SLIDETTT: SLIDE-SPECIFIC TEST-TIME TRAINING FOR GENERALIZABLE WHOLE-SLIDE IMAGES ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multiple instance learning (MIL) has become the dominant paradigm for weakly supervised whole-slide image (WSI) analysis, where large-scale WSIs are divided into patches and their representations are subsequently aggregated for slide-level prediction. However, most existing MIL methods still follow a static inference paradigm, applying a fixed aggregator learned during training to all test WSIs. In the presence of substantial inter-slide heterogeneity and cross-center distribution shifts, such a unified aggregation strategy may limit model generalization. To address this issue, we propose **SlideTTT**, a slide-specific test-time training framework that enables each WSI to adapt its aggregation process before prediction. At its core, we introduce a **Slide TTT Layer** that constructs an inner-learning problem from patch tokens, where key-value pairs dynamically update lightweight fast weights and query tokens retrieve slide-specific representations from the adapted state. The layer combines complementary global and local adaptation: the global branch captures long-range semantic dependencies, while the local branch models short-range spatial interactions through convolutional fast weights. Moreover, the Slide TTT Layer is designed as a modular component that can be readily integrated into other Transformer-based MIL architectures. We evaluate SlideTTT on five pathology tasks using paired internal and external cohorts. SlideTTT achieves the highest average accuracy across internal, external, and overall evaluations, with more pronounced gains on external cohorts, demonstrating the effectiveness of slide-specific adaptation for improving generalization to unseen data. Code is available at `https://anonymous.4open.science/r/SlideTTT`.

1 INTRODUCTION

Whole-slide images (WSIs) typically contain billions of pixels, making direct end-to-end modeling computationally prohibitive. Multiple instance learning (MIL) has therefore become the dominant paradigm for weakly supervised WSI analysis (Campanella et al., 2019). A standard MIL pipeline generally consists of two stages (Ilse et al., 2018). First, a WSI is partitioned into a large number of image patches, which are independently encoded by a pretrained feature extractor to obtain patch-level representations. Second, an MIL aggregator integrates thousands or even tens of thousands of patch features into a slide-level representation for downstream tasks such as classification and prognosis prediction (Yang et al., 2023; Yao et al., 2020). Both stages have been extensively studied in recent years (Xu et al., 2025). On the one hand, pathology foundation models such as UNI, CONCH, and Virchow have substantially improved general-purpose patch representations (Chen et al., 2024; Lu et al., 2024; Vorontsov et al., 2024). On the other hand, a broad range of aggregation architectures, from attention-based MIL to Transformer-based MIL, have been developed to improve slide-level modeling (Lu et al., 2021; Shao et al., 2021). Despite these advances, current two-stage MIL pipelines largely follow the same static inference paradigm: pretrained encoders produce fixed offline patch features, while a trained MIL aggregator processes all unseen WSIs using the same set of parameters (Tang et al., 2024).

This static modeling paradigm implicitly assumes that a single aggregation function learned from the training cohort can be directly generalized to all unseen cases. However, WSIs exhibit substantial distributional heterogeneity (Hetz et al., 2024). Differences in staining protocols, scanners, and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

tissue preparation across medical centers can lead to pronounced cross-center distribution shifts, while variations in tumor morphology, stromal composition, immune infiltration, and tissue organization further introduce considerable inter-slide heterogeneity (Lou et al., 2025). Although existing MIL methods improve representation learning through more sophisticated attention mechanisms, instance-relation modeling, or Transformer based aggregation, they ultimately learn a shared and fixed mapping that is directly applied to all test slides. As a result, these models can learn how to aggregate WSIs over the training distribution, but cannot actively adjust their aggregation behavior according to the particular test case being processed. For pathology data with substantial cross-center and inter-patient variation, such a one-model-fits-all aggregation strategy may become an important bottleneck for generalization (Nunes et al., 2025). This naturally raises the following question: *Can an MIL model first learn from a WSI itself before predicting it?*

Test-Time Training (TTT) provides a natural way to address this problem. Classical TTT constructs an auxiliary learning objective from unlabeled test samples and performs sample-conditioned model updates before prediction, thereby improving robustness under distribution shifts (Sun et al., 2020). More recently, TTT has been reformulated as a sequence modeling mechanism, where TTT layers treat model parameters as an updatable hidden state and construct an inner-learning problem from the input sequence (Sun et al., 2024). ViT^3 further extends this idea to visual modeling by using key-value pairs to dynamically update an inner model and query tokens to retrieve contextual information from the resulting fast weights, enabling efficient input-conditioned token interaction (Han et al., 2026). This paradigm is particularly well suited to WSI analysis, as each slide naturally provides a large collection of patch tokens from which a slide-specific aggregation state can be constructed. However, a WSI contains not only thousands to tens of thousands of highly heterogeneous instances, but also contextual dependencies at different spatial scales. Accurate prediction therefore requires both long-range semantic interactions across distant tissue regions and local structural relationships among spatially neighboring patches. A TTT framework tailored to WSI analysis should thus support slide-specific adaptation while jointly modeling global and local context with scalable computational complexity.

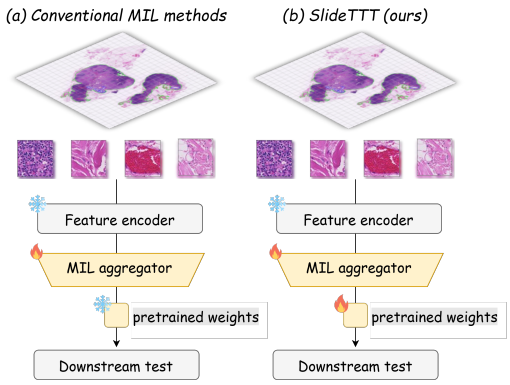

Figure 1: Existing MIL methods rely on fixed aggregators for all test WSIs, whereas SlideTTT allows each slide to adapt its aggregation process through test-time updates.

Based on these observations, we propose **SlideTTT**, a test-time training framework for WSI-level multiple instance learning that enables slide-specific adaptation during aggregation. For each input WSI, SlideTTT constructs an inner-learning problem from its patch tokens, where key-value pairs perform an on-the-fly update of lightweight fast weights and query tokens retrieve slide-conditioned representations from the adapted inner model. This formulation transforms conventional MIL aggregation from a fixed mapping into an input-conditioned process. At the core of SlideTTT is the proposed **Slide TTT Layer**, which performs complementary global and local adaptation. The global branch updates fast weights over the complete token sequence to capture long-range semantic dependencies, while the local branch adapts convolutional fast weights over the two-dimensional patch layout to model short-range spatial interactions. Their outputs are jointly integrated to obtain slide-specific representations without relying on standard quadratic-complexity self-attention. Importantly, the Slide TTT Layer is designed as a modular component and can be readily incorporated into other Transformer-based MIL backbones, extending slide-specific test-time adaptation beyond the original SlideTTT architecture. Overall, SlideTTT provides a dynamic WSI-conditioned aggregation paradigm that jointly exploits global and local context to improve generalization across unseen slides and external cohorts.

2 Related Work

2.1 Multiple Instance Learning for WSI Analysis

Most weakly supervised whole-slide image (WSI) analysis methods follow a two-stage multiple instance learning (MIL) paradigm, where patch representations are first extracted using a pretrained encoder and subsequently aggregated into a slide-level representation. Existing studies have primarily focused on improving the aggregation stage through attention-based instance weighting (Ilse et al., 2018; Lu et al., 2021), graph-based relational reasoning (Li et al., 2024; Fu et al., 2026), and Transformer-based contextual modeling (Shao et al., 2021). These designs substantially improve the interaction and selection of heterogeneous instances, but generally retain a fixed inference rule: once training is completed, the learned aggregator is directly applied to unseen slides without further adaptation. Such an assumption can be restrictive in computational pathology, where variations in staining protocols, scanners, and institutional acquisition procedures often induce substantial distribution shifts across cohorts.

Recent studies have begun to explicitly investigate generalization beyond the training cohort. StructMIL formulates WSI classification under a domain generalization setting and combines prototype-based aggregation with contrastive alignment, adversarial domain confusion, and consistency regularization to improve robustness across scanners and institutions (Yu et al., 2026). Other work has addressed out-of-distribution (OOD) WSI classification by aligning intra-slide and inter-slide patch distributions before Transformer-based MIL aggregation (Sun et al., 2025). More recently, nnMIL improves robustness on external cohorts through randomized patch/feature sampling and sliding-window ensemble inference, demonstrating strong generalization across multiple clinical tasks and pathology foundation models (Luo et al., 2026). Despite their improved robustness, these approaches still learn their adaptation or regularization strategies entirely from the training data and employ a fixed aggregation function at inference time. Consequently, the incoming WSI itself does not actively determine how its instances should be aggregated.

2.2 Test-Time Training for WSI Analysis

Test-Time Training (TTT) aims to further adapt a model using unlabeled samples available at test time, thereby improving generalization under distribution shifts. Classical TTT constructs an auxiliary self-supervised objective from each test sample and updates the model before prediction, allowing the model to extract additional adaptation signals from the current input (Sun et al., 2020). Subsequent studies, including TENT, TTT++, MEMO, and EATA, further developed the broader test-time adaptation paradigm from the perspectives of entropy minimization, improved self-supervised objectives, single-sample adaptation, and efficient and stable model updates (Wang et al., 2020; Liu et al., 2021; Zhang et al., 2022; Niu et al., 2022). More recently, TTT has evolved from conventional test-time model adaptation into a general sequence modeling mechanism. TTT Layers treat a learnable model itself as the hidden state and construct an inner-learning problem from the current sequence, producing input-conditioned fast weights through test-time updates (Sun et al., 2024). ViT^3 further extends this formulation to visual modeling, where key-value pairs dynamically update an inner model and query tokens retrieve contextual information from the updated model state, enabling visual sequence modeling with linear complexity (Han et al., 2026).

Test-time adaptation has also been increasingly explored in computational pathology to mitigate domain shifts caused by variations in staining procedures, scanning devices, and medical centers. Existing studies have investigated this paradigm across histological image registration, segmentation, and pathology foundation models. For example, style and structural guidance have been introduced to alleviate cross-domain variations in histological image registration (Zhou et al., 2024); feature alignment has been employed to improve the robustness of histopathology segmentation models across tissue sources and scanning conditions (Chroni et al., 2024); and uncertainty modeling together with causal intervention has been explored to improve the adaptation of colorectal pathology models under heterogeneous data distributions (Lou et al., 2025). These studies demonstrate that dynamically exploiting unlabeled test data can improve the cross-domain robustness of pathology models. However, existing pathology test-time adaptation methods primarily focus on adapting feature distributions, predictors, or model parameters, while the WSI-level multiple instance aggregation process itself generally remains fixed.

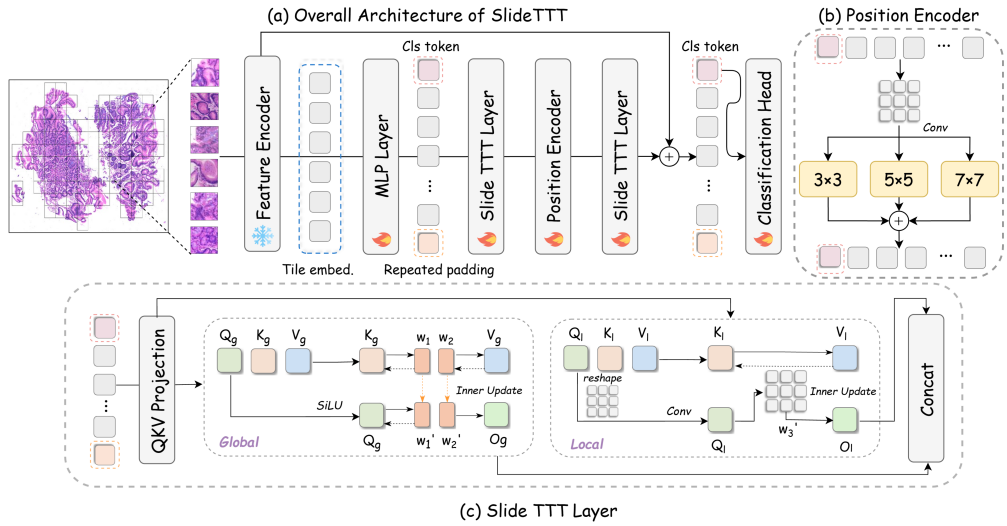

Figure 2: Overview of the proposed SlideTTT framework. (a) Overall architecture of SlideTTT with two slide TTT layers and an intermediate position encoder. (b) Multi-scale position encoder with different kernel sizes. (c) Global-local Slide TTT Layer for slide-specific adaptation.

3 METHODS

Given a WSI $\mathcal{S}$, we follow the standard patch-based paradigm for weakly supervised WSI analysis. The slide is partitioned into N image patches and encoded by a pretrained pathology encoder to obtain a set of patch-level representations $X = \{x_i\}_{i=1}^{N}$, where each $x_i \in \mathbb{R}^d$ denotes a d-dimensional patch feature. The extracted features are then projected into a shared latent space and prepended with a learnable classification token,

$$H_0 = [x_{\text{cls}}; \phi(XW_p + b_p)], \tag{1}$$

where $\phi(\cdot)$ denotes a nonlinear activation function. As illustrated in Figure 2, SlideTTT consists of two independently parameterized Slide TTT Layers with an intermediate multi-scale depthwise positional encoding module,

$$H_2 = \mathcal{T}_2\big(\mathcal{P}(\mathcal{T}_1(H_0))\big), \tag{2}$$

where $\mathcal{T}_1$ and $\mathcal{T}_2$ denote the first and second slide-specific test-time training (TTT) layers, respectively, and $\mathcal{P}$ denotes the multi-scale depthwise positional encoding module.

