Dynamic Policy-Driven Adaptive Multi-Instance Learning for Whole Slide Image Classification

Supplementary Material

Tingting Zheng[1] Kui Jiang[1] Hongxun Yao[1]
[1] Harbin Institute of Technology
23b903051@stu.hit.edu.cn, {jiangkui, h.yao}@hit.edu.cn

In the supplementary materials, we detail **the methods and dataset descriptions, while providing more experimental results and comprehensive discussions**. In Section A, we detail both the Dynamic Policy Instance Selection Scheme (DPIS) and the optimization method used for Selection Fusion Feature Representation (SFFR). Section B elaborates on dataset description and implementation details. In Section C, extensive experiments and in-depth analyses are conducted to verify the contributions of individual components, including the impact of the reward-punishment system on final performance, comparing pseudo-bags level methods with our DPIS scheme, and additional visualization. Finally, Section D investigates thoroughly the performance improvement achieved by our DPIS scheme, highlighting its advantages and potential areas for improvement.

A. More Details for Method

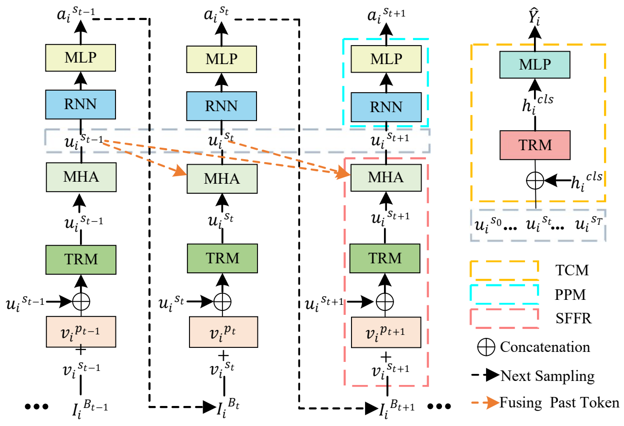


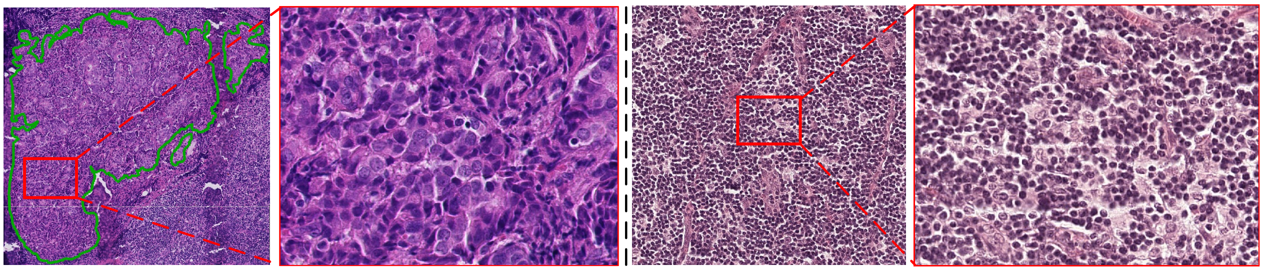
Figure S1. Illustration of the proposed **selection fusion feature representation (SFFR) module**, **proximal policy module** (PPM) and **Transformer classification module (TCM)** in dynamic policy-driven adaptive multi-instance learning (PAMIL) framework. We sample features $v_i^{s_t}$ in the remaining feature space $I_i^{B_t}$ according to the instances indexes $a_i^{s_{t-1}}$. SFFR then takes $v_i^{s_t}$ and an initial token $u_i^{s_t}$ as inputs, refining $u_i^{s_t}$ by utilizing a Transformer module (TRM) and a multi-head attention (MHA) mechanism to fuse $v_i^{s_t}$ and past tokens. Meanwhile, we introduce a Siamese (SIA) structure between $u_i^{s_t}$ and $u_i^{s_{t-1}}$ to enhance the robustness of $u_i^{s_t}$ (as described in Section A.2). Finally, TCM uses a class token (CLS) h_i^{cls} to aggregate tokens for inferring the probability $\hat{Y}_i$ of i^{th} WSI.

