CAPRMIL: Context-Aware Patch Representations for Multiple Instance Learning

Andreas Lolos[*1,2] ANDREASLOLOS@PHYS.UOA.GR
Theofilos Christodoulou[*2] TH.CHRISTODOULOU@ATHENARC.GR
Aris L. Moustakas[1,2] ARISLM@PHYS.UOA.GR
Stergios Christodoulidis[3,4] STERGIOS.CHRISTODOULIDIS@CENTRALESUPELEC.FR
Maria Vakalopoulou[2,3,4] MARIA.VAKALOPOULOU@CENTRALESUPELEC.FR
[1] *National and Kapodistrian University of Athens, Greece*
[2] *Archimedes, Athena Research Center, Greece*
[3] *MICS Laboratory, CentraleSupélec, Université Paris-Saclay*
[4] *IHU PRISM, National Center for Precision Medicine in Oncology, Gustave Roussy*

Editors: Under Review for MIDL 2026

Abstract

In computational pathology, weak supervision has become the standard for deep learning due to the gigapixel scale of WSIs and the scarcity of pixel-level annotations, with Multiple Instance Learning (MIL) established as the principal framework for slide-level model training. In this paper, we introduce a novel setting for MIL methods, inspired by proceedings in Neural Partial Differential Equation (PDE) Solvers. Instead of relying on complex attention-based aggregation, we propose an efficient, aggregator-agnostic framework that removes the complexity of correlation learning from the MIL aggregator. CAPRMIL produces rich context-aware patch embeddings that promote effective correlation learning on downstream tasks. By projecting patch features -extracted using a frozen patch encoder- into a small set of global context/morphology-aware tokens and utilizing multi-head self-attention, CAPRMIL injects global context with linear computational complexity with respect to the bag size. Paired with a simple Mean MIL aggregator, CAPRMIL matches state-of-the-art slide-level performance across multiple public pathology benchmarks, while reducing the total number of trainable parameters by $48\% - 92.8\%$ versus SOTA MILs, lowering FLOPs during inference by $52\% - 99\%$, and ranking among the best models on GPU memory efficiency and training time. Our results indicate that learning rich, context-aware instance representations before aggregation is an effective and scalable alternative to complex pooling for whole-slide analysis. Our code is available at: `https://github.com/mandlos/CAPRMIL`

Keywords: Digital Pathology, Multiple Instance Learning, Context Aware Representations

1. Introduction

Whole Slide Image (WSI) analysis has become the foundation of clinical practice in computational pathology (Alkhalaf et al., 2024; Wang et al., 2024), however their sheer size poses a significant challenge for Deep Learning approaches (Brixtel et al., 2022; Lu et al.,

[*] Contributed equally

2021b; Gadermayr and Tschuchnig, 2024). At the same time, pixel-level annotations are prohibitively expensive and time-consuming, resulting in clinical datasets typically providing only slide-level labels rather than fine-grained annotations. (Lu et al., 2021b; Song et al., 2023; Gadermayr and Tschuchnig, 2024).

To address the computationally prohibitive size of WSIs and the lack of pixel-level annotations, Multiple Instance Learning (MIL) has been established as the standard framework for WSI analysis. The MIL pipeline comprises patch feature extraction, typically adopting pre-trained foundation models (Xiong et al., 2025), followed by aggregation/pooling to produce the slide-level representation for downstream tasks. In recent years, attention-based mechanisms have emerged as a promising approach for a trainable MIL aggregator (Ilse et al., 2018; Wang et al., 2024; Gadermayr and Tschuchnig, 2024), due to their impressive correlation learning capabilities. While effective, approaches that utilize standard attention directly on the patch embeddings face computational bottlenecks due to the quadratic complexity of the attention operator (Shao et al., 2021). Attention-based MIL methods for WSI have also been found to be highly susceptible to overfitting and offer limited interpretability (Zhang et al., 2025b), while often lacking principled uncertainty quantification (Sun et al., 2026; Cui et al., 2022; Lolos et al., 2025), limiting the potential of clinical translation. Therefore, developing aggregation strategies that can effectively model instance interactions, handle the challenges inherent to long sequence processing in WSIs, and provide reliable representations remains an active area of research (Bilal et al., 2023; Fang et al., 2024).