3.1 SLIDE TTT LAYER

The Slide TTT Layer models slide context from both global and local perspectives. Given an input sequence H, the normalized tokens are projected into two groups of query, key, and value representations, denoted as (Q_g, K_g, V_g) and (Q_l, K_l, V_l) for the global and local branches, respectively. The global branch captures long-range dependencies, while the local branch models short-range tissue interactions.

Global slide-specific learning. The global branch captures long-range dependencies across the entire slide using a lightweight gated inner model,

$$f_W(Z) = (ZW_1) \odot \text{SiLU}(ZW_2), \tag{3}$$

where $f_W(\cdot)$ denotes the global inner model parameterized by $W = \{W_1, W_2\}$, Z denotes its input token representation, W_1 and W_2 are learnable projection matrices, $\text{SiLU}(\cdot)$ is the SiLU activation function, and $\odot$ denotes element-wise multiplication.

For each WSI, its global key-value pairs define the inner-learning objective. We adopt a dot-product key-value alignment loss,

$$\mathcal{L}_g = -\frac{1}{M}\sum_{i=1}^{M}\langle f_W(k_{g,i}), v_{g,i}\rangle\,, \tag{4}$$

where $\mathcal{L}_g$ denotes the global inner-learning loss, $k_{g,i}$ and $v_{g,i}$ are the i-th global key and value representations, respectively, M is the number of valid tokens involved in the inner update, and $\langle\cdot,\cdot\rangle$ denotes the inner product.

The shared inner-model parameters are adapted with a single lightweight update,

$$W_j^{\mathcal{S}} = W_j - \eta\widehat{\nabla_{W_j}\mathcal{L}_g}, \qquad j \in \{1,2\}, \tag{5}$$

where η denotes the inner learning rate, $\nabla_{W_j}\mathcal{L}_g$ is the gradient of the global inner-learning loss with respect to W_j, and $\widehat{\nabla}$ denotes the normalized gradient direction. Accordingly, $W_1^{\mathcal{S}}$ and $W_2^{\mathcal{S}}$ serve as slide-specific fast weights and are used only for processing the current slide.

The adapted inner model is then queried by the global query representation Q_g,

$$Y_g = (Q_g W_1^{\mathcal{S}}) \odot \mathrm{SiLU}(Q_g W_2^{\mathcal{S}}), \tag{6}$$

where Q_g denotes the global query representation and Y_g is the resulting slide-specific global feature representation. In this way, different WSIs share the same base parameters W_1 and W_2, while dynamically constructing distinct fast weights during forward propagation.

Local slide-specific learning. Global dependencies alone do not explicitly capture local tissue continuity and neighborhood structure. We therefore introduce a local TTT branch that performs slide-specific short-range modeling through convolutional fast weights. For a WSI containing N patch tokens, we pad the sequence to an approximately square grid with spatial size

$$h = w = \left\lceil\sqrt{N}\right\rceil, \tag{7}$$

where h and w denote the height and width of the reconstructed token grid. The local query, key, and value representations, Q_l, K_l, and V_l, are then reshaped accordingly, while padded positions are masked out from the inner update and subsequent aggregation.

We use a 3×3 depthwise convolution as the local inner model. Similar to the global branch, the local key-value pairs define an inner-learning objective $\mathcal{L}_l$, which is used to adapt the shared convolutional kernel,

$$W_l^{\mathcal{S}} = W_l - \eta\widehat{\nabla_{W_l}\mathcal{L}_l}, \tag{8}$$

where W_l denotes the shared base convolutional kernel, η is the inner learning rate, $\nabla_{W_l}\mathcal{L}_l$ is the gradient of the local inner-learning loss with respect to W_l. The adapted kernel is then applied to the local query representation,

$$Y_l = \mathrm{DWConv}\left(Q_l; W_l^{\mathcal{S}}\right), \tag{9}$$

where Y_l denotes the resulting local contextual representation. Since the classification token has no explicit spatial location, only patch tokens participate in the local branch.

Finally, the global and local outputs are fused by concatenation and linear projection,

$$Y = W_o[Y_g; Y_l] + b_o, \tag{10}$$

where Y_g and Y_l denote the global and local representations, respectively, W_o and b_o are the learnable output projection parameters, and Y is the fused representation. A residual connection is then applied,

$$\mathcal{T}(H) = H + Y, \tag{11}$$

where $\mathcal{T}(\cdot)$ denotes the output of the Slide TTT Layer. In this way, each Slide TTT Layer combines long-range semantic dependencies with local tissue interactions, while adapting both branches to the current WSI.

3.2 Multi-Scale Depthwise Positional Encoding

To incorporate multi-scale positional information, we insert a lightweight spatial enhancement module (MDPE) between the two Slide TTT Layers. The classification token is kept unchanged, while the patch tokens are reshaped into a feature map $F \in \mathbb{R}^{d\times h\times w}$ and enhanced by depthwise convolutions with different receptive fields,

$$\widetilde{F} = F + \mathrm{DWConv}_{3\times 3}(F) + \mathrm{DWConv}_{5\times 5}(F) + \mathrm{DWConv}_{7\times 7}(F). \tag{12}$$

Here, $\widetilde{F}$ denotes the enhanced feature map, and the three kernels capture spatial context at different scales. The enhanced patch features are flattened back into a token sequence and passed to the second Slide TTT Layer together with the unchanged classification token.

After the second TTT layer, the normalized classification token is used as the WSI representation,

$$z = \mathrm{LN}(H_2)_{\mathrm{cls}}, \tag{13}$$

and the final prediction is obtained by

$$\hat{y} = \mathbf{W}_c z + \mathbf{b}_c, \tag{14}$$

where $\mathbf{W}_c$ and $\mathbf{b}_c$ denote the classifier weight matrix and bias vector, respectively.

Conventional MIL aggregators apply a fixed aggregation function to all unseen slides after training. In contrast, SlideTTT performs slide-conditioned inner learning during aggregation, allowing each WSI to dynamically adapt its aggregation state according to the observed patch tokens.

4 Experiments

4.1 Datasets

We evaluate SlideTTT on five slide-level classification tasks across breast, colorectal, and lung cancers, including ER status prediction, PAM50 molecular subtyping, PR status prediction, MSI status prediction, and NSCLC subtype classification. For each task, models are trained and selected on an internal cohort and directly evaluated on an independent external cohort without retraining. To provide a more stringent assessment of cross-cohort generalization, we consistently use the smaller cohort for model development and the larger cohort for external evaluation. Specifically, CPTAC-BRCA and TCGA-BRCA are used for the breast cancer tasks, TCGA-COAD/READ and CPTAC-COAD for MSI prediction, and TCGA-NSCLC and CPTAC-NSCLC for NSCLC classification. Detailed dataset statistics, data splits, and results are provided in the Appendix.

4.2 Implementation Details

Non-overlapping 224×224 patches are extracted at $20\times$ magnification and encoded by the frozen UNI encoder (Chen et al., 2024), yielding 1,024-dimensional patch features. All models are trained with cross-entropy loss and Adam (Kingma & Ba, 2015) using an initial learning rate of 2×10^{-4}, a weight decay of 1×10^{-5}, and a batch size of one slide. Training lasts for at most 50 epochs with early stopping based on validation macro-AUC, and results are reported over five seeds on a single NVIDIA L20 48GB GPU. For SlideTTT, features are processed by two TTT layers with eight global heads and intermediate multi-scale positional encoding using kernel sizes of 3, 5, and 7. Each slide performs a single inner update in both global and local branches with a step size of 1.0. Fast weights are constructed independently for each slide and discarded after the forward pass.

4.3 Downstream Classification Performance

To evaluate the effectiveness and cross-cohort generalization of SlideTTT, we compare it with representative MIL methods across five slide-level classification tasks under both internal and external evaluation settings, as shown in Table 1. SlideTTT achieves the highest overall average accuracy of 78.47% across all ten settings. Compared with the strongest competing methods, SlideTTT improves the average accuracy by 1.10 percentage points on the internal cohorts, 2.52 percentage points on the external cohorts, and 1.96 percentage points overall. The advantage of SlideTTT is

more pronounced under external evaluation. These results indicate that the proposed slide-specific adaptation is particularly beneficial when transferring to unseen cohorts. For the internal cohorts, the performance gains are generally smaller, suggesting that conventional MIL models can already provide competitive representations when the training and test distributions are closely matched. In contrast, under cross-cohort distribution shifts, the slide-specific inner update provides additional adaptive capacity by allowing each WSI to construct its own aggregation state.

Table 1: Comparison of accuracy (Acc, %) across internal and external cohorts. Results are reported as mean_{std} over five random seeds. Avg. denotes the average accuracy across all ten evaluation settings. The best result in each setting is highlighted in bold.

Model	Internal					External					Avg.
	BRCA-ER	BRCA-PAM50	BRCA-PR	MSI	NSCLC	BRCA-ER	BRCA-PAM50	BRCA-PR	MSI	NSCLC	
ABMIL (Ilse et al., 2018)	$83.12_{3.2}$	$82.47_{1.4}$	$75.20_{5.9}$	$70.19_{2.5}$	$91.78_{0.7}$	$76.53_{4.1}$	$54.53_{2.6}$	$67.47_{1.7}$	$74.48_{1.7}$	$89.31_{1.3}$	76.51
CLAM-SB (Lu et al., 2021)	$82.63_{2.2}$	$82.42_{3.0}$	$73.07_{4.7}$	$70.19_{3.4}$	$91.72_{1.2}$	$72.98_{3.6}$	$59.08_{1.7}$	$68.22_{1.9}$	$75.83_{0.9}$	$87.43_{6.0}$	76.36
DSMIL (Li et al., 2021)	$81.05_{3.6}$	$81.32_{2.5}$	$69.60_{5.4}$	$72.88_{2.9}$	$91.17_{0.9}$	$74.79_{3.6}$	$59.28_{1.4}$	$66.40_{3.4}$	$74.03_{5.3}$	$89.33_{1.2}$	75.99
TransMIL (Shao et al., 2021)	$78.42_{3.6}$	$77.63_{4.8}$	$75.73_{5.8}$	$68.68_{4.9}$	$90.92_{1.5}$	$70.60_{5.4}$	$54.09_{3.2}$	$65.58_{4.3}$	$76.15_{0.6}$	$87.43_{4.9}$	74.52
FR-MIL (Chikontwe et al., 2024)	$80.53_{3.1}$	$70.79_{6.4}$	$70.40_{3.3}$	$62.64_{6.3}$	$91.31_{1.0}$	$70.62_{8.3}$	$46.67_{4.9}$	$61.47_{7.3}$	$69.23_{8.4}$	$86.76_{2.9}$	71.04
ILRA (Xiang & Zhang, 2023)	$82.74_{2.4}$	$77.63_{6.2}$	$76.27_{6.6}$	$68.68_{3.4}$	$90.87_{3.5}$	$71.89_{5.4}$	$47.06_{6.2}$	$58.48_{6.5}$	$75.93_{0.8}$	$88.44_{3.0}$	73.80
WiKG (Li et al., 2024)	$81.58_{2.3}$	$80.26_{2.6}$	$70.13_{2.0}$	$67.92_{4.6}$	$91.83_{1.0}$	$59.93_{5.2}$	$49.28_{8.1}$	$53.67_{7.3}$	$67.69_{8.4}$	$87.24_{6.2}$	70.95
AEM (Zhang et al., 2025b)	$83.22_{5.7}$	$78.68_{2.9}$	$70.67_{3.9}$	$72.96_{1.6}$	$91.31_{1.6}$	$75.76_{5.7}$	$55.58_{2.3}$	$68.47_{3.6}$	$73.76_{5.3}$	$82.79_{5.0}$	75.32
AMD-MIL (Ling et al., 2024)	$83.18_{0.6}$	$75.79_{6.4}$	$76.54_{3.2}$	$69.81_{2.7}$	$90.12_{1.0}$	$67.69_{7.0}$	$42.65_{9.3}$	$61.94_{7.7}$	$76.02_{0.0}$	$88.22_{3.7}$	73.19
2DMamba (Zhang et al., 2025a)	$80.79_{3.2}$	$82.37_{2.9}$	$72.00_{7.2}$	$69.06_{7.1}$	$91.68_{0.7}$	$68.54_{5.8}$	$58.09_{2.5}$	$65.65_{3.8}$	$75.29_{0.7}$	$89.34_{3.8}$	75.28
SlideTTT	$\mathbf{83.42_{2.4}}$	$\mathbf{82.89_{2.6}}$	$\mathbf{76.80_{5.4}}$	$\mathbf{73.21_{4.7}}$	$\mathbf{91.92_{1.5}}$	$\mathbf{79.48_{1.1}}$	$\mathbf{59.63_{2.4}}$	$\mathbf{71.36_{1.8}}$	$\mathbf{76.38_{0.6}}$	$\mathbf{89.56_{2.5}}$	**78.47**

4.4 Transferability across Transformer-based MIL backbones.

To further evaluate whether the proposed Slide TTT Layer can serve as a plug-in module for different Transformer-based MIL architectures, we integrate it into two representative backbones, DGR and MHIM, and evaluate the resulting models on the internal and external cohorts of BRCA-ER, BRCA-PAM50, and NSCLC. As shown in Figure 3, incorporating the Slide TTT Layer improves balanced accuracy in most settings. The gains are particularly consistent on BRCA-PAM50, where both DGR and MHIM benefit from the additional slide-specific adaptation under internal and external evaluation, with a maximum absolute improvement of 11.64 percentage points. Clear improvements are also observed for DGR on BRCA-ER and NSCLC, and for MHIM on the external BRCA-ER cohort. Although a few settings show slight performance degradation, the overall results indicate that the Slide TTT Layer is not tied to the original SlideTTT architecture and can be flexibly integrated into other Transformer-based MIL frameworks.