A.1. Dynamic Policy Instance Selection Scheme

As described in **Section 3.3** of the main paper, we integrate the dynamic instance sampling and experience-based learning of proximal policy module (PPM) G_p into a unified framework to facilitate decision-making, forming a novel dynamic policy instance selection (DPIS) scheme. The DPIS aims to select the informative samples from the remaining instances, guided by previous knowledge and decision-making feedback. As shown in Figure S1, the corrective past information and the current features token $u_i^{s_t}$ are packed into a recurrent neural network (RNN) G_p^{RNN} [13] to fully explore the temporal dependencies, followed by a multi-layer perceptron (MLP) G_p^{MLP} to derive the relation index for the next sampling. This process is expressed as

$$P_i^t(a_i^{s_t}|u_i^{s_t}) = G_p^{\text{MLP}}(G_p^{\text{RNN}}(u_i^{s_t})), \tag{1}$$

Where $a_i^{s_t}$ represents the indexes of the next sampling. Considering the application convenience, and the connectivity and proximity between the current and the remaining instances, we advise three different schemes to optimise instance sampling. They are greedy policy-based max similarity scheme (GMSS), greedy policy-based hybrid similarity scheme (GHSS) and policy-optimized linear interpolation instances scheme (LIIS).



(a) Tumor WSI

(b) Normal WSI

Figure S2. **Illustration of whole slide images (WSIs)**. The green annotation indicates the tumor regions. The red bounding boxes in the images indicate that the cropped instances within these boxes share spatial and contextual consistency.

Greedy Policy-based Max Similarity Scheme (GMSS). Since the similarity between the current instances and the remaining instances is taken into consideration, we introduce the max-feature similarity with a ξ-greedy policy [12]. Specifically, when the probability $P_i^t \in \mathbb{R}^{1\times 1}$ is above the given threshold ξ, we choose M indexes closest to $u_i^{s_t}$ in remaining features (more M details and settings are in the Sections 3.2 and 4.6 of the main paper). Otherwise, M instance indexes are randomly selected for the next step. This scheme balances feature relevance and historical data exploration, efficiently capturing robustness features without over-reliance on any single instance. However, this method relies heavily on the accuracy of pseudo-bag labels, which may lead to false positives or false negatives.

Greedy Policy-based Hybrid Similarity Scheme (GHSS). To mitigate the risk of GMSS, we employ a hybrid similarity-greedy approach. Unlike GMSS, GHSS selects the top $\frac{M}{2}$ indexes nearest to $u_i^{s_t}$ and another $\frac{M}{2}$ indexes away, at a distance of $2 \times M$. It greatly alleviates the overfitting. Additionally, to improve efficiency, instances selected at time t are masked in subsequent selections. ξ is initially set to 0 and increases linearly to 0.9 along with iterations, shifting DPIS from broad exploration to targeted selection as knowledge accumulates.

Policy-optimized Linear Interpolation Instances Scheme (LIIS). Existing works [19, 20] have demonstrated that applying reinforcement learning (RL) directly can produce more discriminative features or instances to facilitate classification performance. However, it is non-trivial to identify salient instances from many instances for the next sampling. In addition, Figure S2 displays that the adjacent instances share spatial and context consistency. Therefore, we advise a policy-optimized linear interpolation instances scheme (LIIS) for sampling neighboring instances in the remaining space. Specifically, PPM aggregates past experience and current feature into $u_i^{s_t}$ to estimate potential sampling regions likely to contain crucial instances, expressed as probabilities $P_i^t \in \mathbb{R}^{1\times M}$. We then employ linear interpolation between instances of the remaining space according to the indices (locations) of current instances and probabilities $P_i^t \in \mathbb{R}^{1\times M}$, and select the top M probability indices for the next step. LIIS eliminates similarity calculations and is more flexible in adapting to various bag sizes.

A.2. Selection Fusion Feature Representation

As illustrated in **Section 3.4** of the main manuscript, to achieve general and robust token $u_i^{s_t}$, the selection fusion feature representation (SFFR) module is equipped with a more stable and efficient Siamese (SIA) structure [4] among $u_i^{s_{t-1}}$ and $u_i^{s_t}$. As shown in Figure S3, the Simsiam (SIA) G_{SIA} takes the current token $u_i^{s_t}$ and past token $u_i^{s_{t-1}}$ as inputs. These tokens are processed through a multi-layer perceptron (MLP) $G_{\text{SIA}}^{\text{MLP}}$ to transform $u_i^{s_t}$ into $p_i^{s_t}$, defined as $p_i^{s_t} = G_{\text{SIA}}^{\text{MLP}}(u_i^{s_t})$. The SIA contrastive loss [4] aims to minimize the negative cosine similarity between $p_i^{s_t}$ and $u_i^{s_{t-1}}$ for enriching the informational content of $u_i^{s_t}$. This process is expressed as

$$\mathcal{D}(p_i^{s_t}, u_i^{s_{t-1}}) = -\frac{p_i^{s_t}}{\|p_i^{s_t}\|_2} \cdot \frac{u_i^{s_{t-1}}}{\|u_i^{s_{t-1}}\|_2}, \tag{2}$$

where $\|\cdot\|_2$ is L_2-norm. The core component of SIA is to stop-gradient (Stop-grad) operation to avoid model collapse. Eq. 2 is modified as

$$\mathcal{D}(p_i^{s_t}, u_i^{s_{t-1}}) = D(p_i^{s_t}, \text{Stop-grad}(u_i^{s_{t-1}})). \tag{3}$$

