A Proof

Proof. Since the bag of instances $\boldsymbol{X}$ is sampled from the probability distribution $\mu(\boldsymbol{x})$, we have the upper bound for Wasserstein distance between $\boldsymbol{X}$ and μ [30],

$$\mathbb{E}[\mathcal{W}_p(\boldsymbol{X}, \mu)] \leq K^{-\frac{1}{d_\mu}}, \tag{6}$$

where K is the number of samples. Next we define the function $\sigma(\boldsymbol{x})$ as a small perturbation function $\sigma(\boldsymbol{x}) = \boldsymbol{x} + \boldsymbol{\delta}$ and let $\tilde{\boldsymbol{X}} = \{\sigma(\boldsymbol{x}) : \boldsymbol{x} \in \boldsymbol{X}\}$. Using the triangle inequality, we have

$$\mathbb{E}[\mathcal{W}_p(\boldsymbol{X}, \tilde{\boldsymbol{X}})] \leq \mathbb{E}[\mathcal{W}_p(\boldsymbol{X}, \mu)] + \mathbb{E}[\mathcal{W}_p(\tilde{\boldsymbol{X}}, \mu)] \leq 2K^{-\frac{1}{d_\mu}} + C, \tag{7}$$

where C is a constant as a result of the perturbation. As the function $S(\cdot)$ is Lipschitz continuous, we have

$$|S(\boldsymbol{X}) - S(\tilde{\boldsymbol{X}})| \leq L \cdot \mathbb{E}[\mathcal{W}_p(\boldsymbol{X}, \tilde{\boldsymbol{X}})] \leq O(L \cdot K^{-\frac{1}{d_\mu}}). \tag{8}$$

Similar to TransMIL [5], let $\Phi : \mathcal{X} \to \mathbb{R}^n$ be any invertible map, where its inverse mapping is expressed as $\Phi^{-1} : \mathbb{R}^d \to \mathcal{X}$. Then we have:

$$S(\Phi^{-1}(\Phi_{\boldsymbol{X} \in \mathcal{X}}(\{\sigma(\boldsymbol{x}) : \boldsymbol{x} \in \boldsymbol{X}\}))) = S(\Phi^{-1}(\Phi_{\tilde{\boldsymbol{X}} \in \mathcal{X}}(\tilde{\boldsymbol{X}}))) = S(\tilde{\boldsymbol{X}}). \tag{9}$$

Let $\gamma = S \circ \Phi^{-1}$. As $|S(\boldsymbol{X}) - S(\tilde{\boldsymbol{X}})| \leq O(L \cdot K^{-\frac{1}{d_\mu}})$, we have

$$|S(\boldsymbol{X}) - \gamma(\Phi_{\boldsymbol{X} \in \mathcal{X}}\{\sigma(\boldsymbol{x}) : \boldsymbol{x} \in \boldsymbol{X}\})| \leq O(L \cdot K^{-\frac{1}{d_\mu}}). \tag{10}$$

□

In this proof, the transformation $\Phi(\cdot) = \mathcal{A}(\cdot)$. This proof could be easily extended to the representations $\boldsymbol{H}$ by assuming a probability measure over the instance representations $\boldsymbol{h}$ and replacing $\boldsymbol{X}$ with $\boldsymbol{H}$. In this case, the transformation $\Phi(\boldsymbol{H}) = \sum_{k=1}^{K} a_k \boldsymbol{h}_k$.

Table 4: Results on general MIL datasets. Experiments were run 5 times and the average classification accuracy (± a standard error of a mean) is reported.