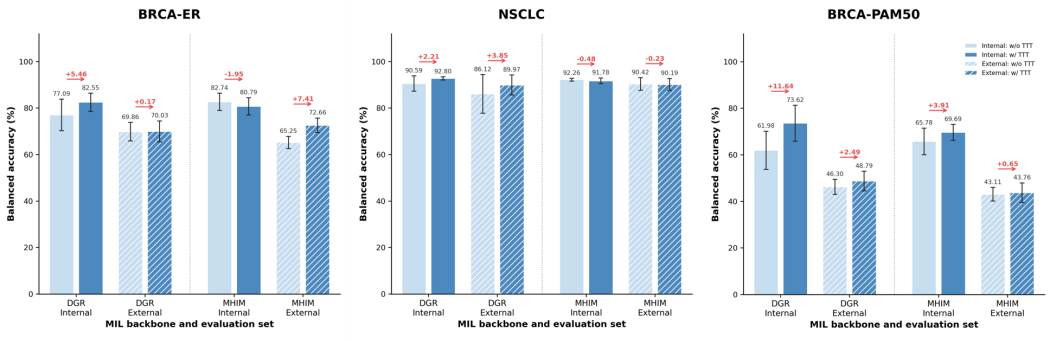

Figure 3: Performance comparison of Transformer-based MIL backbones with and without the proposed Slide TTT Layer.

4.5 Ablation Study

Effectiveness of inner adaptation and spatial enhancement. To investigate the effectiveness of slide-specific inner adaptation and spatial modeling, we conduct ablation studies on different inner-update configurations and spatial enhancement strategies. As shown in Table 2, introducing the global update improves the average accuracy from 43.98% to 69.81%, demonstrating the effectiveness of test-time inner adaptation. Further incorporating the local branch increases the average

accuracy to 75.28%, indicating that global dependencies and local contextual information provide complementary cues for slide-specific representation learning. In addition, multi-scale positional modeling improves the average accuracy from 70.31% without MDPE to 75.28%. The varying performance of single-scale variants across tasks suggests that a fixed receptive field is insufficient to consistently capture heterogeneous tissue structures, whereas combining fine- and coarse-grained spatial information provides more complementary contextual cues.

Table 2: Ablation study of inner-update configurations and spatial enhancement strategies on the external cohorts using accuracy (Acc, %). Results are reported as mean_{std} over five random seeds. The best result within each ablation group is highlighted in bold.

Dataset	Inner Update Configuration				Spatial Enhancement			
	No Update	Local Only	Global Only	Global + Local	No MDPE	Fine-Scale	Coarse-Scale	Multi-Scale
BRCA-ER	$22.08_{0.0}$	$66.75_{25.0}$	$74.76_{2.0}$	$\mathbf{79.48_{1.1}}$	$71.36_{4.4}$	$72.23_{3.7}$	$74.79_{5.7}$	$\mathbf{79.48_{1.1}}$
BRCA-PAM50	$32.33_{17.7}$	$38.09_{18.4}$	$52.80_{5.5}$	$\mathbf{59.63_{2.4}}$	$55.64_{6.1}$	$48.81_{13.3}$	$55.53_{5.4}$	$\mathbf{59.63_{2.4}}$
BRCA-PR	$39.13_{16.2}$	$46.38_{19.8}$	$55.67_{13.4}$	$\mathbf{71.36_{1.8}}$	$60.97_{12.6}$	$62.34_{8.1}$	$58.23_{18.9}$	$\mathbf{71.36_{1.8}}$
MSI	$76.02_{0.0}$	$65.61_{23.3}$	$\mathbf{76.83_{1.8}}$	$76.38_{0.6}$	$76.02_{0.0}$	$76.11_{0.5}$	$76.20_{0.4}$	$\mathbf{76.38_{0.6}}$
NSCLC	$50.35_{1.9}$	$50.35_{1.9}$	$88.98_{3.9}$	$\mathbf{89.56_{2.5}}$	$87.56_{3.9}$	$89.97_{3.9}$	$\mathbf{90.48_{1.6}}$	$89.56_{2.5}$
Avg.	43.98	53.44	69.81	**75.28**	70.31	69.89	71.05	**75.28**

Effectiveness of the inner-update loss. We investigate the choice of loss function used to optimize the fast weights during the inner update, comparing dot-product alignment, mean squared error (MSE), and cosine alignment. As shown in Table 3, the dot-product objective achieves the best accuracy across all five internal test sets, demonstrating the most consistent overall performance. This may be attributed to its more direct update signal, which is less dependent on the current prediction residual than MSE and avoids the feature-magnitude normalization introduced by cosine alignment. Such a property is particularly suitable for our single-step inner-update setting.

Table 3: Internal-test accuracy (Acc, %) with standard deviation under different inner-update objectives.

Dataset	Dot Alignment	MSE	Cosine Alignment
BRCA-ER	$\mathbf{83.42_{2.4}}$	$80.53_{1.7}$	$81.84_{2.2}$
BRCA-PAM50	$\mathbf{82.89_{2.6}}$	$81.84_{5.4}$	$77.11_{3.0}$
BRCA-PR	$\mathbf{76.80_{5.4}}$	$64.27_{7.0}$	$71.47_{6.3}$
MSI	$\mathbf{73.21_{4.7}}$	$70.19_{0.8}$	$62.26_{14.9}$
NSCLC	$\mathbf{91.92_{1.5}}$	$91.15_{1.2}$	$89.90_{1.3}$

Effectiveness of the Number of Attention Heads. We further study the effect of the number of attention heads. As shown in Figure 4, performance remains relatively stable across different head settings on most datasets. Moderate head numbers have more stable performance, while larger settings such as 16 or 32 heads introduce greater variability. Although the optimal choice varies across tasks, eight heads provide the most balanced performance across internal and external cohorts.

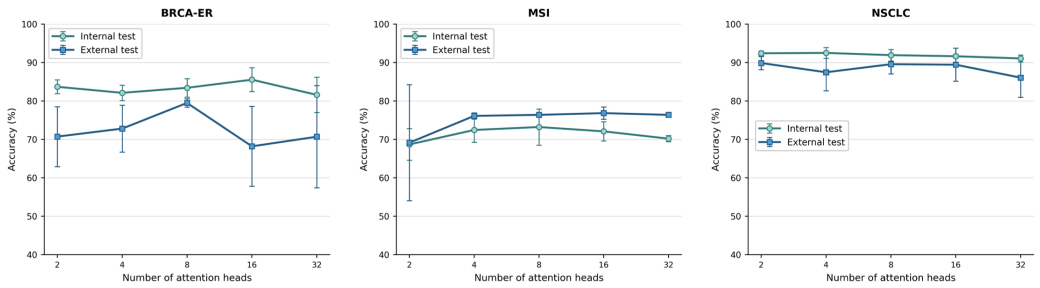

Figure 4: Effect of the number of attention heads on internal and external test accuracy. Results are reported as mean ± standard deviation over five random seeds for BRCA-ER, MSI, and NSCLC.

Effectiveness of token selection. We investigate how token selection affects both inner adaptation and final slide-level aggregation. As shown in Table 4, using the CLS and patch tokens for the inner

update consistently performs better than using patch tokens alone, indicating that the CLS token provides complementary global information during slide-specific adaptation. In contrast, for final prediction, retaining only the adapted CLS token generally performs better than further combining it with mean-pooled patch features. This suggests that the adapted CLS token already serves as an effective slide-level summary, while direct patch averaging may introduce redundant information.

Table 4: Ablation of token selection for inner updates and slide-level pooling strategies using accuracy (Acc, %). Results are reported as mean_{std} over five random seeds. The best result within each evaluation setting is highlighted in bold.

Dataset	Inner-Update Tokens				Slide-Level Pooling			
	Internal		External		Internal		External	
	CLS + Patch	Patch Only	CLS + Patch	Patch Only	CLS	CLS + mean(Patch)	CLS	CLS + mean(Patch)
BRCA-ER	$\mathbf{83.42_{2.4}}$	$82.37_{2.4}$	$\mathbf{79.48_{1.1}}$	$74.91_{5.2}$	$\mathbf{83.42_{2.4}}$	$80.79_{3.0}$	$\mathbf{79.48_{1.1}}$	$75.78_{5.8}$
BRCA-PAM50	$\mathbf{82.89_{2.6}}$	$78.42_{4.2}$	$\mathbf{59.63_{2.4}}$	$49.53_{11.1}$	$\mathbf{82.89_{2.6}}$	$77.63_{1.9}$	$\mathbf{59.63_{2.4}}$	$54.81_{5.1}$
BRCA-PR	$\mathbf{76.80_{5.4}}$	$73.33_{8.5}$	$\mathbf{71.36_{1.8}}$	$52.55_{15.0}$	$76.80_{5.4}$	$\mathbf{84.27_{4.2}}$	$\mathbf{71.36_{1.8}}$	$61.87_{9.1}$
MSI	$\mathbf{73.21_{4.7}}$	$68.68_{2.5}$	$\mathbf{76.38_{0.6}}$	$75.93_{0.2}$	$\mathbf{73.21_{4.7}}$	$66.04_{14.6}$	$\mathbf{76.38_{0.6}}$	$75.29_{2.4}$
NSCLC	$91.92_{1.5}$	$\mathbf{92.31_{0.7}}$	$\mathbf{89.56_{2.5}}$	$86.63_{2.7}$	$\mathbf{91.92_{1.5}}$	$90.67_{0.8}$	$\mathbf{89.56_{2.5}}$	$84.41_{2.8}$
Avg.	**81.65**	79.02	**75.28**	67.91	**81.65**	79.88	**75.28**	70.43

Inference Efficiency. We evaluate the inference overhead introduced by slide-specific inner updates. As shown in Figure 5, SlideTTT maintains latency comparable to conventional MIL methods while showing only limited growth as the number of patches increases. It also preserves low GPU memory consumption across different WSI sizes, using only about 1.1 GiB at 30,000 patches. These results indicate that slide-specific adaptation introduces only modest computational and memory overhead during inference.

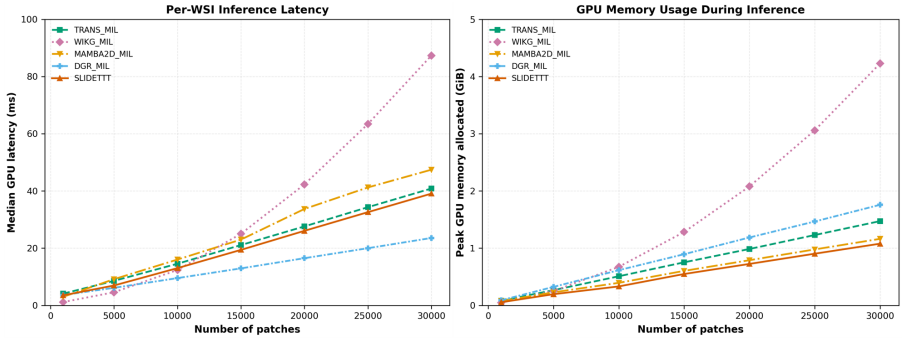

Figure 5: Inference efficiency of SlideTTT under different WSI sizes. We report per-WSI inference latency (left) and peak GPU memory usage (right) as the number of input patches increases.

5 Conclusion and Future Work

In this work, we propose SlideTTT, a slide-specific test-time training framework for whole-slide image analysis. Instead of applying a fixed aggregator to all test slides, SlideTTT constructs an inner-learning problem from each WSI and dynamically updates lightweight fast weights during inference, enabling input-conditioned slide-specific aggregation. At the core of SlideTTT, the proposed Slide TTT Layer combines complementary global and local adaptation, where the global branch captures long-range semantic dependencies and the local branch models short-range spatial interactions through convolutional fast weights. Extensive experiments on five pathology tasks with paired internal and external cohorts demonstrate the effectiveness of slide-specific adaptation, with particularly notable gains under external evaluation. We further show that the Slide TTT Layer can be integrated into other Transformer-based MIL backbones, supporting its modularity beyond the original SlideTTT architecture. In future work, we will explore extending this plug-in adaptation mechanism to a broader range of MIL frameworks and WSI analysis scenarios.

REFERENCES

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.

Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature medicine*, 30(3):850–862, 2024.

Philip Chikontwe, Meejeong Kim, Jaehoon Jeong, Hyun Jung Sung, Heounjeong Go, Soo Jeong Nam, and Sang Hyun Park. Fr-mil: distribution re-calibration-based multiple instance learning with transformer for whole slide image classification. *IEEE Transactions on Medical Imaging*, 44(1):409–421, 2024.

Evgenia Tatiani Chroni, Konstantinos M Dafnis, Georgios Chantzialexiou, Eric Cosatto, and Dimitris Metaxas. Improving test-time adaptation for histopathology image segmentation: Gradient-to-parameter ratio guided feature alignment. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pp. 1–5. IEEE, 2024.