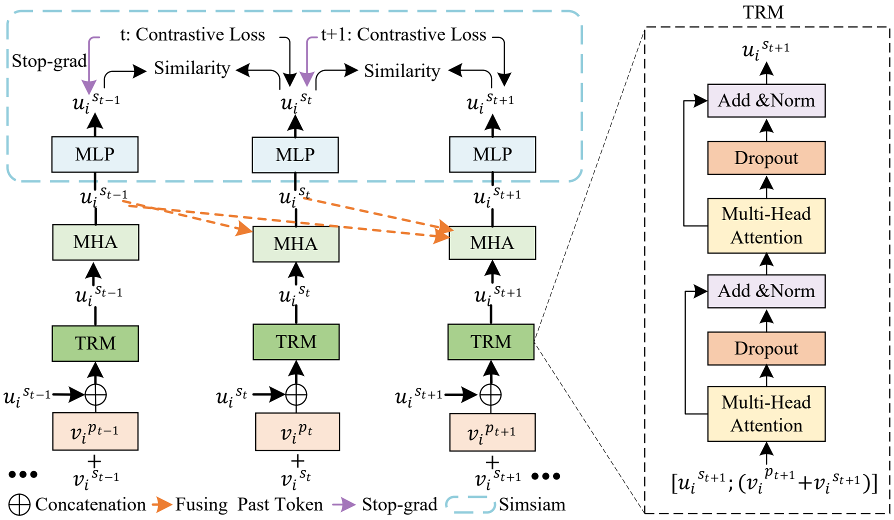


Figure S3. **Simsiam architecture in our proposed selection fusion feature representation (SFFR) module.** SFFR consists of a Transformer module (TRM), a multi-head attention (MHA) mechanism, and a Simsiam (SIA) architecture. SFFR takes a vector $v_i^{p_t}$ to learn spatial information in $v_i^{s_t}$. The TRM and MHA are designed to employ an initial token $u_i^{s_t}$, aiming to capture local, global, and historical representations through the interactions between the current $v_i^{s_t}$ and past tokens $\{u_i^{s_k}\}_{k=1}^{t-1}$. Finally, we introduce SIA contrastive loss among $u_i^{s_t}$ and $u_i^{s_{t-1}}$ from a multi-layer perceptron (MLP) for more robust and discriminative feature $u_i^{s_t}$ to facilitate the decision-making.

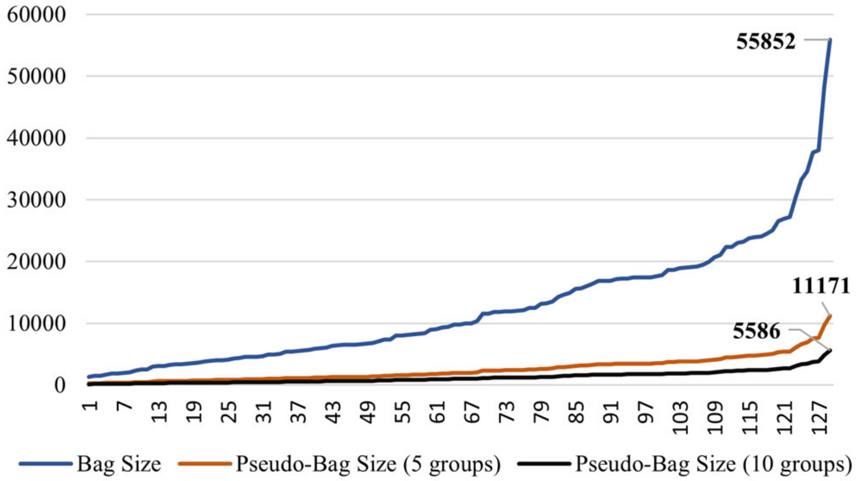


Figure S4. **Illustration of different grouping schemes on the CAMELYON16 test set.** The horizontal axis represents WSI index. The vertical axis denotes the number of bag or pseudo-bag instances per WSI using different grouping schemes. The 5 or 10 represent by a pseudo-packet scheme to split each bag into 5 or 10 groups.