Method	MUSK1	MUSK2	FOX	TIGER	ELEPHANT
Attention	0.892±0.090	0.858±0.106	0.615±0.096	0.839±0.054	0.868±0.054
Attention-Gated	0.900±0.088	0.863±0.094	0.603±0.068	**0.845±0.046**	0.857±0.064
CLAM	0.900±0.136	0.860±0.128	0.610±0.128	0.805±0.052	0.860±0.080
RAM-MIL	**0.911±0.130**	**0.870±0.142**	**0.645±0.117**	0.820±0.040	**0.879±0.096**

B General MIL dataset

Table 4 presents the performance of RAM-MIL on general MIL datasets [31, 32], offering a comparison with baseline methods. The results indicate that OT-based retrieval generally enhances the classification performance. The sole exception is observed with the TIGER dataset, where both CLAM and RAM-MIL are outperformed. This discrepancy might be attributed to CLAM, as RAM-MIL uses CLAM as a pretrained model for attention weights and bag representation extraction. Nonetheless, RAM-MIL still improves over its CLAM baseline on TIGER. Note that our primary focus lies on the more challenging WSI datasets, hence our models are not extensively optimized for general datasets. The data in these general datasets are typically of lower dimensionality and present less challenging conditions. Therefore, any potential underperformance in these contexts should not detract from the strength of our models in handling the WSI data.

C Experiment Details of WSI Classification

We present the experimental details, ablation studies and analysis step-by-step.

MIL Pre-training. For the backbone MIL model we use the the same parameter setup as CLAM. The model parameters are updated via the Adam optimizer with an L2 weight decay of 1e-5 and a learning rate of 2e-4. Each result is obtained with 10-fold splits of training/validation/testing sets.

Table 5: Ablation study for the percentage of instances used on CAMELYON16 and CAMELYON17.

	In-Domain (CAM16)		Out-of-Domain (CAM17)	
	AUC	Accuracy	AUC	Accuracy
10% attention Retr_I	0.9440±0.037	0.8975±0.052	-	-
10% attention Retr_IO	0.9365±0.052	**0.9200±0.050**	**0.7974±0.054**	0.7433±0.073
10% attention Retr_O	0.9414±0.046	0.8975±0.056	0.7775±0.050	0.7392±0.063
20% attention Retr_I	**0.9451±0.036**	0.8925±0.050	-	-
20% attention Retr_IO	0.9341±0.051	0.8925±0.053	0.7651±0.056	0.7714±0.030
20% attention Retr_O	0.9419±0.048	0.9175±0.051	0.7681±0.058	**0.7795±0.021**

Neighbor Selection. After pre-training the MIL model, we obtain the slide-level feature and the attention scores predicted by the network. As computing the optimal transport distance based on all instances is time-consuming, we approximate the distance with a part of samples in a bag.

$$d_{\text{OT}}(\mu,\nu) = \min_{\boldsymbol{T}\in\mathcal{T}(\boldsymbol{\alpha},\tilde{\boldsymbol{\alpha}})} \sum_{i=1}^{|\boldsymbol{\alpha}|}\sum_{j=1}^{|\tilde{\boldsymbol{\alpha}}|} c(\boldsymbol{h}_i,\tilde{\boldsymbol{h}}_j)T_{ij} + \beta\cdot\sum_{ij} T_{ij}\log T_{ij} \tag{11}$$

$$s.t., \boldsymbol{T}^\top \mathbf{1}_K = \boldsymbol{\alpha}, \boldsymbol{T}\mathbf{1}_{\tilde{K}} = \tilde{\boldsymbol{\alpha}}, \boldsymbol{T}\geq 0$$

where $\boldsymbol{\alpha}$ and $\tilde{\boldsymbol{\alpha}}$ are the new attention vector obtained by selecting top η% from $\boldsymbol{a}$ and $\tilde{\boldsymbol{a}}$. In other words, we approximate a bag with η% of instances with the highest attention values generated by pretrained MIL model. As shown in Table 5, we set $\eta = 10$ and $\eta = 20$ for the ablation study. In this experiment, we use Regularization term of 0.5 and Max number of iterations 1000.

It is observed from Table 5, improving the percentage of data improves the performance on Retr_I and Retr_O. On Retr_IO, the performance is saturated when using only 10% of all patches. Differentiating in-domain and out-of-domain data for retrieval could be easily accomplished by representing a bag by a few amount of instances.

Table 6: Ablation study for different merge functions on CAMELYON16 and CAMELYON17.