Mingxi Fu, Xitong Ling, Yuxuan Chen, Jiawen Li, Fanglei Fu, Huaitian Yuan, Tian Guan, Yonghong He, and Lianghui Zhu. Deformable attention graph representation learning for histopathology whole slide image analysis. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6471–6475. IEEE, 2026.

Dongchen Han, Yining Li, Tianyu Li, Zixuan Cao, Ziming Wang, Jun Song, Yu Cheng, Bo Zheng, and Gao Huang. Vit 3: Unlocking test-time training in vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 51–61, 2026.

Martin J Hetz, Tabea-Clara Bucher, and Titus J Brinker. Multi-domain stain normalization for digital pathology: A cycle-consistent adversarial network for whole slide images. *Medical Image Analysis*, 94:103149, 2024.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pp. 2127–2136. Pmlr, 2018.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. URL `https://arxiv.org/abs/1412.6980`.

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 *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 14313–14323. IEEE, 2021.

Jiawen Li, Yuxuan Chen, Hongbo Chu, Qiehe Sun, Tian Guan, Anjia Han, and Yonghong He. Dynamic graph representation with knowledge-aware attention for histopathology whole slide image analysis. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11323–11332. IEEE, 2024.

Xitong Ling, Minxi Ouyang, Yizhi Wang, Xinrui Chen, Renao Yan, Hongbo Chu, Junru Cheng, Tian Guan, Sufang Tian, Xiaoping Liu, et al. Agent aggregator with mask denoise mechanism for histopathology whole slide image analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 2795–2803, 2024.

Yuejiang Liu, Parth Kothari, Bastien Van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. Ttt++: When does self-supervised test-time training fail or thrive? *Advances in Neural Information Processing Systems*, 34:21808–21820, 2021.

Shenghan Lou, Genshen Mo, Xiao Zhang, Hao Wang, Hao Li, Keru Ma, Huiying Li, Xinyue Zhang, Meihong Yan, Haonan Xie, et al. Uncertainty-aware and causal test-time adaptive foundation model for robust colorectal cancer pathology diagnosis. *NPJ Digital Medicine*, 8(1):784, 2025.

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.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature medicine*, 30(3):863–874, 2024.

Xiangde Luo, Jinxi Xiang, Yuanfeng Ji, and Ruijiang Li. nnmil: a generalizable multiple instance learning framework for computational pathology. *Nature Biomedical Engineering*, pp. 1–17, 2026.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International conference on machine learning*, pp. 16888–16905. PMLR, 2022.

João D Nunes, Diana Montezuma, Domingos Oliveira, Tania Pereira, Inti Zlobec, Isabel Macedo Pinto, and Jaime S Cardoso. Bridging domain gaps in computational pathology: A comparative study of adaptation strategies. *Sensors*, 25(9):2856, 2025.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021.

Kai Sun, Kai Huang, Jiaqi Huang, Maoxu Zhou, and Gang Yu. Accurate classification of pathological whole-slide images for out-of-distribution generalization. *International Journal of Intelligent Systems*, 2025(1):9988577, 2025.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pp. 9229–9248. PMLR, 2020.

Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, et al. Learning to (learn at test time): Rnns with expressive hidden states. *arXiv preprint arXiv:2407.04620*, 2024.

Wenhao Tang, Fengtao Zhou, Sheng Huang, Xiang Zhu, Yi Zhang, and Bo Liu. Feature re-embedding: Towards foundation model-level performance in computational pathology. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11343–11352. IEEE, 2024.

Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nature medicine*, 30(10): 2924–2935, 2024.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020.

Jinxi Xiang and Jun Zhang. Exploring low-rank property in multiple instance learning for whole slide image classification. *ICLR*, 2(6):7, 2023.

Hongming Xu, Mingkang Wang, Duanbo Shi, Huamin Qin, Yunpeng Zhang, Zaiyi Liu, Anant Madabhushi, Peng Gao, Fengyu Cong, and Cheng Lu. When multiple instance learning meets foundation models: Advancing histological whole slide image analysis. *Medical Image Analysis*, 101: 103456, 2025.

Litao Yang, Deval Mehta, Sidong Liu, Dwarikanath Mahapatra, Antonio Di Ieva, and Zongyuan Ge. Tpmil: Trainable prototype enhanced multiple instance learning for whole slide image classification. *arXiv preprint arXiv:2305.00696*, 2023.

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.

Zhenjun Yu, Zhelin Xia, Donghao Xu, Zhiyuan Zhang, Lingling Zhang, Peng Zhang, Liang Wu, Bibo Wang, Helin Wang, and Zhenxiong Zhao. Structure-aware generalization for heterogeneous histopathology via prototype-based multiple instance learning. *npj Digital Medicine*, 9(1):114, 2026.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3583–3592. IEEE, 2025a.

Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. *Advances in neural information processing systems*, 35:38629–38642, 2022.

Yunlong Zhang, Honglin Li, Yuxuan Sun, Zhongyi Shui, Jingxiong Li, Chenglu Zhu, and Lin Yang. Aem: attention entropy maximization for multiple instance learning based whole slide image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 45–55. Springer, 2025b.

Shenglong Zhou, Zhiwei Xiong, and Feng Wu. Test-time adaptation via style and structure guidance for histological image registration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 7677–7685, 2024.

A Appendix

AI Use Statement

Generative AI tools were used to assist with language editing, grammatical refinement, and improving the clarity and presentation of the manuscript. All scientific ideas, methodological designs, experimental results, analyses, and conclusions were developed and verified by the authors. All AI-assisted text was manually reviewed and revised by the authors, who take full responsibility for the final content of this work.

Ethics Statement

This study uses existing pathology datasets for methodological research on whole-slide image analysis and does not involve the prospective recruitment of human participants. All datasets were used in accordance with their corresponding data-access and usage requirements. The proposed method is intended for research purposes and is not designed for direct clinical decision-making without further validation.

Reproducibility Statement

To facilitate reproducibility, we provide detailed descriptions of the SlideTTT architecture, slide-specific inner-update procedure, training settings, dataset splits, and evaluation protocols in the main paper and Appendix. All reported experiments are conducted using consistent data splits and five random seeds. We additionally provide an anonymous implementation of SlideTTT, including the model and evaluation pipeline, to support reproduction of the reported results.

A.1 Datasets

We evaluate our method on five slide-level classification tasks spanning breast, colorectal, and lung cancers. For breast cancer, we consider binary estrogen receptor (ER) status prediction, four-class PAM50 molecular subtyping, and binary progesterone receptor (PR) status prediction. CPTAC-BRCA serves as the internal cohort, containing 384, 379, and 372 WSIs for ER, PAM50, and PR, respectively, while TCGA-BRCA serves as the corresponding external cohort with 806, 972, and 803 WSIs. For colorectal cancer, we evaluate binary microsatellite instability (MSI) status prediction using TCGA-COAD/READ as the internal cohort with 265 WSIs and CPTAC-COAD as the external cohort with 221 WSIs. For lung cancer, we evaluate binary histological subtype classification of non-small cell lung cancer (NSCLC), distinguishing lung adenocarcinoma (LUAD) and lung squamous

cell carcinoma (LUSC). TCGA-NSCLC serves as the internal cohort with 1,040 WSIs, and CPTAC-NSCLC serves as the external cohort with 630 WSIs. Detailed dataset statistics and data splits are summarized in Table 5.

Table 5: Dataset statistics for the five downstream classification tasks. Models are trained and selected on the internal cohort and directly evaluated on the paired external cohort without retraining.

Task	Internal → External	Train	Val.	Test	External
BRCA ER	CPTAC-BRCA → TCGA-BRCA	269	39	76	806
BRCA PAM50	CPTAC-BRCA → TCGA-BRCA	265	38	76	972
BRCA PR	CPTAC-BRCA → TCGA-BRCA	260	37	75	803
COAD MSI	TCGA-COAD/READ → CPTAC-COAD	186	26	53	221
NSCLC	TCGA-NSCLC → CPTAC-NSCLC	728	104	208	630

A.2 RESULTS

Table 6: Comparison of AUC (%) across internal and external cohorts. Results are reported as mean_{std} over five random seeds. Avg. denotes the average performance across all ten evaluation settings.

Model	Internal					External					Avg.
	BRCA-ER	BRCA-PAM50	BRCA-PR	MSI	NSCLC	BRCA-ER	BRCA-PAM50	BRCA-PR	MSI	NSCLC	
ABMIL	$91.84_{1.6}$	$92.65_{0.6}$	$84.77_{4.9}$	$63.99_{1.9}$	$97.24_{0.5}$	$79.12_{3.1}$	$69.66_{1.3}$	$69.70_{2.7}$	$57.55_{5.5}$	$96.60_{0.6}$	80.32
CLAM-SB	$91.55_{1.2}$	$92.60_{0.8}$	$84.13_{3.7}$	$62.18_{5.1}$	$97.18_{0.3}$	$80.89_{1.0}$	$70.58_{0.7}$	$72.70_{1.3}$	$58.84_{8.3}$	$96.62_{0.8}$	80.73
DSMIL	$89.85_{2.0}$	$92.69_{0.3}$	$76.97_{2.3}$	$\mathbf{64.57}_{3.4}$	$97.13_{0.3}$	$77.26_{1.0}$	$69.54_{1.9}$	$65.36_{5.1}$	$58.98_{5.9}$	$96.59_{0.3}$	78.89
TransMIL	$88.67_{1.9}$	$91.65_{1.3}$	$83.65_{5.7}$	$61.82_{6.5}$	$97.30_{0.5}$	$74.53_{3.1}$	$68.37_{2.1}$	$68.06_{2.4}$	$58.53_{4.0}$	$96.55_{0.5}$	78.51
FR-MIL	$87.57_{2.0}$	$85.18_{3.3}$	$77.95_{5.4}$	$56.55_{4.8}$	$96.61_{0.6}$	$71.20_{2.3}$	$63.46_{3.0}$	$65.92_{2.6}$	$56.17_{6.0}$	$95.61_{1.1}$	75.62
ILRA	$90.29_{2.6}$	$92.06_{2.1}$	$84.29_{5.8}$	$61.42_{4.7}$	$96.86_{0.6}$	$69.68_{3.3}$	$63.97_{2.3}$	$63.92_{4.1}$	$46.59_{9.0}$	$95.75_{1.5}$	76.48
WiKG	$91.52_{1.0}$	$93.10_{0.4}$	$79.58_{4.2}$	$63.72_{4.6}$	$97.15_{0.4}$	$72.47_{2.4}$	$69.42_{2.1}$	$69.72_{1.9}$	$58.83_{10.8}$	$96.60_{1.0}$	79.21
AEM	$92.36_{1.4}$	$92.26_{1.3}$	$82.41_{4.7}$	$63.45_{4.7}$	$96.94_{0.4}$	$78.20_{2.3}$	$69.95_{1.1}$	$70.23_{2.5}$	$58.28_{16.0}$	$95.55_{1.8}$	79.96
AMD-MIL	$92.10_{0.9}$	$92.40_{1.4}$	$\mathbf{86.58}_{1.9}$	$61.52_{2.6}$	$96.87_{0.4}$	$76.80_{2.4}$	$69.09_{2.7}$	$71.69_{4.3}$	$59.11_{8.4}$	$96.60_{0.8}$	80.28
2DMamba	$91.12_{1.0}$	$92.74_{1.6}$	$80.85_{6.0}$	$62.30_{1.7}$	$97.22_{0.5}$	$78.04_{1.5}$	$70.12_{2.1}$	$68.49_{1.4}$	$57.01_{3.7}$	$96.50_{1.1}$	79.44
SlideTTT	$\mathbf{92.61}_{1.0}$	$\mathbf{93.29}_{1.1}$	$85.58_{6.9}$	$64.46_{9.8}$	$\mathbf{97.45}_{0.8}$	$\mathbf{82.03}_{0.9}$	$\mathbf{70.83}_{1.2}$	$\mathbf{73.78}_{1.1}$	$\mathbf{59.18}_{20.0}$	$\mathbf{96.68}_{2.2}$	**81.59**

Table 7: Comparison of F1 score (%) across internal and external cohorts. Results are reported as mean_{std} over five random seeds. Avg. denotes the average performance across all ten evaluation settings.