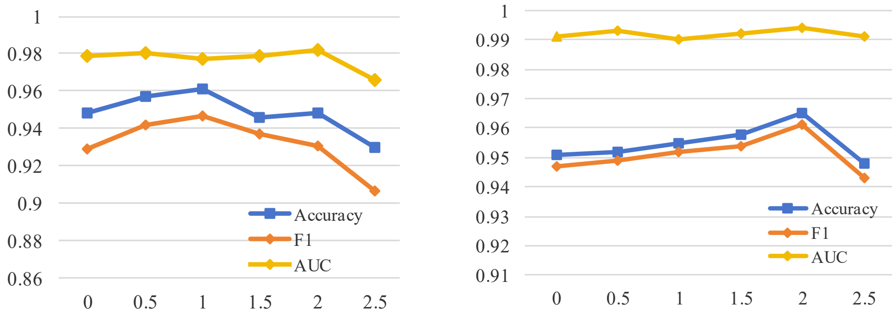
The final SIA contrast loss is defined as

$$\mathcal{L}_{\text{SIA}}^{i} = \frac{1}{T}\sum_{t=1}^{T}\left[\frac{1}{2}\mathcal{D}(p_i^{s_t}, u_i^{s_{t-1}}) + \frac{1}{2}\mathcal{D}(p_i^{s_{t-1}}, u_i^{s_t})\right]. \tag{4}$$

B. More Dataset Description and Implementation Details

B.1. Dataset Description

The CAMELYON16 dataset [2] consists of 270 training whole slide images (WSIs) (159 normal and 111 tumor) and 129 testing WSIs (80 normal and 49 tumor). Following TransMIL [14], we use CLAM [10] to identify tissues on WSIs and



(a) CAMELYON16

(b) TCGA Lung

Figure S5. **Accuracy, F1 and AUC scores with different reward values on the CAMELYON16 and TCGA Lung datasets**. The horizontal axis indicates the different reward r_i^* values.

obtain non-overlapping 256×256 instances at $20\times$ magnification. The TCGA Lung cancer dataset [16] contains two cancer subtypes: Lung Adenocarcinoma (LUAD) and Lung Squamous Cell Carcinoma (LUSC). It provides 541 LUAD slides from 478 cases and 512 LUSC slides from 478 distinct cases. We adopt the same pre-processing as DSMIL [8] for 1046 WSIs, segmenting each WSI into non-overlapping patches of 224×224 at $20\times$ magnification.

After pre-processing, the TCGA Lung dataset comprises approximately 4.1 million instances, averaging $4,000$ instances per bag. The CAMELYON16 dataset yields about 4.6 million instances, with an average of roughly $13,600$ instances per bag. As shown in Figure S4, the test set shows a range of bag sizes, with maximum and minimum extents of $55,852$ and $1,347$, respectively. Notably, positive instances in the tumor WSIs are scarce, presenting a significant challenge to traditional bag and pseudo-bags based multi-instance learning (MIL) methods.

B.2. Implementation Details

We adopt the same settings as [14, 15, 17, 18] for a fair comparison due to discrepancies of dataset splits. The CAMELYON16 official training set is further randomly divided into training and validation sets at $9:1$. Following [10, 14, 15, 17], we modify a pre-trained ResNet50 (without its last convolutional module, as trained on ImageNet [5]) to extract 1024-dimensional feature vectors from each instance using global average pooling. For the TCGA Lung dataset, the dataset is randomly split into training, validation, and testing sets with ratios of $65:10:25$. We employ the SimCLR [11] with a ResNet18 [6] encoder to obtain 512-dimension feature vectors from each patch.

AdaMax optimizer [1] with a weight decay of $1e-5$ and the initial learning rate of $1e-4$ are used. To enhance model generalization and label relevance, we employ a cosine scheduler [9] for dynamic adjustment of the weights λ_{STL} and λ_{SIA} in the Eq.(9) (detailed in main paper). During the initial phase of training, a high value (e.g., 0.5) is assigned to λ_{SIA} encouraging robust feature exploration by the SFFR module. As the training progresses, λ_{SIA} is gradually reduced while λ_{STL} is increased, guiding the model from broad feature exploration to more precise feature refinement. To prevent potential bias, a maximum limit of 0.1 is set for λ_{STL}. The loss $\mathcal{L}_{\text{WSL}}^{i}$ function provides consistent and stable supervisory information throughout the training, ensuring that predictions align with the actual WSI labels. With the above settings, we train our PAMIL with 300 epochs with batch size 1 on one NVIDIA 2080Ti GPU.