At the same time, we identify that neural Partial Differential Equation (PDE) Solvers (Li et al., 2020; Hao et al., 2023; Wu et al., 2024) face a similar challenge: how to achieve efficient and reliable correlation learning in large-scale inputs. Solving PDEs often includes modeling complex phenomena that may cause long-distance interactions, on domains discretized into millions of mesh points (Grossmann et al., 2024). Attention-based methods have been used in PDE modeling, but they also face prohibitive computational cost and degraded correlation learning due to the large scale of the input (Katharopoulos et al., 2020; Wu et al., 2024). Therefore, we assume that ideas that have successfully tackled these problems in the domain of Surrogate PDE Solvers could provide new insights in digital pathology.

In this work, we introduce CAPRMIL , a novel and efficient attention-based MIL framework for WSI analysis, proposing a paradigm shift by removing the complexity of correlation learning from the MIL aggregator, using context-aware patch representations. Following the architecture of Transolver (Wu et al., 2024; Luo et al., 2025), which shows promising results in efficient PDE modeling, we leverage Multi-Head Self-Attention (MSA) over a small set of global context-aware tokens, achieving linear computational complexity with respect to the input and promoting effective correlation learning on downstream tasks. More precisely, our main contributions are summarized as follows:

(i) We propose a novel and efficient MIL setting based on the Transolver architecture. Tackling the challenge of the large dimensionality of the input, CAPRMIL introduces a bottleneck before the attention operator, which consists of: (1) soft clustering of the patch embeddings and (2) aggregating each cluster into a context-aware token. By utilizing MSA over the context-aware tokens, CAPRMIL achieves linear computational complexity with respect to the bag size and produces rich morphology/context-aware patch representations.

(ii) A highly parameter-efficient formulation. Our approach performs on par with current state-of-the-art MIL heads, while reducing the total number of trainable parameters by 48% compared to ABMIL and up to 92.8% compared to SOTA transformer-based MILs. This significantly reduces the computational requirements during training and inference in terms of time, FLOPS, and memory utilization.

(iii) A scalable, aggregator-agnostic formulation that can be adapted in multiple MIL heads. Our formulation is independent of the MIL aggregator, and it can be applied in different commonly used MIL settings with small computational overhead.

We challenge CAPRMIL on various publicly available computational pathology datasets. Paired with a simple MeanMIL aggregator, our method matches SOTA performance, while achieving leading efficiency, highlighting a highly efficient and adaptable MIL framework.

2. Related Work

MIL-based frameworks for digital pathology. During the last years, many different MIL settings have been introduced and extensively tested in different settings (Shao et al., 2025). Depending on the mechanism of aggregation that they are using, they can be grouped into different categories. Among the most popular attention-based methods, we can note ABMIL (Ilse et al., 2018), CLAM (Lu et al., 2021b) and DSMIL (Li et al., 2021). Moreover, TransMIL (Shao et al., 2021) was among the first to introduce a transformer network specifically for WSI, in order to model both morphological and spatial correlations. Building on top of this, DGRMIL (Zhu et al., 2025) utilizes a set of learnable ”global vectors” to summarize distinct morphological patterns and computes cross-attention between the instances and these global vectors, effectively achieving linear scaling. Finally, probabilistic-based MIL such as the one introduced at Cui et al. (Cui et al., 2022) argued that standard attention scores are unreliable proxies for interpretability, proposing BayesMIL, where they introduce a probabilistic instance-wise attention module that yields patch-level uncertainty estimates. Similarly, Lolos et al. targeted the lack of uncertainty estimation in deterministic models and introduced SGPMIL (Lolos et al., 2025), a framework to learn a posterior distribution over attention scores.

Attention-based Neural PDE Solvers. Solving Partial Differential Equations (PDEs) is fundamental to modeling complex phenomena in science and engineering. While traditional numerical approaches such as the Finite Element Method (FEM) offer high accuracy, they typically require discretization of the domain into high-resolution meshes -often containing millions of mesh points-, resulting in prohibitive computational costs (Grossmann et al., 2024). Consequently, deep learning-based neural operators have emerged as efficient surrogates, capable of learning the mapping between model state and solution fields directly from data (Li et al., 2020; Lu et al., 2021a; Wu et al., 2024). Transformer architectures have been increasingly utilized in neural PDE solvers due to their ability to model global dependencies (Li et al., 2022). However, they often face computational bottlenecks due to the quadratic complexity of standard self-attention (Katharopoulos et al., 2020; Luo et al., 2025). Furthermore, simply applying attention to individual mesh points may fail to capture the intricate high-order physical correlations governing the system, as the model can become overwhelmed by low-level geometric details, thus preventing effective relation learning (Wu et al., 2022). We identify that challenges inherent to long-sequence processing, such as computational complexity and efficient correlation learning, are common in both large-scale physical simulations and WSI analysis. Surprisingly, the use of neural PDE solvers has not been explored in digital pathology, to the best of our knowledge.