Model	Internal					External					Avg.
	BRCA-ER	BRCA-PAM50	BRCA-PR	MSI	NSCLC	BRCA-ER	BRCA-PAM50	BRCA-PR	MSI	NSCLC	
ABMIL	$80.03_{4.1}$	$72.02_{1.5}$	$74.04_{6.0}$	$49.73_{4.2}$	$91.47_{0.7}$	$65.74_{2.9}$	$39.75_{1.5}$	$57.76_{6.5}$	$43.94_{2.0}$	$88.50_{1.3}$	66.30
CLAM-SB	$80.09_{2.5}$	$73.06_{3.5}$	$72.82_{4.8}$	$49.85_{10.0}$	$91.38_{1.2}$	$68.34_{2.4}$	$\mathbf{42.23}_{0.7}$	$64.94_{1.0}$	$44.56_{3.5}$	$87.23_{6.3}$	67.45
DSMIL	$79.17_{3.9}$	$71.89_{3.2}$	$68.59_{7.1}$	$50.55_{5.8}$	$91.16_{0.9}$	$67.32_{1.7}$	$40.65_{1.5}$	$57.40_{5.5}$	$44.59_{7.3}$	$88.33_{1.2}$	65.97
TransMIL	$77.32_{3.8}$	$65.58_{7.4}$	$75.62_{5.9}$	$46.95_{8.2}$	$90.91_{1.5}$	$62.42_{2.4}$	$38.03_{2.0}$	$61.65_{2.1}$	$43.35_{6.1}$	$87.29_{5.1}$	64.91
FR-MIL	$79.69_{3.4}$	$56.50_{6.6}$	$70.00_{3.4}$	$48.00_{4.6}$	$90.28_{1.0}$	$62.08_{4.7}$	$33.40_{2.7}$	$58.76_{5.7}$	$43.23_{3.5}$	$86.69_{2.9}$	62.86
ILRA	$80.07_{2.4}$	$63.95_{12.1}$	$75.76_{7.1}$	$47.64_{5.5}$	$90.79_{3.6}$	$60.25_{4.5}$	$29.48_{2.9}$	$52.91_{7.1}$	$44.49_{3.2}$	$88.39_{3.1}$	63.37
WiKG	$79.78_{3.2}$	$72.34_{2.6}$	$69.95_{1.9}$	$50.69_{9.1}$	$90.81_{0.9}$	$57.13_{3.7}$	$35.23_{3.9}$	$53.10_{7.5}$	$41.74_{4.2}$	$87.06_{6.5}$	63.78
AEM	$80.44_{6.0}$	$65.26_{3.1}$	$68.76_{8.0}$	$48.36_{1.9}$	$90.30_{1.6}$	$67.28_{3.2}$	$35.64_{5.3}$	$57.99_{5.3}$	$44.49_{3.6}$	$82.45_{5.3}$	64.10
AMD-MIL	$\mathbf{82.33}_{1.1}$	$67.33_{6.2}$	$75.62_{3.3}$	$46.93_{8.2}$	$90.09_{1.0}$	$62.16_{4.1}$	$31.22_{7.3}$	$60.93_{7.0}$	$43.19_{0.0}$	$88.14_{3.8}$	64.79
2DMamba	$78.84_{4.5}$	$72.08_{3.8}$	$70.21_{10.6}$	$49.13_{3.4}$	$91.35_{0.7}$	$64.05_{3.9}$	$40.28_{2.3}$	$61.66_{1.6}$	$43.97_{1.5}$	$89.28_{3.8}$	66.09
SlideTTT	$81.51_{3.4}$	$\mathbf{73.86}_{5.5}$	$\mathbf{76.56}_{5.3}$	$\mathbf{50.74}_{13.3}$	$\mathbf{91.91}_{1.5}$	$\mathbf{70.98}_{0.7}$	$41.38_{1.6}$	$\mathbf{65.65}_{2.1}$	$\mathbf{44.72}_{2.5}$	$\mathbf{89.55}_{2.5}$	**68.69**

Table 8: Comparison of balanced accuracy (BAcc, %) across internal and external cohorts. Results are reported as mean_{std} over five random seeds. Avg. denotes the average performance across all ten evaluation settings.

Model	Internal					External					Avg.
	BRCA-ER	BRCA-PAM50	BRCA-PR	MSI	NSCLC	BRCA-ER	BRCA-PAM50	BRCA-PR	MSI	NSCLC	
ABMIL	$79.84_{4.8}$	$70.82_{2.2}$	$75.56_{6.3}$	$53.07_{2.9}$	$91.44_{0.7}$	$67.34_{5.9}$	$43.65_{1.7}$	$59.94_{6.5}$	$49.50_{1.2}$	$88.69_{1.3}$	67.98
CLAM-SB	$79.54_{2.7}$	$\mathbf{71.71}_{3.2}$	$73.41_{4.8}$	$52.72_{7.0}$	$91.39_{1.1}$	$69.01_{0.9}$	$\mathbf{47.84}_{0.9}$	$64.03_{1.4}$	$50.09_{1.9}$	$87.79_{5.7}$	68.75
DSMIL	$79.72_{4.4}$	$70.56_{3.2}$	$69.28_{6.6}$	$54.32_{4.7}$	$91.14_{0.8}$	$70.05_{1.3}$	$45.94_{2.6}$	$58.59_{4.5}$	$50.18_{6.8}$	$89.46_{1.1}$	67.92
TransMIL	$77.67_{4.0}$	$62.60_{6.5}$	$75.90_{5.9}$	$53.45_{5.0}$	$90.91_{1.5}$	$65.91_{3.9}$	$42.85_{1.8}$	$62.66_{1.4}$	$50.00_{3.0}$	$87.77_{4.6}$	66.97
FR-MIL	$78.43_{3.6}$	$54.93_{5.8}$	$70.62_{3.6}$	$50.54_{2.6}$	$90.24_{1.1}$	$64.44_{1.5}$	$35.25_{4.4}$	$61.52_{2.8}$	$49.22_{1.3}$	$87.06_{2.7}$	64.22
ILRA	$80.11_{2.3}$	$63.76_{11.3}$	$76.14_{6.9}$	$52.70_{4.8}$	$90.28_{3.6}$	$61.75_{5.2}$	$34.93_{2.2}$	$58.49_{5.5}$	$50.46_{1.6}$	$88.51_{3.0}$	65.71
WiKG	$79.30_{3.9}$	$70.09_{3.8}$	$71.33_{2.5}$	$53.39_{5.4}$	$90.80_{1.0}$	$66.84_{1.5}$	$43.54_{1.6}$	$61.51_{3.6}$	$50.35_{1.7}$	$87.52_{5.9}$	67.47
AEM	$80.50_{6.1}$	$63.35_{3.0}$	$70.41_{5.8}$	$54.00_{1.3}$	$90.34_{1.6}$	$69.19_{2.5}$	$39.77_{6.3}$	$59.44_{3.6}$	$50.10_{1.0}$	$83.30_{4.8}$	66.04
AMD-MIL	$\mathbf{82.53}_{1.7}$	$67.96_{6.7}$	$76.58_{2.7}$	$52.84_{3.9}$	$90.55_{1.0}$	$68.80_{4.1}$	$40.35_{5.2}$	$64.77_{4.1}$	$50.00_{0.0}$	$88.55_{3.5}$	68.29
2DMamba	$78.53_{5.2}$	$70.06_{3.3}$	$71.78_{9.0}$	$53.39_{2.7}$	$91.33_{0.7}$	$70.11_{0.7}$	$44.46_{1.7}$	$62.99_{1.7}$	$49.91_{0.7}$	$89.14_{3.6}$	68.17
SlideTTT	$80.79_{4.4}$	$71.48_{6.3}$	$\mathbf{77.22}_{4.4}$	$\mathbf{55.63}_{7.8}$	$\mathbf{91.88}_{1.6}$	$\mathbf{71.69}_{1.8}$	$45.80_{1.6}$	$\mathbf{65.55}_{2.6}$	$\mathbf{50.75}_{1.2}$	$\mathbf{89.69}_{2.6}$	**70.05**

Table 9: Ablation study of different inner objectives using accuracy (Acc, %). Results are reported as mean_{std} over five random seeds. Avg. denotes the average accuracy across all ten internal and external evaluation settings. The best result in each setting is highlighted in bold.

Inner Objective	BRCA-ER		BRCA-PAM50		BRCA-PR		MSI		NSCLC		Avg.
	Int.	Ext.	Int.	Ext.	Int.	Ext.	Int.	Ext.	Int.	Ext.	
Dot alignment	$\mathbf{83.42}_{2.4}$	$\mathbf{79.48}_{1.1}$	$\mathbf{82.89}_{2.6}$	$\mathbf{59.63}_{2.4}$	$\mathbf{76.80}_{5.4}$	$\mathbf{71.36}_{1.8}$	$\mathbf{73.21}_{4.7}$	$76.38_{0.6}$	$\mathbf{91.92}_{1.5}$	$89.56_{2.5}$	**77.95**
MSE	$80.53_{1.7}$	$69.50_{16.4}$	$81.84_{5.4}$	$55.06_{5.0}$	$64.27_{7.0}$	$60.27_{11.3}$	$70.19_{0.8}$	$74.84_{2.4}$	$91.15_{1.2}$	$\mathbf{90.19}_{2.3}$	73.78
Cosine alignment	$81.84_{2.2}$	$70.55_{7.7}$	$77.11_{3.0}$	$57.43_{3.3}$	$71.47_{6.3}$	$51.53_{16.4}$	$62.26_{14.9}$	$\mathbf{76.83}_{1.8}$	$89.90_{1.3}$	$86.38_{2.2}$	72.53

Table 10: Ablation study of different inner objectives using macro-AUC (%). Results are reported as mean_{std} over five random seeds. Avg. denotes the average macro-AUC across all ten internal and external evaluation settings. The best result in each setting is highlighted in bold.

Inner Objective	BRCA-ER		BRCA-PAM50		BRCA-PR		MSI		NSCLC		Avg.
	Int.	Ext.	Int.	Ext.	Int.	Ext.	Int.	Ext.	Int.	Ext.	
Dot alignment	$\mathbf{92.61}_{1.0}$	$\mathbf{82.03}_{0.9}$	$93.29_{1.1}$	$70.83_{1.2}$	$\mathbf{85.58}_{6.9}$	$\mathbf{73.78}_{1.1}$	$64.46_{9.8}$	$59.18_{20.0}$	$\mathbf{97.45}_{0.8}$	$\mathbf{96.68}_{2.2}$	**81.59**
MSE	$88.77_{2.6}$	$74.69_{5.9}$	$\mathbf{94.01}_{2.4}$	$\mathbf{71.25}_{1.2}$	$75.57_{7.6}$	$61.69_{5.7}$	$61.11_{13.9}$	$55.77_{13.0}$	$96.89_{1.1}$	$\mathbf{96.68}_{1.9}$	77.64
Cosine alignment	$89.93_{1.0}$	$73.61_{3.9}$	$92.95_{1.4}$	$71.16_{2.0}$	$83.23_{3.8}$	$65.55_{6.0}$	$\mathbf{65.03}_{9.1}$	$\mathbf{62.05}_{18.9}$	$96.69_{0.3}$	$95.88_{1.1}$	79.61

Table 11: Ablation study of different inner objectives using balanced accuracy (BAcc, %). Results are reported as mean_{std} over five random seeds. Avg. denotes the average balanced accuracy across all ten internal and external evaluation settings. The best result in each setting is highlighted in bold.

Inner Objective	BRCA-ER		BRCA-PAM50		BRCA-PR		MSI		NSCLC		Avg.
	Int.	Ext.	Int.	Ext.	Int.	Ext.	Int.	Ext.	Int.	Ext.	
Dot alignment	$\mathbf{80.79}_{4.4}$	$\mathbf{71.69}_{1.8}$	$\mathbf{71.48}_{6.3}$	$45.80_{1.6}$	$\mathbf{77.22}_{4.4}$	$\mathbf{64.08}_{4.0}$	$\mathbf{55.62}_{7.8}$	$50.75_{1.2}$	$\mathbf{91.88}_{1.5}$	$89.69_{2.6}$	**69.90**
MSE	$78.71_{2.1}$	$65.98_{4.2}$	$69.27_{7.8}$	$\mathbf{45.97}_{1.2}$	$62.70_{8.4}$	$54.61_{4.4}$	$50.62_{1.4}$	$51.42_{5.4}$	$91.09_{1.2}$	$\mathbf{90.35}_{2.1}$	66.07
Cosine alignment	$78.72_{3.1}$	$63.99_{7.5}$	$64.88_{3.0}$	$44.05_{1.7}$	$71.84_{5.1}$	$57.63_{6.4}$	$53.11_{3.0}$	$\mathbf{53.64}_{7.8}$	$89.89_{1.4}$	$86.56_{2.0}$	66.43

Table 12: Ablation study of different inner objectives using macro-F1 (%). Results are reported as mean_{std} over five random seeds. Avg. denotes the average macro-F1 across all ten internal and external evaluation settings. The best result in each setting is highlighted in bold.

Inner Objective	BRCA-ER		BRCA-PAM50		BRCA-PR		MSI		NSCLC		Avg.
	Int.	Ext.	Int.	Ext.	Int.	Ext.	Int.	Ext.	Int.	Ext.	
Dot alignment	$\mathbf{81.51}_{3.4}$	$\mathbf{70.98}_{0.7}$	$\mathbf{73.86}_{5.5}$	$\mathbf{41.38}_{1.6}$	$\mathbf{76.56}_{5.3}$	$\mathbf{61.39}_{9.7}$	$\mathbf{50.74}_{13.3}$	$44.72_{2.5}$	$\mathbf{91.91}_{1.5}$	$89.55_{2.5}$	**68.26**
MSE	$79.01_{1.8}$	$61.91_{12.1}$	$71.91_{7.6}$	$40.49_{3.2}$	$59.18_{14.1}$	$48.40_{8.2}$	$42.38_{2.8}$	$46.53_{8.5}$	$91.13_{1.2}$	$\mathbf{90.17}_{2.3}$	63.11
Cosine alignment	$79.69_{2.8}$	$60.16_{5.5}$	$66.76_{2.6}$	$39.97_{1.3}$	$70.85_{6.9}$	$47.60_{17.5}$	$45.29_{9.6}$	$\mathbf{48.77}_{11.5}$	$89.88_{1.3}$	$86.24_{2.3}$	63.52

Table 13: Ablation of token selection for inner updates and slide-level pooling strategies using macro-AUC (%). Results are reported as mean_{std} over five random seeds. The best result within each evaluation setting is highlighted in bold.

