analysis.

To further understand the effect of the sampling ratio on representation quality, we visualize the learned slide-level embeddings using t-SNE. As illustrated in Figure 5, representations learned with a moderate sampling ratio ($q = 0.2$) form more compact and better-separated class clusters than those obtained with a high sampling ratio ($q = 0.9$). When a large proportion of patches is retained, the embeddings become more scattered and exhibit increased overlap between classes, suggesting that excessive inclusion of weakly in-

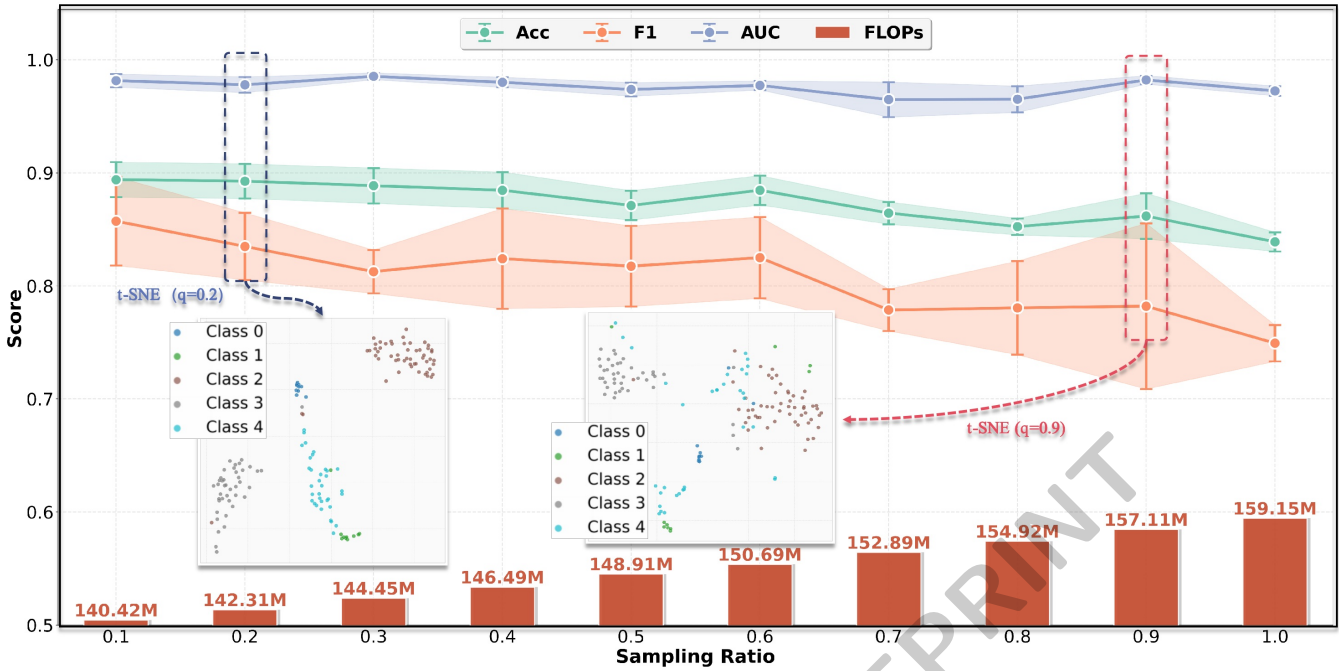

Figure 5: The classification performance on DHMC dataset obtained by PMDMIL when varying the sampling ratio from 0.1 to 1.0.

Method	Acc	F1	AUC
PMDMIL	**0.8940**	**0.8572**	**0.9815**
PMDMIL w/o Memory-guided Sampling	0.8175	0.7554	0.9713
PMDMIL w/o Instance-level Supervision	0.8295	0.7449	0.9538
PMDMIL w/o Differential Attention	0.8738	0.8490	0.9795

Table 2: Ablation study on the DHMC dataset. We evaluate the impact of memory guidance, instance-level supervision, and differential attention. Removing any component leads to consistent performance degradation, demonstrating their complementary roles.

formative patches may dilute discriminative structure in the learned representations.

This analysis demonstrates that PMDMIL benefits from a principled instance selection strategy. By retaining only a compact subset of informative patches, the model is able to preserve discriminative representations while significantly reducing computational cost. This analysis provides empirical support for the central "less is more" principle underlying PMDMIL, showing that effective instance selection is more important than exhaustive aggregation for whole-slide image classification.

4.5 Ablation Study

We conduct ablation experiments on the DHMC dataset to analyze the contribution of key components in PMDMIL, with results summarized in Table 2. Removing memory-guided sampling leads to a substantial performance drop across all metrics, indicating that the class-wise prototype memory provides an important global reference for identifying informative patches beyond proportional sampling alone. Eliminating the auxiliary instance-level supervision also consistently degrades performance, suggesting that instance-level scores serve as a useful regularization signal under slide-level supervision. Replacing differential attention with standard attention results in a smaller but noticeable performance decrease, highlighting the benefit of explicitly modeling relative differences between group representations. These findings indicate that memory guidance, instance-level supervision, and differential-attention refinement play complementary roles in improving the robustness and effectiveness of PMDMIL.