	In-Domain (CAM16)		Out-of-Domain (CAM17)	
	AUC	Accuracy	AUC	Accuracy
Merge_{add}(2-feats) Retr_I	0.9409±0.038	0.9000±0.049	-	-
Merge_{add}(2-feats) Retr_IO	0.9341±0.051	0.8925±0.053	0.7651±0.056	0.7714±0.030
Merge_{add}(2-feats) Retr_O	0.9414±0.046	0.8975±0.056	0.7775±0.050	0.7392±0.063
Merge_{add}(3-feats) Retr_I	0.9383±0.050	0.9175±0.051	-	-
Merge_{add}(3-feats) Retr_IO	0.9313±0.044	0.9000±0.045	0.7641±0.059	0.7553±0.043
Merge_{add}(3-feats) Retr_O	0.9391±0.051	0.9175±0.045	0.7644±0.059	0.7754±0.022
Merge_{convex}(2-feats) Retr_I	0.9451±0.036	0.8925±0.050	-	-
Merge_{convex}(2-feats) Retr_IO	0.9365±0.052	0.9200±0.050	0.7974±0.054	0.7433±0.073
Merge_{convex}(2-feats) Retr_O	0.9419±0.048	0.9175±0.051	0.7681±0.058	0.7795±0.021
Merge_{convex}(3-feats) Retr_I	0.9398±0.043	0.8975±0.052	-	-
Merge_{convex}(3-feats) Retr_IO	0.9435±0.038	0.8975±0.052	0.7652±0.052	0.7714±0.030
Merge_{convex}(3-feats) Retr_O	0.9417±0.048	0.9050±0.050	0.7690±0.056	0.7755±0.021

Merge Function. Table 6 presents the results using different merge functions. To generate the bag representations, we employ two merge functions: 1) simple addition, referred to as Merge_{add}; 2) convex combination, referred to as Merge_{convex}. Additionally, '2-feats' and '3-feats' refer to bag representation that are merged with 1 nearest neighbor or 2 nearest neighbors, respectively. For convex combination, '2-feats' uses coefficients of 0.6 and 0.4, while '3-feats' uses coefficients of $0.6, 0.2$ and 0.2, where the greatest coefficient corresponds to the original representation. This experiment is done with $\eta = 10$.

It is derived from Table 6 that using 1 nearest neighbor and convex combination presents the best performance. Using 2 nearest neighbors and addition presents the similar results.

Classification Training. Finally, we train a single logistic regression classifier using the merged representation. The Adam optimizer is used to update the model parameters, with a L2 weight decay of 1e-4 and a learning rate of 2e-4. The models are trained for a minimum of 40 epochs and up to a maximum of 200 epochs if the early stopping criterion is not met. This criterion involves monitoring the validation loss each epoch and if it has not decreased from the previous low for over 15 consecutive epochs, early stopping is used.

D Patch-level results on tumor slides of CAMELYON16.

Table 7: Patch-level results on tumor slides of CAMELYON16.

	P-Prec.(↑)	FROC(↑)
DSMIL	0.1030	0.4443
CLAM	0.6068	0.4792
TransMIL	0.1726	0.4797
Bayes-MIL	**0.8107**	0.4919
RAM-MIL	0.6114	**0.5281**

The Tumor-Precision is calculated by the precision of classifying the tumor patches. The Patch-Precision is calculated by averaging the precision of classifying both normal and tumor patches. The Patch-FROC is defined as the average sensitivity (recall) at 6 predefined false positive rates: 1/4, 1/2, 1, 2, 4 and 8 FPs per WSI.

In Table 7, RAM-MIL presents the second best precision and the best FROC on the patch-level segmentation. This indicates that using transport matrix for interpreting the patch-level classification achieves the best overall performance in the trade-off of false positive rate and recall. By contrast, Bayes-MIL could only obtain a high precision, which reduces the number of false alarm. However, for the application of medical WSI, reducing false negative (better recall and FROC) is supposed to be more important as classifying a positive instance to be negative is unacceptable in the application of medical pregnosis or diagnosis.