Dataset	Inner-Update Tokens				Slide-Level Pooling			
	Internal		External		Internal		External	
	CLS + Patch	Patch Only	CLS + Patch	Patch Only	CLS	CLS + mean(Patch)	CLS	CLS + mean(Patch)
BRCA-ER	$\mathbf{92.61}_{(1.0)}$	$91.40_{(0.7)}$	$\mathbf{82.03}_{(0.9)}$	$78.90_{(3.4)}$	$\mathbf{92.61}_{(1.0)}$	$90.05_{(3.2)}$	$\mathbf{82.03}_{(0.9)}$	$75.08_{(2.8)}$
BRCA-PAM50	$\mathbf{93.29}_{(1.1)}$	$92.29_{(2.9)}$	$\mathbf{70.83}_{(1.2)}$	$68.91_{(5.8)}$	$93.29_{(1.1)}$	$\mathbf{94.94}_{(1.0)}$	$\mathbf{70.83}_{(1.2)}$	$68.31_{(0.8)}$
BRCA-PR	$\mathbf{85.58}_{(6.9)}$	$83.82_{(7.6)}$	$\mathbf{73.78}_{(1.1)}$	$69.71_{(2.5)}$	$85.58_{(6.9)}$	$\mathbf{93.14}_{(3.4)}$	$\mathbf{73.78}_{(1.1)}$	$70.05_{(1.7)}$
MSI	$\mathbf{64.46}_{(9.8)}$	$64.29_{(3.4)}$	$59.18_{(20.0)}$	$\mathbf{66.90}_{(5.8)}$	$64.46_{(9.8)}$	$\mathbf{66.28}_{(4.5)}$	$59.18_{(20.0)}$	$\mathbf{66.78}_{(4.5)}$
NSCLC	$\mathbf{97.45}_{(0.8)}$	$97.44_{(0.3)}$	$\mathbf{96.68}_{(2.2)}$	$96.12_{(0.8)}$	$\mathbf{97.45}_{(0.8)}$	$96.17_{(0.7)}$	$\mathbf{96.68}_{(2.2)}$	$95.01_{(1.8)}$
Avg.	**86.68**	85.85	**76.50**	76.11	86.68	**88.12**	**76.50**	75.05

Table 14: Ablation of token selection for inner updates and slide-level pooling strategies using macro-F1 (%). Results are reported as mean_{std} over five random seeds. The best result within each evaluation setting is highlighted in bold.

Dataset	Inner-Update Tokens				Slide-Level Pooling			
	Internal		External		Internal		External	
	CLS + Patch	Patch Only	CLS + Patch	Patch Only	CLS	CLS + mean(Patch)	CLS	CLS + mean(Patch)
BRCA-ER	$\mathbf{81.51}_{(3.4)}$	$80.65_{(3.6)}$	$\mathbf{70.98}_{(0.7)}$	$68.11_{(3.5)}$	$\mathbf{81.51}_{(3.4)}$	$78.16_{(3.6)}$	$\mathbf{70.98}_{(0.7)}$	$63.55_{(2.1)}$
BRCA-PAM50	$\mathbf{73.86}_{(5.5)}$	$68.89_{(5.5)}$	$\mathbf{41.38}_{(1.6)}$	$35.05_{(6.6)}$	$\mathbf{73.86}_{(5.5)}$	$67.15_{(2.2)}$	$\mathbf{41.38}_{(1.6)}$	$34.01_{(3.7)}$
BRCA-PR	$\mathbf{76.56}_{(5.3)}$	$73.19_{(8.5)}$	$\mathbf{61.39}_{(9.7)}$	$49.53_{(16.4)}$	$76.56_{(5.3)}$	$\mathbf{84.05}_{(4.3)}$	$\mathbf{61.39}_{(9.7)}$	$59.01_{(7.5)}$
MSI	$\mathbf{50.74}_{(13.3)}$	$44.23_{(7.0)}$	$44.72_{(2.5)}$	$\mathbf{47.85}_{(10.4)}$	$50.74_{(13.3)}$	$\mathbf{56.78}_{(11.2)}$	$44.72_{(2.5)}$	$\mathbf{49.16}_{(8.8)}$
NSCLC	$91.91_{(1.5)}$	$\mathbf{92.30}_{(0.7)}$	$\mathbf{89.55}_{(2.5)}$	$86.56_{(2.8)}$	$\mathbf{91.91}_{(1.5)}$	$90.66_{(0.8)}$	$\mathbf{89.55}_{(2.5)}$	$84.30_{(2.9)}$
Avg.	**74.92**	71.85	**61.60**	57.42	74.92	**75.36**	**61.60**	58.01

Table 15: Ablation of token selection for inner updates and slide-level pooling strategies using balanced accuracy (BAcc, %). Results are reported as mean_{std} over five random seeds. The best result within each evaluation setting is highlighted in bold.

Dataset	Inner-Update Tokens				Slide-Level Pooling			
	Internal		External		Internal		External	
	CLS + Patch	Patch Only	CLS + Patch	Patch Only	CLS	CLS + mean(Patch)	CLS	CLS + mean(Patch)
BRCA-ER	$\mathbf{80.79}_{(4.4)}$	$80.33_{(4.5)}$	$\mathbf{71.69}_{(1.8)}$	$71.50_{(2.3)}$	$\mathbf{80.79}_{(4.4)}$	$76.94_{(3.5)}$	$\mathbf{71.69}_{(1.8)}$	$63.89_{(3.0)}$
BRCA-PAM50	$\mathbf{71.48}_{(6.3)}$	$67.51_{(4.9)}$	$\mathbf{45.80}_{(1.6)}$	$42.46_{(4.7)}$	$\mathbf{71.48}_{(6.3)}$	$64.13_{(2.4)}$	$\mathbf{45.80}_{(1.6)}$	$40.21_{(2.2)}$
BRCA-PR	$\mathbf{77.22}_{(4.4)}$	$74.00_{(8.2)}$	$\mathbf{64.08}_{(4.0)}$	$59.52_{(6.2)}$	$77.22_{(4.4)}$	$\mathbf{84.15}_{(4.4)}$	$\mathbf{64.08}_{(4.0)}$	$62.62_{(2.3)}$
MSI	$\mathbf{55.62}_{(7.8)}$	$51.32_{(2.9)}$	$50.75_{(1.2)}$	$\mathbf{53.30}_{(7.4)}$	$55.62_{(7.8)}$	$\mathbf{59.36}_{(9.4)}$	$50.75_{(1.2)}$	$\mathbf{54.04}_{(6.9)}$
NSCLC	$91.88_{(1.5)}$	$\mathbf{92.29}_{(0.6)}$	$\mathbf{89.69}_{(2.6)}$	$86.89_{(2.5)}$	$\mathbf{91.88}_{(1.5)}$	$90.67_{(0.8)}$	$\mathbf{89.69}_{(2.6)}$	$84.77_{(2.7)}$
Avg.	**75.40**	73.09	**64.40**	62.73	**75.40**	75.05	**64.40**	61.11

Table 16: Ablation of different spatial enhancement strategies using accuracy (Acc, %). Results are reported as mean_{std} over five random seeds. Fine- and coarse-scale MDPE denote spatial enhancement with relatively small and large receptive fields, respectively. The best result within each evaluation setting is highlighted in bold.

Dataset	Internal				External			
	No MDPE	Fine-Scale	Coarse-Scale	Multi-Scale	No MDPE	Fine-Scale	Coarse-Scale	Multi-Scale
BRCA-ER	$84.47_{2.9}$	$83.95_{3.0}$	$\mathbf{85.53_{0.9}}$	$83.42_{2.4}$	$71.36_{4.4}$	$72.23_{3.7}$	$74.79_{5.7}$	$\mathbf{79.48_{1.1}}$
BRCA-PAM50	$79.47_{3.2}$	$82.63_{3.1}$	$79.74_{3.4}$	$\mathbf{82.89_{2.6}}$	$55.64_{6.1}$	$48.81_{13.3}$	$55.53_{5.4}$	$\mathbf{59.63_{2.4}}$
BRCA-PR	$74.67_{10.3}$	$\mathbf{77.33_{3.4}}$	$74.67_{2.8}$	$76.80_{5.4}$	$60.97_{12.6}$	$62.34_{8.1}$	$58.23_{18.9}$	$\mathbf{71.36_{1.8}}$
MSI	$69.81_{0.0}$	$67.92_{6.5}$	$70.57_{1.7}$	$\mathbf{73.21_{4.7}}$	$76.02_{0.0}$	$76.11_{0.5}$	$76.20_{0.4}$	$\mathbf{76.38_{0.6}}$
NSCLC	$91.63_{1.9}$	$91.91_{0.9}$	$91.15_{2.6}$	$\mathbf{91.92_{1.5}}$	$87.56_{3.9}$	$89.97_{3.9}$	$\mathbf{90.48_{1.6}}$	$89.56_{2.5}$
Avg.	80.01	80.75	80.33	**81.65**	70.31	69.89	71.05	**75.28**

Table 17: Ablation of different spatial enhancement strategies using macro-AUC (%). Results are reported as mean_{std} over five random seeds. Fine- and coarse-scale MDPE denote spatial enhancement with relatively small and large receptive fields, respectively. The best result within each evaluation setting is highlighted in bold.

Dataset	Internal				External			
	No MDPE	Fine-Scale	Coarse-Scale	Multi-Scale	No MDPE	Fine-Scale	Coarse-Scale	Multi-Scale
BRCA-ER	$91.83_{(1.7)}$	$89.44_{(3.8)}$	$92.33_{(0.4)}$	$\mathbf{92.61}_{(1.0)}$	$74.07_{(5.2)}$	$78.21_{(2.5)}$	$80.72_{(1.8)}$	$\mathbf{82.03}_{(0.9)}$
BRCA-PAM50	$92.35_{(1.8)}$	$93.88_{(0.9)}$	$\mathbf{94.21}_{(1.4)}$	$93.29_{(1.1)}$	$69.35_{(3.8)}$	$69.45_{(4.8)}$	$69.83_{(1.7)}$	$\mathbf{70.83}_{(1.2)}$
BRCA-PR	$85.82_{(8.5)}$	$\mathbf{87.83}_{(4.3)}$	$84.46_{(4.5)}$	$85.58_{(6.9)}$	$67.30_{(2.7)}$	$72.44_{(1.4)}$	$68.61_{(5.2)}$	$\mathbf{73.78}_{(1.1)}$
MSI	$59.63_{(13.1)}$	$\mathbf{66.89}_{(6.7)}$	$61.72_{(9.7)}$	$64.46_{(9.8)}$	$\mathbf{59.56}_{(16.8)}$	$59.32_{(13.0)}$	$54.09_{(11.0)}$	$59.18_{(20.0)}$
NSCLC	$97.40_{(0.4)}$	$96.93_{(1.0)}$	$97.30_{(0.5)}$	$\mathbf{97.45}_{(0.8)}$	$95.10_{(2.0)}$	$96.70_{(1.3)}$	$\mathbf{97.03}_{(0.9)}$	$96.68_{(2.2)}$
Avg.	85.41	**86.99**	86.00	86.68	73.08	75.22	74.06	**76.50**

Table 18: Ablation of different spatial enhancement strategies using macro-F1 (%). Results are reported as mean_{std} over five random seeds. Fine- and coarse-scale MDPE denote spatial enhancement with relatively small and large receptive fields, respectively. The best result within each evaluation setting is highlighted in bold.

Dataset	Internal				External			
	No MDPE	Fine-Scale	Coarse-Scale	Multi-Scale	No MDPE	Fine-Scale	Coarse-Scale	Multi-Scale
BRCA-ER	$83.32_{(3.3)}$	$82.74_{(3.4)}$	$\mathbf{84.64}_{(1.3)}$	$81.51_{(3.4)}$	$63.93_{(3.9)}$	$66.41_{(3.3)}$	$67.85_{(3.2)}$	$\mathbf{70.98}_{(0.7)}$
BRCA-PAM50	$69.17_{(5.0)}$	$73.65_{(4.1)}$	$69.05_{(7.5)}$	$\mathbf{73.86}_{(5.5)}$	$37.39_{(2.2)}$	$34.29_{(8.5)}$	$38.51_{(2.4)}$	$\mathbf{41.38}_{(1.6)}$
BRCA-PR	$73.76_{(10.8)}$	$\mathbf{77.08}_{(3.3)}$	$74.04_{(2.4)}$	$76.56_{(5.3)}$	$54.67_{(11.6)}$	$60.86_{(7.1)}$	$51.86_{(17.8)}$	$\mathbf{61.39}_{(9.7)}$
MSI	$41.11_{(0.0)}$	$47.09_{(8.6)}$	$46.12_{(6.9)}$	$\mathbf{50.74}_{(13.3)}$	$43.19_{(0.0)}$	$\mathbf{46.83}_{(6.5)}$	$44.97_{(2.5)}$	$44.72_{(2.5)}$
NSCLC	$91.59_{(2.0)}$	$91.89_{(0.9)}$	$91.08_{(2.8)}$	$\mathbf{91.91}_{(1.5)}$	$87.55_{(3.9)}$	$89.95_{(3.9)}$	$\mathbf{90.47}_{(1.6)}$	$89.55_{(2.5)}$
Avg.	71.79	74.49	72.99	**74.92**	57.35	59.67	58.73	**61.60**

Table 19: Ablation of different spatial enhancement strategies using balanced accuracy (BAcc, %). Results are reported as mean_{std} over five random seeds. Fine- and coarse-scale MDPE denote spatial enhancement with relatively small and large receptive fields, respectively. The best result within each evaluation setting is highlighted in bold.