The Transolver Architecture. To address the prohibitive computational cost and degraded correlation learning due to the large size of the input, Wu et al. introduced Transolver (Wu et al., 2024), a ”Transformer-based PDE solver for General Geometries”, which was later scaled by Luo et al. (Luo et al., 2025). Their architecture introduces Physics-Attention, proposing that a domain discretized to N mesh points can be decomposed into a set of $M \ll N$ physically consistent clusters (“slices”), which can then be aggregated into “physics-aware tokens”, forming a compact latent representation of distinct physical states. Standard Multi-Head Self-Attention (MSA) can then be applied to these tokens for correlation modeling with complexity $O(M^2)$, achieving linear scaling with respect to the number of mesh points. By explicitly modeling ”physical states” rather than individual points, the model becomes more robust to geometric variations and discretization artifacts (Wu et al., 2024; Luo et al., 2025), while the learned slices have been shown to correspond to meaningful physical regions, enhancing the model's interpretability and generalization capability (Wu et al., 2024; Luo et al., 2025). Drawing a parallel to digital pathology, both neural PDE solvers and MIL models face the fundamental challenge of efficiently learning correlations over massive sequences of instances (mesh points in PDEs, patches in WSIs). Viewed through this lens, Transolver's Physics-Attention constitutes a promising approach to facilitate efficient global corelation modeling, by projecting the high-dimensional input space onto a compact set of latent variables.

3. Methodology

In this work, we propose CAPRMIL , a novel and efficient MIL framework, designed to overcome the limitations of standard attention in WSI analysis. Unlike prior methods that treat patches as isolated units, CAPRMIL projects patch embeddings into morphology units via soft clustering, and aggregates them into a compact set of context-aware, low-dimensional global tokens, over which self-attention is performed. Global contextual information is then propagated back to the patch embeddings via context broadcasting. By attending to the tokens rather than patch embeddings, CAPRMIL achieves linear scaling with respect to the bag size, while maintaining strong representational capacity and high parameter efficiency.

3.1. Model Architecture

The CAPRMIL framework for WSI consists of three sequential stages (Figure 1): (1) an initial projection of WSI patches into patch embeddings using a pre-trained encoder as frozen backbone, (2) a stack of CAPRMIL Blocks that use multi-head self-attention over global token representations to produce context/morphology-aware patch embeddings, and (3) a final MIL aggregation and classification head to produce the slide-level prediction.

3.1.1. Feature Projection

A WSI is represented as a bag $X \in \mathbb{R}^{B \times N \times D_{in}}$ of N patch embeddings of dimension D_{in}, with batch size B. These embeddings are projected into a latent space of dimension $D \ll$

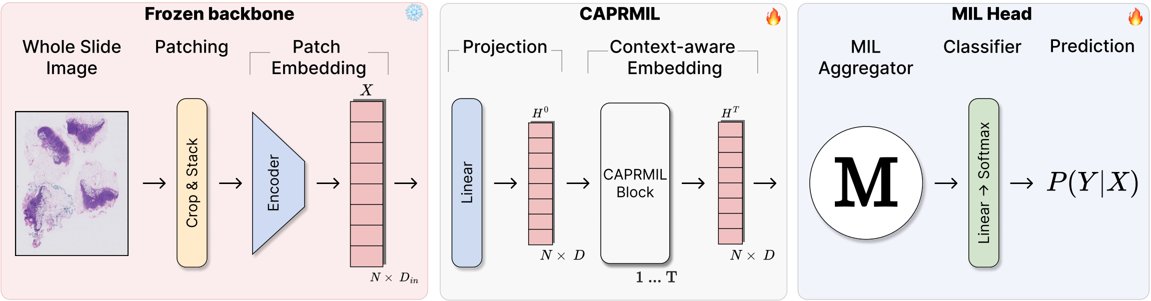


Figure 1: CAPRMIL Framework. The WSI is tessellated into patches which are encoded into patch embeddings using a frozen backbone. After a linear projection, T consecutive CAPRMIL Blocks augment global context to yield context-aware patch embeddings. A MIL aggregator and classifier then produce the final slide-level prediction.