5 Conclusion

In this work, we proposed PMDMIL, a proportional memory-guided multiple instance learning framework for whole-slide image classification. Motivated by the sparsity and spatial structure of diagnostic evidence in WSIs, PMDMIL performs group-wise proportional instance selection to reduce redundant computation while preserving spatial organization, following a principled "less is more" paradigm. Experiments on five public benchmarks demonstrate that PMDMIL consistently outperforms representative MIL baselines while processing only a fraction of patches, and ablation analyses confirm the complementary roles of memory-guided sampling and differential-attention refinement. In conclusion, PMDMIL provides an effective and scalable solution for WSI classification.

Acknowledgments

This work is supported by grants from the National Natural Science Foundation of China (Grants No. 62322215, 62532017 and 62402488), and Natural Science Foundation of Hunan Province (Grants No. 2026JJ30018).

References

[Adnan *et al.*, 2022] Mohammed Adnan, Shivam Kalra, Jesse C Cresswell, Graham W Taylor, and Hamid R Tizhoosh. Federated learning and differential privacy for medical image analysis. *Scientific reports*, 12(1):1953, 2022.

[Blanchard *et al.*, 2021] Pierre Blanchard, Desmond J Higham, and Nicholas J Higham. Accurately computing the log-sum-exp and softmax functions. *IMA Journal of Numerical Analysis*, 41(4):2311–2330, 2021.

[Brancati *et al.*, 2022] Nadia Brancati, Anna Maria Anniciello, Pushpak Pati, Daniel Riccio, Giosuè Scognamiglio, Guillaume Jaume, Giuseppe De Pietro, Maurizio Di Bonito, Antonio Foncubierta, Gerardo Botti, et al. Bracs: A dataset for breast carcinoma subtyping in h&e histology images. *Database*, 2022:baac093, 2022.

[Bulten *et al.*, 2022] Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen, Peter Ström, Hans Pinckaers, Kunal Nagpal, Yuannan Cai, David F Steiner, Hester Van Boven, Robert Vink, et al. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. *Nature medicine*, 28(1):154–163, 2022.

[Campanella *et al.*, 2019] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019.

[Caron *et al.*, 2020] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

[Chen *et al.*, 2022] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16144–16155, 2022.

[Chen *et al.*, 2024] Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Bowen Chen, Andrew Zhang, Daniel Shao, Andrew H Song, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 2024.

[Coudray *et al.*, 2018] Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L Moreira, Narges Razavian, and Aristotelis Tsirigos. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature medicine*, 24(10):1559–1567, 2018.

[Di *et al.*, 2022] Donglin Di, Changqing Zou, Yifan Feng, Haiyan Zhou, Rongrong Ji, Qionghai Dai, and Yue Gao. Generating hypergraph-based high-order representations of whole-slide histopathological images for survival prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5800–5815, 2022.

[Fang *et al.*, 2024] Heng Fang, Sheng Huang, Wenhao Tang, Luwen Huangfu, and Bo Liu. Sam-mil: A spatial contextual aware multiple instance learning approach for whole slide image classification. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 6083–6092, 2024.

[Fang *et al.*, 2025] Jiaxiang Fang, Shiqiang Ma, Shengfeng He, and Fei Guo. Self-support prototype-aware for few-shot semantic segmentation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[Hammoud and Ghanem, 2025] Hasan Abed Al Kader Hammoud and Bernard Ghanem. Diffclip: Differential attention meets clip. *arXiv preprint arXiv:2503.06626*, 2025.

[He *et al.*, 2012] Lei He, L Rodney Long, Sameer Antani, and George R Thoma. Histology image analysis for carcinoma detection and grading. *Computer methods and programs in biomedicine*, 107(3):538–556, 2012.

[He *et al.*, 2026] Yifeng He, Jinbo Xie, Suquan Zhong, Changxin Zhan, Fazhong Dai, Hongshen Lai, Mancun Wang, Yanyan He, Harsh Patel, Zhe-Sheng Chen, et al. A deep learning-generated mixed tumor–stroma ratio for prognostic stratification and multi-omics profiling in bladder cancer. *Research*, 9:1053, 2026.

[Ho *et al.*, 2023] Stella Ho, Ming Liu, Lan Du, Longxiang Gao, and Yong Xiang. Prototype-guided memory replay for continual learning. *IEEE transactions on neural networks and learning systems*, 35(8):10973–10983, 2023.

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

[Li *et al.*, 2021] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2021.

[Lu *et al.*, 2021] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021.

[Snell *et al.*, 2017] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

[Song *et al.*, 2024] Andrew H Song, Richard J Chen, Tong Ding, Drew FK Williamson, Guillaume Jaume, and Faisal Mahmood. Morphological prototyping for unsupervised

slide representation learning in computational pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11566–11578, 2024.

[Srinidhi *et al.*, 2021] Chetan L Srinidhi, Ozan Ciga, and Anne L Martel. Deep neural network models for computational histopathology: A survey. *Medical image analysis*, 67:101813, 2021.