C. More Experiments

C.1. Effects of Reward

We construct an in-depth exploration of how rewards affect instance sampling robustness and labeling relevance in the DPIS scheme. As shown in Table S5, compared to using only the penalty term, appropriately increasing the reward significantly enhances model performance across all metrics and datasets. For positive WSIs in the CAMELYON16 dataset, which occupy small tumor portions [8, 14], the challenge lies in prompting the model to focus on positive instances without overfitting. Therefore, a slightly smaller reward guides attention from negative to tumor instances. In contrast, since the TCGA lung dataset comprises over 80% tumor areas [14], a larger reward ensures the model selects the most informative instances for achieving the best performance. However, a larger reward excessively emphasizes the role of labels, which limits the generalizability of the model, resulting in higher false positive or false negative rates. To this end, we set the reward $r_i^* = 1$ for the CAMELYON16 dataset and $r_i^* = 2$ for the TCGA Lung dataset.

Table S1. **Comparison of the pseudo-bags-level method DTFD and the bag-level method ABMIL on the CAMELYON16 dataset.** The numbers in **bold** indicate the best performance. Δ denotes the performance improvement by using Eq. 5 to predict the WSI probability compared to the prediction using only h_i^{cls} in Eq. 6. "TPB@AvgTop1", "TPB@AvgTop3" and "TPB@AvgTop5" indicate using $\hat{y}_{i,\max}$, $\text{avg}(\hat{y}_{i,1:3})$ and $\text{avg}(\hat{y}_{i,1:5})$ to predict tumor WSI label probability, and otherwise by $\hat{Y}_i$. Relative improvement ratio: $\Delta = \frac{(\text{Eq } 5 - \text{Eq } 6) \times 100\%}{1 - \text{Eq } 6}$

Metrics	Methods	Eq. 6	TPB@AvgTop1	TPB@AvgTop3	TPB@AvgTop5	Eq. 5	Δ (%) ↑
Accuracy	ABMIL [7] (Baseline)	0.845	—	—	—	—	—
	DTFD-AFS [17]	0.798	0.845	0.837	0.807	0.822	11.7
	DTFD-MaxMinS	0.822	0.884	0.837	0.791	0.833	6.51
	DTFD-MaxS	0.845	0.899	0.853	0.791	0.847	1.25
	DPIS-GMSS	0.915	**0.985**	**0.977**	**0.977**	**0.963**	**57.1**
	DPIS-GHSS	0.923	0.954	0.946	0.946	0.942	25.0
	DPIS-LIIS	**0.954**	0.969	0.961	0.961	0.961	16.6
F1	ABMIL (Baseline)	0.779	—	—	—	—	—
	DTFD-AFS	0.759	0.825	0.814	0.748	0.786	11.2
	DTFD-MaxMinS	0.758	0.854	0.784	0.703	0.775	6.95
	DTFD-MaxS	0.756	0.854	0.771	0.640	0.755	-0.334
	DPIS-GMSS	0.876	**0.980**	**0.969**	**0.969**	**0.949**	**58.4**
	DPIS-GHSS	0.894	0.939	0.928	0.928	0.922	26.7
	DPIS-LIIS	**0.936**	0.958	0.947	0.947	0.947	17.4
AUC	ABMIL (Baseline)	0.839	—	—	—	—	—
	DTFD-AFS	0.896	0.896	0.832	0.807	0.858	-36.3
	DTFD-MaxMinS	0.858	0.951	0.917	0.896	0.906	33.5
	DTFD-MaxS	0.865	0.983	0.977	0.969	0.948	61.8
	DPIS-GMSS	0.905	**0.996**	**0.993**	**0.992**	0.972	**70.0**
	DPIS-GHSS	**0.944**	0.990	0.976	0.970	0.970	46.0
	DPIS-LIIS	0.944	0.992	0.987	0.987	**0.977**	59.6

C.2. Comparison with Pseudo-bags-level Method

To validate the effectiveness of our proposed PAMIL method in sampling discriminative instances and fusing historical information for enhanced decision-making representation, we conduct comparison between PAMIL and the pseudo-bags-level DTFD [17] method by using Eq. 5 and 6 for predicting WSI label final probabilities on the CAMELYON16 dataset. The DTFD methods are built from officially released code. We employ bag-level attention-based multi-instance learning (ABMIL) [7] method as the baseline. Each model is trained for 300 epoch with base size 1 on a single NVIDIA 2080Ti. In the testing phase, the final decision-making $\hat{Y}_i'$ for X_i is depicted as