Dataset	Internal				External			
	No MDPE	Fine-Scale	Coarse-Scale	Multi-Scale	No MDPE	Fine-Scale	Coarse-Scale	Multi-Scale
BRCA-ER	$83.09_{(3.7)}$	$82.40_{(3.6)}$	$\mathbf{84.73}_{(2.0)}$	$80.79_{(4.4)}$	$67.29_{(4.8)}$	$71.27_{(3.6)}$	$71.46_{(3.1)}$	$\mathbf{71.69}_{(1.8)}$
BRCA-PAM50	$66.16_{(5.7)}$	$71.01_{(5.0)}$	$65.80_{(7.0)}$	$\mathbf{71.48}_{(6.3)}$	$43.77_{(1.9)}$	$41.67_{(6.2)}$	$42.93_{(2.7)}$	$\mathbf{45.80}_{(1.6)}$
BRCA-PR	$74.02_{(10.8)}$	$77.11_{(3.0)}$	$74.62_{(1.7)}$	$\mathbf{77.22}_{(4.4)}$	$60.05_{(5.3)}$	$\mathbf{65.73}_{(2.8)}$	$59.89_{(7.7)}$	$64.08_{(4.0)}$
MSI	$50.00_{(0.0)}$	$52.55_{(4.2)}$	$52.31_{(3.2)}$	$\mathbf{55.62}_{(7.8)}$	$50.00_{(0.0)}$	$\mathbf{51.74}_{(3.4)}$	$50.76_{(1.1)}$	$50.75_{(1.2)}$
NSCLC	$91.54_{(2.0)}$	$91.84_{(0.9)}$	$91.05_{(2.7)}$	$\mathbf{91.88}_{(1.5)}$	$87.66_{(3.8)}$	$90.10_{(3.7)}$	$\mathbf{90.52}_{(1.6)}$	$89.69_{(2.6)}$
Avg.	72.96	74.98	73.70	**75.40**	61.75	64.10	63.11	**64.40**

Table 20: Ablation of different numbers of heads using accuracy (Acc, %). Results are reported as mean_{std} over five random seeds. The best result within each evaluation setting is highlighted in bold.

Dataset	Internal					External				
	$H=2$	$H=4$	$H=8$	$H=16$	$H=32$	$H=2$	$H=4$	$H=8$	$H=16$	$H=32$
BRCA-ER	$83.68_{1.8}$	$82.11_{2.0}$	$83.42_{2.4}$	$\mathbf{85.53_{3.1}}$	$81.58_{4.6}$	$70.72_{7.8}$	$72.80_{6.1}$	$\mathbf{79.48_{1.1}}$	$68.19_{10.4}$	$70.69_{13.3}$
BRCA-PAM50	$\mathbf{82.89_{1.6}}$	$80.00_{3.0}$	$\mathbf{82.89_{2.6}}$	$\mathbf{82.89_{4.3}}$	$77.89_{6.0}$	$57.98_{4.3}$	$55.02_{5.0}$	$\mathbf{59.63_{2.4}}$	$53.77_{5.0}$	$53.31_{2.6}$
BRCA-PR	$66.93_{11.8}$	$\mathbf{80.27_{4.0}}$	$76.80_{5.4}$	$72.27_{4.4}$	$76.80_{8.1}$	$57.41_{14.5}$	$\mathbf{66.45_{6.3}}$	$71.36_{1.8}$	$60.80_{15.3}$	$63.64_{5.8}$
MSI	$68.68_{4.1}$	$72.45_{3.2}$	$\mathbf{73.21_{4.7}}$	$72.08_{2.5}$	$70.19_{0.8}$	$69.14_{15.1}$	$76.11_{0.8}$	$76.38_{0.6}$	$\mathbf{76.83_{1.6}}$	$76.38_{0.6}$
NSCLC	$92.40_{0.6}$	$\mathbf{92.50_{1.4}}$	$91.92_{1.5}$	$91.63_{0.5}$	$91.06_{0.9}$	$\mathbf{89.87_{1.7}}$	$87.46_{4.8}$	$89.56_{2.5}$	$89.43_{4.3}$	$86.06_{5.1}$
Avg.	78.92	81.47	**81.65**	80.88	79.50	69.02	71.57	**75.28**	69.80	70.02

Table 21: Ablation of different numbers of heads using macro-AUC (%). Results are reported as mean_{std} over five random seeds. The best result within each evaluation setting is highlighted in bold.

Dataset	Internal					External				
	$H=2$	$H=4$	$H=8$	$H=16$	$H=32$	$H=2$	$H=4$	$H=8$	$H=16$	$H=32$
BRCA-ER	$91.56_{0.6}$	$92.00_{0.8}$	$\mathbf{92.61_{1.0}}$	$91.74_{1.9}$	$91.01_{1.4}$	$78.17_{2.2}$	$76.51_{4.4}$	$\mathbf{82.03_{0.9}}$	$77.97_{3.2}$	$77.61_{2.8}$
BRCA-PAM50	$\mathbf{94.83_{1.6}}$	$93.98_{1.0}$	$93.29_{1.1}$	$93.49_{0.8}$	$92.62_{1.9}$	$\mathbf{73.07_{1.8}}$	$70.96_{3.8}$	$70.83_{1.2}$	$70.59_{2.0}$	$71.20_{1.5}$
BRCA-PR	$79.56_{9.1}$	$\mathbf{90.00_{4.8}}$	$85.58_{6.9}$	$81.25_{3.0}$	$86.63_{8.4}$	$65.40_{7.8}$	$69.77_{4.5}$	$\mathbf{73.78_{1.1}}$	$67.61_{5.3}$	$69.37_{1.3}$
MSI	$63.55_{5.3}$	$\mathbf{67.06_{4.4}}$	$64.46_{9.8}$	$55.00_{11.0}$	$59.70_{9.1}$	$\mathbf{68.47_{8.5}}$	$59.69_{7.3}$	$59.18_{20.0}$	$59.06_{12.2}$	$67.07_{8.6}$
NSCLC	$\mathbf{97.65_{0.5}}$	$97.21_{0.4}$	$97.45_{0.8}$	$97.29_{0.6}$	$96.68_{0.6}$	$96.90_{0.9}$	$\mathbf{96.99_{1.6}}$	$96.68_{2.2}$	$96.86_{1.6}$	$95.36_{2.8}$
Avg.	85.43	**88.05**	86.68	83.75	85.33	76.40	74.78	**76.50**	74.42	76.12

Table 22: Ablation of different numbers of heads using macro-F1 (%). Results are reported as mean_{std} over five random seeds. The best result within each evaluation setting is highlighted in bold.

Dataset	Internal					External				
	$H=2$	$H=4$	$H=8$	$H=16$	$H=32$	$H=2$	$H=4$	$H=8$	$H=16$	$H=32$
BRCA-ER	$82.48_{2.2}$	$80.27_{2.7}$	$81.51_{3.4}$	$\mathbf{84.32_{3.7}}$	$79.48_{6.6}$	$65.55_{5.5}$	$65.22_{3.9}$	$\mathbf{70.98_{0.7}}$	$62.81_{7.9}$	$63.62_{9.5}$
BRCA-PAM50	$73.57_{3.2}$	$71.37_{3.0}$	$73.86_{5.5}$	$\mathbf{74.63_{6.9}}$	$67.30_{9.9}$	$\mathbf{42.35_{2.6}}$	$38.55_{3.1}$	$41.38_{1.6}$	$38.95_{1.6}$	$39.12_{1.7}$
BRCA-PR	$65.25_{13.0}$	$\mathbf{80.10_{4.1}}$	$76.56_{5.3}$	$71.83_{4.5}$	$75.99_{9.5}$	$48.27_{13.5}$	$\mathbf{61.52_{4.5}}$	$61.39_{9.7}$	$52.86_{13.8}$	$60.67_{3.5}$
MSI	$51.86_{9.9}$	$\mathbf{55.43_{11.2}}$	$50.74_{13.3}$	$50.67_{10.5}$	$42.38_{2.8}$	$45.35_{5.5}$	$44.99_{2.0}$	$44.72_{2.5}$	$\mathbf{46.38_{6.1}}$	$44.72_{2.5}$
NSCLC	$92.39_{0.6}$	$\mathbf{92.48_{1.4}}$	$91.91_{1.5}$	$91.62_{0.6}$	$91.02_{0.9}$	$\mathbf{89.85_{1.7}}$	$87.29_{5.0}$	$89.55_{2.5}$	$89.37_{4.4}$	$85.93_{5.3}$
Avg.	73.11	**75.93**	74.92	74.61	71.23	58.27	59.51	**61.60**	58.07	58.81

Table 23: Ablation of different numbers of heads using balanced accuracy (BAcc, %). Results are reported as mean_{std} over five random seeds. The best result within each evaluation setting is highlighted in bold.

Dataset	Internal					External				
	$H=2$	$H=4$	$H=8$	$H=16$	$H=32$	$H=2$	$H=4$	$H=8$	$H=16$	$H=32$
BRCA-ER	$82.32_{2.8}$	$79.59_{3.1}$	$80.79_{4.4}$	$\mathbf{83.94_{4.1}}$	$79.16_{6.8}$	$71.27_{2.1}$	$69.02_{4.7}$	$\mathbf{71.69_{1.8}}$	$69.84_{5.9}$	$68.27_{4.9}$
BRCA-PAM50	$71.69_{3.7}$	$69.05_{4.4}$	$71.48_{6.3}$	$\mathbf{73.28_{6.3}}$	$66.87_{10.7}$	$\mathbf{48.02_{1.5}}$	$44.46_{4.3}$	$45.80_{1.6}$	$45.17_{3.2}$	$45.82_{1.7}$
BRCA-PR	$67.09_{11.4}$	$\mathbf{80.14_{4.1}}$	$77.22_{4.4}$	$72.42_{3.4}$	$76.37_{9.2}$	$57.54_{6.9}$	$62.55_{3.9}$	$\mathbf{64.08_{4.0}}$	$59.09_{5.9}$	$63.25_{1.8}$
MSI	$55.93_{5.5}$	$\mathbf{58.28_{7.5}}$	$55.62_{7.8}$	$55.17_{5.8}$	$50.62_{1.4}$	$\mathbf{53.23_{4.8}}$	$50.71_{1.1}$	$50.75_{1.2}$	$51.70_{3.3}$	$50.75_{1.2}$
NSCLC	$92.35_{0.6}$	$\mathbf{92.47_{1.4}}$	$91.88_{1.5}$	$91.60_{0.6}$	$91.00_{0.9}$	$\mathbf{89.92_{1.7}}$	$87.63_{4.4}$	$89.69_{2.6}$	$89.63_{4.1}$	$86.37_{4.9}$
Avg.	73.88	**75.91**	75.40	75.28	72.80	64.00	62.87	**64.40**	63.09	62.89

Table 24: Ablation of different inner-update configurations using accuracy (Acc, %). Results are reported as mean_{std} over five random seeds. The best result within each evaluation setting is highlighted in bold.

Dataset	Internal				External			
	No Update	Local Only	Global Only	Global + Local	No Update	Local Only	Global Only	Global + Local
BRCA-ER	$38.16_{0.0}$	$57.11_{10.6}$	$\mathbf{85.79_{3.7}}$	$83.42_{2.4}$	$22.08_{0.0}$	$66.75_{25.0}$	$74.76_{2.0}$	$\mathbf{79.48_{1.1}}$
BRCA-PAM50	$33.42_{14.1}$	$40.26_{7.9}$	$80.53_{5.0}$	$\mathbf{82.89_{2.6}}$	$32.33_{17.7}$	$38.09_{18.4}$	$52.80_{5.5}$	$\mathbf{59.63_{2.4}}$
BRCA-PR	$47.20_{4.2}$	$49.07_{5.1}$	$73.60_{9.1}$	$\mathbf{76.80_{5.4}}$	$39.13_{16.2}$	$46.38_{19.8}$	$55.67_{13.4}$	$\mathbf{71.36_{1.8}}$
MSI	$69.81_{0.0}$	$61.89_{17.7}$	$71.70_{1.9}$	$\mathbf{73.21_{4.7}}$	$76.02_{0.0}$	$65.61_{23.3}$	$\mathbf{76.83_{1.8}}$	$76.38_{0.6}$
NSCLC	$50.19_{1.1}$	$50.19_{1.1}$	$\mathbf{92.40_{1.1}}$	$91.92_{1.5}$	$50.35_{1.9}$	$50.35_{1.9}$	$88.98_{3.9}$	$\mathbf{89.56_{2.5}}$
Avg.	47.76	51.70	80.80	**81.65**	43.98	53.44	69.81	**75.28**

Table 25: Ablation of different inner-update configurations using macro-AUC (%). Results are reported as mean_{std} over five random seeds. The best result within each evaluation setting is highlighted in bold.