[Vaidya *et al.*, 2025] Anurag Vaidya, Andrew Zhang, Guillaume Jaume, Andrew H Song, Tong Ding, Sophia J Wagner, Ming Y Lu, Paul Doucet, Harry Robertson, Cristina Almagro-Perez, et al. Molecular-driven foundation model for oncologic pathology. *arXiv preprint arXiv:2501.16652*, 2025.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[Wang *et al.*, 2025] Xiangxue Wang, Yufei Zhou, Cristian Barrera, Yijiang Chen, Bolin Song, Cheng Lu, Kailin Yang, Shlomo Koyfman, James Lewis, Kurt Scahlper, et al. A deep learning framework to identify prognostically relevant cancer regions (dover) within whole slide histopathology images. *Cancer letters*, page 218134, 2025.

[Wei *et al.*, 2025] Hongkai Wei, Yang Yang, Shijie Sun, Mingtao Feng, Rong Wang, and Xianfeng Han. Lmttm-vmi: Linked memory token turing machine for 3d volumetric medical image classification. *Computer Methods and Programs in Biomedicine*, 262:108640, 2025.

[Wu *et al.*, 2018] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018.

[Xiang and Zhang, 2023] Jinxi Xiang and Jun Zhang. Exploring low-rank property in multiple instance learning for whole slide image classification. In *The Eleventh International Conference on Learning Representations*, 2023.

[Yao *et al.*, 2020] Jiawen Yao, Xinliang Zhu, Jitendra Jonnagaddala, Nicholas Hawkins, and Junzhou Huang. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical image analysis*, 65:101789, 2020.

[Ye *et al.*, 2024] Tianzhu Ye, Li Dong, Yuqing Xia, Yutao Sun, Yi Zhu, Gao Huang, and Furu Wei. Differential transformer. *arXiv preprint arXiv:2410.05258*, 2024.

[Yu *et al.*, 2023] Jin-Gang Yu, Zihao Wu, Yu Ming, Shule Deng, Yuanqing Li, Caifeng Ou, Chunjiang He, Baiye Wang, Pusheng Zhang, and Yu Wang. Prototypical multiple instance learning for predicting lymph node metastasis of breast cancer from whole-slide pathological images. *Medical Image Analysis*, 85:102748, 2023.

[Yu *et al.*, 2025] Jikai Yu, Hongda Chen, Lianxin Hu, Boyuan Wu, Shicheng Zhou, Jiayun Zhu, Yizhen Jiang, Shuwen Han, and Zefeng Wang. Exploring multi-instance learning in whole slide imaging: Current and future perspectives. *Pathology-Research and Practice*, page 156006, 2025.

[Zhang *et al.*, 2024] Yunlong Zhang, Honglin Li, Yunxuan Sun, Sunyi Zheng, Chenglu Zhu, and Lin Yang. Attention-challenging multiple instance learning for whole slide image classification. In *European conference on computer vision*, pages 125–143. Springer, 2024.

[Zhang *et al.*, 2025a] Andrew Zhang, Guillaume Jaume, Anurag Vaidya, Tong Ding, and Faisal Mahmood. Accelerating data processing and benchmarking of ai models for pathology. *arXiv preprint arXiv:2502.06750*, 2025.

[Zhang *et al.*, 2025b] Jingwei Zhang, Anh Tien Nguyen, Xi Han, Vincent Quoc-Huy Trinh, Hong Qin, Dimitris Samaras, and Mahdi S Hosseini. 2dmamba: Efficient state space model for image representation with applications on giga-pixel whole slide image classification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3583–3592, 2025.

[Zheng *et al.*, 2023] Tingting Zheng, Weixing Chen, Shuqin Li, Hao Quan, Mingchen Zou, Song Zheng, Yue Zhao, Xinghua Gao, and Xiaoyu Cui. Learning how to detect: A deep reinforcement learning method for whole-slide melanoma histopathology images. *Computerized Medical Imaging and Graphics*, 108:102275, 2023.

[Zheng *et al.*, 2024] Tingting Zheng, Kui Jiang, and Hongxun Yao. 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*, pages 8028–8037, 2024.

[Zheng *et al.*, 2025] Tingting Zheng, Hongxun Yao, Kui Jiang, Yi Xiao, and Sicheng Zhao. Gmmamba: Group masking mamba for whole slide image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9935–9944, 2025.

[Zhou *et al.*, 2023] Tianfei Zhou, Liulei Li, Gustav Bredell, Jianwu Li, Jan Unkelbach, and Ender Konukoglu. Volumetric memory network for interactive medical image segmentation. *Medical Image Analysis*, 83:102599, 2023.

[Zhu *et al.*, 2021] Mengdan Zhu, Bing Ren, Ryland Richards, Matthew Suriawinata, Naofumi Tomita, and Saeed Hassanpour. Development and evaluation of a deep neural network for histologic classification of renal cell carcinoma on biopsy and surgical resection slides. *Scientific reports*, 11(1):7080, 2021.