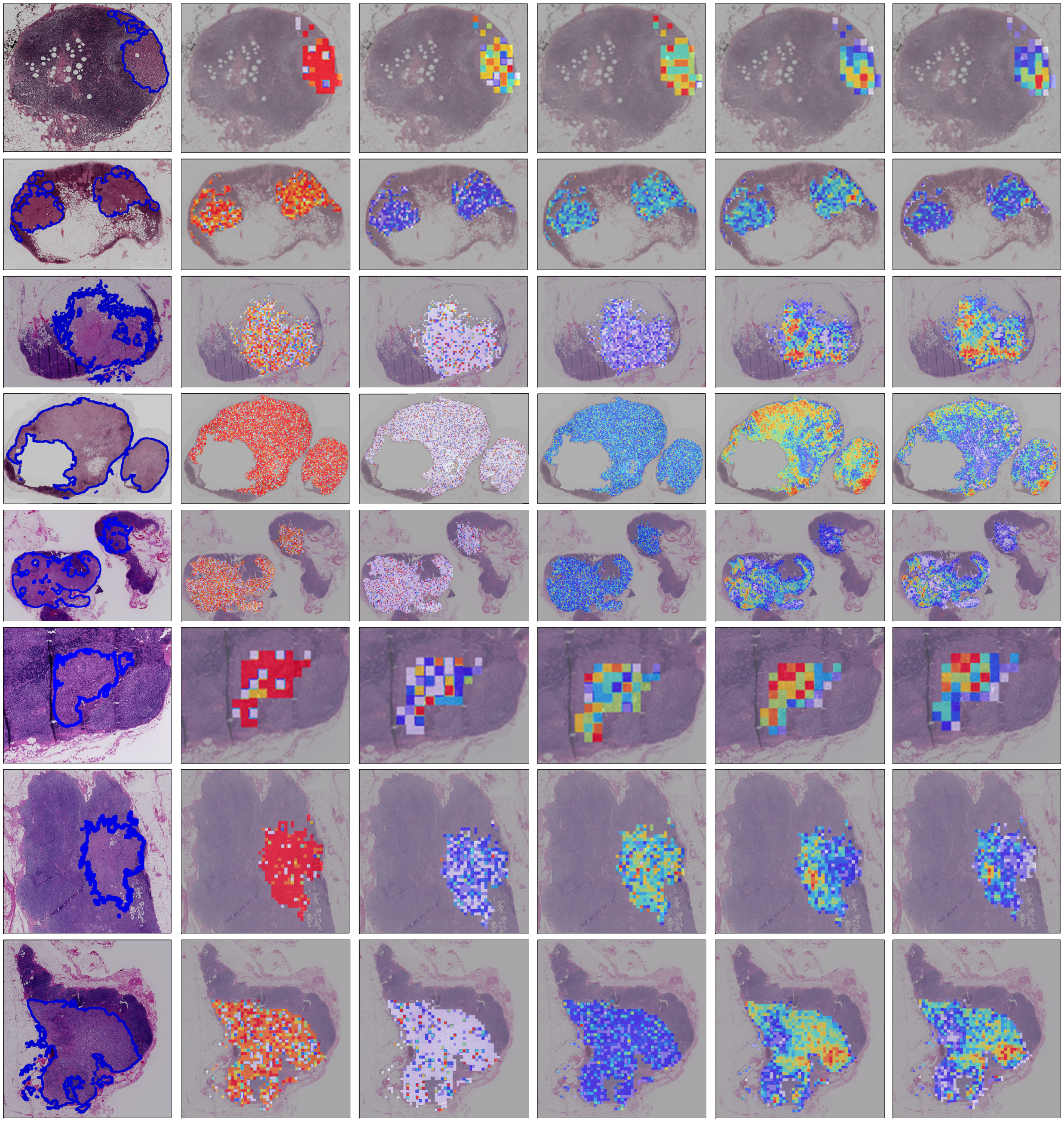
$$\hat{Y}_i' = \begin{cases} \frac{\hat{y}_{i,\max} + \text{avg}(\hat{y}_{i,1:3}) + \text{avg}(\hat{y}_{i,1:5}) + \hat{Y}_i}{4}, & \text{if } Y_i = 1, \\ \hat{Y}_i, & \text{if } Y_i = 0. \end{cases} \tag{5}$$

$$\hat{Y}_i' = \hat{Y}_i = G_c{}^{\text{MLP}}(h_i^{cls}), \tag{6}$$

Where $G_c{}^{\text{MLP}}$ is a MLP in the Transformer classification module (TCM). $\{\hat{y}_{i,t}\}_{t=1}^{T}$ are drawn from a MLP G_s^{MLP} to infer the category of $\{u_i^{s_t}\}_{t=1}^{T}$. $\hat{y}_{i,\max}$, $\text{avg}(\hat{y}_{i,1:3})$ and $\text{avg}(\hat{y}_{i,1:5})$ denote the top-1 max, and averages of the top 3 and 5 in $\{\hat{y}_{i,t}\}_{t=1}^{T}$, respectively. $\hat{Y}_i$ indicates that class token (CLS) h_i^{cls} is employed to predict WSI label.