Dataset	Internal				External			
	No Update	Local Only	Global Only	Global + Local	No Update	Local Only	Global Only	Global + Local
BRCA-ER	$56.52_{4.2}$	$53.37_{3.1}$	$92.58_{0.6}$	$\mathbf{92.61_{1.0}}$	$46.10_{2.4}$	$46.33_{2.3}$	$78.10_{2.3}$	$\mathbf{82.03_{0.9}}$
BRCA-PAM50	$51.54_{2.7}$	$50.36_{4.7}$	$92.54_{2.9}$	$\mathbf{93.29_{1.1}}$	$50.00_{1.7}$	$50.01_{1.6}$	$\mathbf{71.91_{1.9}}$	$70.83_{1.2}$
BRCA-PR	$50.98_{4.8}$	$55.85_{4.4}$	$81.56_{11.6}$	$\mathbf{85.58_{6.9}}$	$49.92_{2.5}$	$47.90_{1.9}$	$68.09_{3.1}$	$\mathbf{73.78_{1.1}}$
MSI	$49.07_{1.3}$	$56.10_{3.7}$	$\mathbf{67.30_{3.6}}$	$64.46_{9.8}$	$51.56_{2.1}$	$45.70_{1.0}$	$57.46_{18.1}$	$\mathbf{59.18_{20.0}}$
NSCLC	$49.98_{2.1}$	$52.35_{3.0}$	$97.10_{1.5}$	$\mathbf{97.45_{0.8}}$	$47.62_{1.6}$	$50.34_{4.0}$	$\mathbf{97.30_{0.4}}$	$96.68_{2.2}$
Avg.	51.62	53.61	86.22	**86.68**	49.04	48.06	74.57	**76.50**

Table 26: Ablation of different inner-update configurations using macro-F1 (%). Results are reported as mean_{std} over five random seeds. The best result within each evaluation setting is highlighted in bold.

Dataset	Internal				External			
	No Update	Local Only	Global Only	Global + Local	No Update	Local Only	Global Only	Global + Local
BRCA-ER	$27.62_{0.0}$	$36.09_{4.7}$	$\mathbf{84.62_{4.4}}$	$81.51_{3.4}$	$18.09_{0.0}$	$38.65_{11.5}$	$67.69_{1.7}$	$\mathbf{70.98_{0.7}}$
BRCA-PAM50	$12.17_{4.3}$	$14.26_{2.1}$	$71.29_{6.4}$	$\mathbf{73.86_{5.5}}$	$11.69_{4.9}$	$13.24_{5.2}$	$37.58_{1.9}$	$\mathbf{41.38_{1.6}}$
BRCA-PR	$32.02_{1.9}$	$32.85_{2.3}$	$73.44_{9.2}$	$\mathbf{76.56_{5.3}}$	$27.44_{7.3}$	$30.71_{9.0}$	$53.50_{13.3}$	$\mathbf{61.39_{9.7}}$
MSI	$41.11_{0.0}$	$37.53_{8.0}$	$47.94_{7.1}$	$\mathbf{50.74_{13.3}}$	$43.19_{0.0}$	$38.42_{10.7}$	$\mathbf{46.30_{6.9}}$	$44.72_{2.5}$
NSCLC	$33.42_{0.5}$	$33.42_{0.5}$	$\mathbf{92.38_{1.1}}$	$91.91_{1.5}$	$33.48_{0.9}$	$33.48_{0.9}$	$88.90_{4.0}$	$\mathbf{89.55_{2.5}}$
Avg.	29.27	30.83	73.93	**74.92**	26.78	30.90	58.79	**61.60**

Table 27: Ablation of different inner-update configurations using balanced accuracy (BAcc, %). Results are reported as mean_{std} over five random seeds. The best result within each evaluation setting is highlighted in bold.

Dataset	Internal				External			
	No Update	Local Only	Global Only	Global + Local	No Update	Local Only	Global Only	Global + Local
BRCA-ER	$50.00_{0.0}$	$50.00_{0.0}$	$\mathbf{84.28_{4.9}}$	$80.79_{4.4}$	$50.00_{0.0}$	$50.00_{0.0}$	$70.96_{3.2}$	$\mathbf{71.69_{1.8}}$
BRCA-PAM50	$25.00_{0.0}$	$25.00_{0.0}$	$68.09_{6.3}$	$\mathbf{71.48_{6.3}}$	$25.00_{0.0}$	$25.00_{0.0}$	$45.41_{2.7}$	$\mathbf{45.80_{1.6}}$
BRCA-PR	$50.00_{0.0}$	$50.00_{0.0}$	$73.79_{9.1}$	$\mathbf{77.22_{4.4}}$	$50.00_{0.0}$	$50.00_{0.0}$	$60.87_{6.1}$	$\mathbf{64.08_{4.0}}$
MSI	$50.00_{0.0}$	$50.00_{0.0}$	$53.48_{3.6}$	$\mathbf{55.62_{7.8}}$	$50.00_{0.0}$	$50.00_{0.0}$	$\mathbf{51.70_{3.8}}$	$50.75_{1.2}$
NSCLC	$50.00_{0.0}$	$50.00_{0.0}$	$\mathbf{92.34_{1.1}}$	$91.88_{1.5}$	$50.00_{0.0}$	$50.00_{0.0}$	$89.25_{3.7}$	$\mathbf{89.69_{2.6}}$
Avg.	45.00	45.00	74.40	**75.40**	45.00	45.00	63.64	**64.40**

Table 28: Accuracy (Acc, %) comparison of TTT across different MIL backbones. Results are reported as mean_{std} over five random seeds. The better result within each backbone and evaluation setting is highlighted in bold.

Dataset	Internal				External			
	DGR		MHIM		DGR		MHIM	
	w/o TTT	w/ TTT	w/o TTT	w/ TTT	w/o TTT	w/ TTT	w/o TTT	w/ TTT
BRCA-ER	$80.00_{3.3}$	$\mathbf{83.16_{4.8}}$	$82.89_{4.1}$	$\mathbf{83.42_{2.7}}$	$\mathbf{74.49_{8.3}}$	$69.60_{9.6}$	$71.56_{6.2}$	$\mathbf{73.20_{6.0}}$
BRCA-PAM50	$77.89_{5.1}$	$\mathbf{82.11_{4.4}}$	$79.47_{5.8}$	$\mathbf{83.16_{1.1}}$	$\mathbf{51.56_{9.8}}$	$50.99_{8.6}$	$49.98_{10.0}$	$\mathbf{52.78_{6.3}}$
BRCA-PR	$\mathbf{72.27_{10.3}}$	$\mathbf{72.27_{4.5}}$	$71.20_{5.4}$	$\mathbf{71.73_{7.1}}$	$\mathbf{66.13_{3.6}}$	$63.34_{7.2}$	$\mathbf{67.25_{1.6}}$	$60.37_{7.3}$
MSI	$\mathbf{70.94_{1.7}}$	$70.57_{5.4}$	$69.81_{3.3}$	$\mathbf{70.57_{4.7}}$	$\mathbf{76.02_{0.0}}$	$73.85_{3.1}$	$70.50_{7.9}$	$\mathbf{76.65_{1.4}}$
NSCLC	$90.58_{3.2}$	$\mathbf{92.88_{0.7}}$	$\mathbf{92.31_{0.6}}$	$91.83_{1.2}$	$85.78_{8.7}$	$\mathbf{89.75_{4.5}}$	$\mathbf{90.25_{2.7}}$	$89.97_{2.6}$
Avg.	78.34	**80.20**	79.14	**80.14**	**70.80**	69.51	69.91	**70.59**

Table 29: Macro-F1 (%) comparison of TTT across different MIL backbones. Results are reported as mean_{std} over five random seeds. The better result within each backbone and evaluation setting is highlighted in bold.

Dataset	Internal				External			
	DGR		MHIM		DGR		MHIM	
	w/o TTT	w/ TTT	w/o TTT	w/ TTT	w/o TTT	w/ TTT	w/o TTT	w/ TTT
BRCA-ER	$76.95_{6.2}$	$\mathbf{82.13_{4.7}}$	$\mathbf{82.17_{4.1}}$	$81.65_{3.4}$	$\mathbf{66.45_{4.2}}$	$64.01_{7.0}$	$62.58_{1.8}$	$\mathbf{67.72_{4.9}}$
BRCA-PAM50	$63.78_{10.7}$	$\mathbf{74.52_{6.9}}$	$67.46_{7.8}$	$\mathbf{72.48_{2.7}}$	$37.70_{5.5}$	$\mathbf{40.19_{5.7}}$	$35.97_{5.4}$	$\mathbf{37.08_{3.9}}$
BRCA-PR	$68.30_{18.7}$	$\mathbf{70.71_{6.9}}$	$69.91_{7.5}$	$\mathbf{71.42_{7.1}}$	$\mathbf{58.54_{10.0}}$	$57.29_{8.9}$	$\mathbf{61.65_{3.9}}$	$54.89_{7.0}$
MSI	$46.13_{8.1}$	$\mathbf{53.62_{11.5}}$	$\mathbf{58.11_{7.3}}$	$51.50_{11.7}$	$43.19_{0.0}$	$\mathbf{46.99_{6.9}}$	$\mathbf{53.42_{5.8}}$	$46.36_{7.1}$
NSCLC	$90.50_{3.3}$	$\mathbf{92.86_{0.7}}$	$\mathbf{92.29_{0.6}}$	$91.80_{1.2}$	$85.32_{9.6}$	$\mathbf{89.71_{4.6}}$	$\mathbf{90.24_{2.7}}$	$89.96_{2.6}$
Avg.	69.13	**74.77**	**73.99**	73.77	58.24	**59.64**	**60.77**	59.20

Table 30: Macro-AUC (%) comparison of TTT across different MIL backbones. Results are reported as mean_{std} over five random seeds. The better result within each backbone and evaluation setting is highlighted in bold.

Dataset	Internal				External			
	DGR		MHIM		DGR		MHIM	
	w/o TTT	w/ TTT	w/o TTT	w/ TTT	w/o TTT	w/ TTT	w/o TTT	w/ TTT
BRCA-ER	$90.52_{6.0}$	$\mathbf{92.22_{1.9}}$	$91.47_{2.8}$	$\mathbf{93.10_{1.4}}$	$\mathbf{80.58_{1.1}}$	$78.92_{3.5}$	$73.79_{1.2}$	$\mathbf{78.27_{3.1}}$
BRCA-PAM50	$93.54_{1.2}$	$\mathbf{94.20_{1.2}}$	$92.26_{1.5}$	$\mathbf{94.79_{1.2}}$	$72.17_{2.4}$	$\mathbf{72.43_{2.8}}$	$69.52_{2.5}$	$\mathbf{71.18_{1.4}}$
BRCA-PR	$81.85_{7.0}$	$\mathbf{85.06_{3.2}}$	$83.06_{5.3}$	$\mathbf{85.62_{4.9}}$	$71.26_{1.2}$	$\mathbf{72.87_{1.5}}$	$69.17_{1.3}$	$\mathbf{71.46_{2.7}}$
MSI	$\mathbf{67.23_{8.5}}$	$66.25_{5.9}$	$\mathbf{70.03_{4.8}}$	$67.53_{6.6}$	$61.56_{11.2}$	$\mathbf{66.66_{4.5}}$	$62.61_{7.1}$	$\mathbf{72.11_{6.4}}$
NSCLC	$97.86_{0.3}$	$\mathbf{97.99_{0.3}}$	$\mathbf{97.56_{0.3}}$	$97.46_{0.4}$	$\mathbf{98.20_{0.7}}$	$97.50_{1.3}$	$\mathbf{97.74_{1.0}}$	$97.34_{1.7}$
Avg.	86.20	**87.14**	86.88	**87.70**	76.75	**77.68**	74.57	**78.07**

Table 31: Balanced accuracy (BAcc, %) comparison of TTT across different MIL backbones. Results are reported as mean_{std} over five random seeds. The better result within each backbone and evaluation setting is highlighted in bold.

Dataset	Internal				External			
	DGR		MHIM		DGR		MHIM	
	w/o TTT	w/ TTT	w/o TTT	w/ TTT	w/o TTT	w/ TTT	w/o TTT	w/ TTT
BRCA-ER	$77.09_{6.8}$	$\mathbf{82.55_{3.9}}$	$\mathbf{82.74_{3.7}}$	$80.79_{3.7}$	$69.86_{4.0}$	$\mathbf{70.03_{4.5}}$	$65.25_{2.6}$	$\mathbf{72.66_{3.1}}$
BRCA-PAM50	$61.98_{8.2}$	$\mathbf{73.62_{7.7}}$	$65.78_{5.7}$	$\mathbf{69.69_{3.4}}$	$46.30_{3.2}$	$\mathbf{48.79_{4.2}}$	$43.11_{2.9}$	$\mathbf{43.76_{4.2}}$
BRCA-PR	$71.47_{12.4}$	$\mathbf{71.82_{6.0}}$	$70.65_{6.5}$	$\mathbf{71.69_{7.2}}$	$61.96_{6.6}$	$\mathbf{62.28_{6.0}}$	$\mathbf{62.70_{3.8}}$	$60.73_{5.5}$
MSI	$52.58_{4.3}$	$\mathbf{57.64_{7.0}}$	$\mathbf{59.58_{5.3}}$	$56.57_{7.7}$	$50.00_{0.0}$	$\mathbf{51.80_{4.3}}$	$\mathbf{56.31_{4.7}}$	$51.71_{3.8}$
NSCLC	$90.59_{3.3}$	$\mathbf{92.80_{0.7}}$	$\mathbf{92.26_{0.6}}$	$91.78_{1.2}$	$86.12_{8.3}$	$\mathbf{89.97_{4.3}}$	$\mathbf{90.42_{2.7}}$	$90.19_{2.6}$
Avg.	70.74	**75.69**	**74.20**	74.10	62.85	**64.57**	63.56	**63.81**