Results in the terms of the Area Under Curve (AUC), accuracy, F1 score (F1) and relative improvement ratio Δ are presented in Table S1, where a higher Δ indicates the presence of label-related and robust pseudo-bag representations in $\{u_i^{s_t}\}_{t=1}^{T}$ for accurate bag h_i^{cls} prediction. As shown in Table S1, our proposed PAMIL method significantly outperforms all competitors (Eq. 6). In particular, considering pseudo-bags predictions (Eq. 5), we observe a remarkable enhancement in classification accuracy compared to DFTD. This improvement is attributed to the DPIS scheme, which samples the most informative instances from the sampling bag. However, previous MIL methods are barely explore mutual relationship between sampling, feature representation and decision-making. Although DTFD-MaxS gains a slight advantage over DPIS-GHSS and DPIS-LIIS in terms of AUC improvement rate, it suffers from lower accuracy than DPIS-GHSS and DPIS-LIIS by 9.5% and 11.4%, respectively (referring to the results of DPIS-GHSS, DPIS-LIIS, and DTFD-MaxS in Eq. 5). In addition, since ABMIL is poor in identifying salient information from thousands of instances, it fails to surpass the pseudo-bags-level approach on the AUC. These results demonstrate that our strategy effectively balances specific feature representations, alleviating the shortcomings of both bag-level and pseudo-bags-level methods.

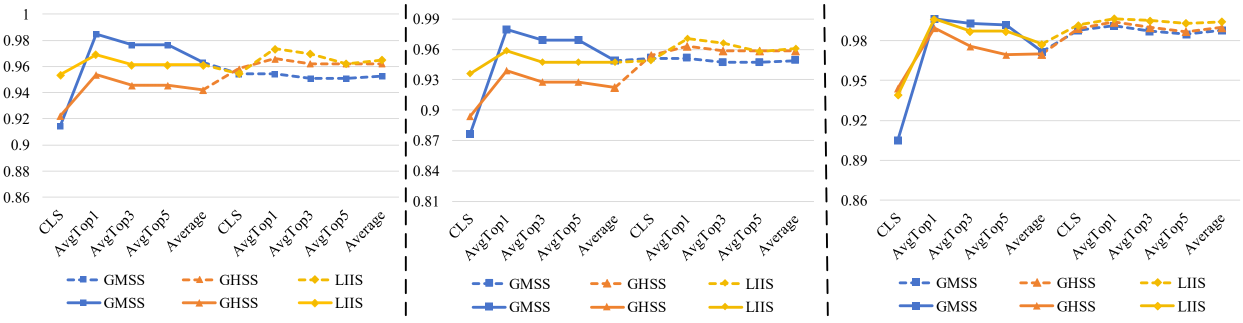
C.3. Additional Visualization



(a) Input Tumor Region (b) PAMIL (Ours) (c) K-Means Grouping (d) Random Grouping (e) DTFD (f) ABMIL

Figure S6. **Visual comparisons with pseudo-bags schemes, pseudo-bags-level method (DTFD) and bags-level method (ABMIL) on CAMELYON16 test set**. Our PAMIL presents more attention to tumor instances. The pseudo-bags schemes (c) and (d) are K-Means grouping and random grouping split each bag of 10 groups for training TRM in SFFR and TCM under the cross-entropy loss at the level of WSI (detailed in the main manuscript Eq.(7)). The blue outline indicates the tumor region. Attention score (0-1): indicates model focus on tumor instances, higher values show more attention, not positive probability.

To further demonstrate robustness and effectiveness, besides the aforementioned quantitative comparisons, a visual comparison of the pseudo-bag schemes, the pseudo-bags-level method DTFD, and the bag-level method ABMIL is presented in Figure S6. Our method, K-Means, and random grouping visualize attention scores from the final Transform layer in the



(a) Accuracy (b) F1 (c) AUC

Figure S7. **Different sampling schemes performance on CAMELYON16 and TCGA Lung datasets**. The solid and dot lines represent the CAMELYON16 dataset and TCGA Lung dataset, respectively. "CLS", "AvgTop1", "AvgTop3" and "AvgTop5" indicate using h_i^{cls}, $\hat{y}_{i,\max}$, avg$(\hat{y}_{i,1:3})$ and avg$(\hat{y}_{i,1:5})$ to predict tumor WSI label probability. Average is their average result.

TCM. DTFD and ABMIL generate heat maps based on normalized attention scores. As we can see, our method effectively focuses on tumor instances compared to all competitors, providing more convincing evidence for the high accuracy shown in Table S1. Specifically, PAMIL guides the model toward sampling salient instances by combining relationships between historical information and feedback correction mechanisms for more precise predictions. Although the attention mechanism demonstrates impressive performance in extracting key information, capturing label-relevant instances from long sequences using a sigmoid layer is still a challenging task, as evidenced by slightly diminished attention to tumor instances in ABMIL compared to DTFD. In addition, random grouping and prior context-based clustering (K-Means) tend to scatter a small fraction of tumor instances across wide groups, which leads to the model not fully exploring tumor information, thus shifting to learning more normal tissue features. Consequently, these methods struggle with the risk of over-fitting or false negatives due to inadequate exploration of the mutual relationships among instance sampling, feature representation, and decision-making, resulting in an imbalanced focus on either tumor or normal instances.

D. Discussions on DPIS Scheme

As shown in Figure S7, we carry out comparative analyses of various sampling schemes on CAMELYON16 and TCGA Lung datasets. It is observed that selecting the max-feature similarity instances (GMSS) significantly improves performance on the CAMELYON16 dataset while focusing on fewer label-related instances (GHSS) enhances classification precision in the TCGA Lung dataset. Specifically, since CAMELYON16 positive WSIs contain fewer tumor tissues, robust token representation is achieved by exploring feature correlations and integrating historical data. The GMSS effectively aggregates the most informative bag representation via sampling instances most similar to the token for facilitating precise decision-making. The GHSS discards 50% of stable and specific features introducing more noise into sub-bags, which leads to an increase in false negatives. In contrast, since the TCGA lung dataset contains a large number of tumor regions with positive WSIs, LIIS and GHSS consider non-optimal instances and employ contrast loss and reward-penalty mechanisms for generalized features, thus considerably improving inference robustness. Overall, extensive experiments in WSI tasks have verified the effectiveness of our proposed PAMIL method.

Limitations Due to the complex biological structures and specific tumor distribution in WSIs, like other MIL methods [3, 10, 15, 17], our proposed GHSS scheme also considers a balanced mixture of salient and non-salient instances proportions. However, we plan to optimize the state token representation and sampling schemes for a variety of WSI datasets.

References

[1] Kingma DP Ba J Adam et al. A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 1412, 2014. 4

[2] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, 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. 3

[3] 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. 7

[4] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, pages 15750–15758, 2021. 2

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 4

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 4

[7] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *ICML*, pages 2127–2136. PMLR, 2018. 5

[8] 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 *CVPR*, pages 14318–14328, 2021. 4

[9] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *ICLR*. OpenReview.net, 2017. 4

[10] 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. 3, 4, 7

[11] Faisal Mahmood, Richard J. Chen, and Nicholas J. Durr. Unsupervised reverse domain adaptation for synthetic medical images via adversarial training. *IEEE Trans. Medical Imaging*, 37(12):2572–2581, 2018. 4

[12] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 2

[13] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997. 1

[14] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, and Xiangyang Ji. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems*, 34:2136–2147, 2021. 3, 4

[15] Wenhao Tang, Sheng Huang, Xiaoxian Zhang, Fengtao Zhou, Yi Zhang, and Bo Liu. Multiple instance learning framework with masked hard instance mining for whole slide image classification. In *CVPR*, pages 4078–4087, 2023. 4, 7

[16] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. Review the cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia*, 2015(1):68–77, 2015. 4

[17] Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E. Coupland, and Yalin Zheng. DTFD-MIL: double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *CVPR*, pages 18780–18790, 2022. 4, 5, 7

[18] Ruijie Zhang, Qiaozhe Zhang, Yingzhuang Liu, Hao Xin, Yan Liu, and Xinggang Wang. Multi-level multiple instance learning with transformer for whole slide image classification. *arXiv preprint arXiv:2306.05029*, 2023. 4

[19] Boxuan Zhao, Jun Zhang, Deheng Ye, Jian Cao, Xiao Han, Qiang Fu, and Wei Yang. Rlogist: fast observation strategy on whole-slide images with deep reinforcement learning. In *AAAI*, volume 37, pages 3570–3578, 2023. 2

[20] Zhonghang Zhu, Lequan Yu, Wei Wu, Rongshan Yu, Defu Zhang, and Liansheng Wang. Murcl: Multi-instance reinforcement contrastive learning for whole slide image classification. *IEEE Trans. Medical Imaging*, 2022. 2