D_{in} via a learnable linear layer followed by Layer Normalization (LN), GELU activation, and Dropout, yielding patch representations $\mathbf{H}^{(0)}$ as input to the first CAPRMIL Block:

$$\mathbf{H}^{(0)} = \text{Dropout}(\text{GELU}(\text{LN}(\text{Linear}(\mathbf{X})))) \in \mathbb{R}^{B \times N \times D}$$

3.1.2. The CAPRMIL Block

To capture high-order correlations without the quadratic cost of standard self-attention, the CAPRMIL Block adopts the Transolver architecture, performing attention over low-dimensional, context-aware global tokens to achieve linear complexity with respect to the bag size. As illustrated in Figure 2a, it follows a Transformer encoder-style design with H CAPRMIL heads and shared projection matrices across heads, augmenting patch embeddings with global context to produce rich, morphology-aware representations, formulated as:

$$\mathbf{H}' = \mathbf{H}^{(l-1)} + \text{Dropout}(\text{CAPRMIL Attention}(\text{LN}(\mathbf{H}^{(l-1)})))$$

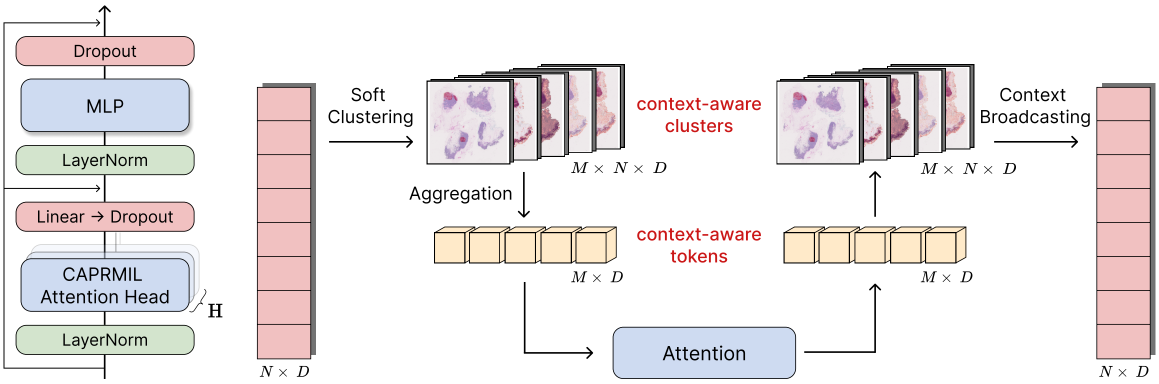
$$\mathbf{H}^{(l)} = \mathbf{H}' + \text{Dropout}(\text{MLP}(\text{LN}(\mathbf{H}')))$$

for $l \in [1, T]$, for T consecutive CAPRMIL Blocks. The MLP comprises two linear layers with GELU activation. CAPRMIL attention returns the concatenated output of all heads.

3.1.3. CAPRMIL Attention head

CAPRMIL adopts the Physics-Attention mechanism from Transolver to enable efficient correlation learning on large-scale inputs. As illustrated in Figure 2b, it operates in four stages: (1) soft clustering of patch representations into morphology-aware clusters, (2) aggregation into morphology-aware tokens, (3) self-attention over these tokens, and (4) broadcasting the transited tokens back to the input space, producing context-aware patch representations.

Soft Clustering. To achieve linear scaling with respect to the bag size, CAPRMIL replaces attention over patches with attention over a compact set of morphology-aware tokens, produced via soft clustering followed by aggregation. Let $\mathbf{H} \in \mathbb{R}^{B \times N \times D}$ denote the input patch representations. $\mathbf{H}$ is mapped via two learnable projections into



(a) CAPRMIL Block

(b) CAPRMIL Attention head

Figure 2: (a) The CAPRMIL Block follows a transformer encoder architecture with multihead self-attention. Each CAPRMIL Block contains H CAPRMIL Attention heads and their output is concatenated. Skip connections are implemented after every Dropout. (b) The CAPRMIL Attention head projects patch embeddings into M clusters, via soft clustering. Each cluster is aggregated into a context-aware token and attention is applied to the set of M tokens. The transited tokens are projected back to the input latent space via context broadcasting.

$\mathbf{x}, \mathbf{f} \in \mathbb{R}^{B\times N\times(HD_{\text{head}})}$, $D_{\text{head}} = D/H$, by linear layers $\mathbf{W}_x, \mathbf{W}_f \in \mathbb{R}^{D\times(HD_{\text{head}})}$ and reshaped into $\tilde{\mathbf{x}}, \tilde{\mathbf{f}} \in \mathbb{R}^{B\times H\times N\times D_{\text{head}}}$. All N patch embeddings are then softly assigned to M context-aware clusters per head. A learnable projection $\mathbf{W}_{\text{cluster}} \in \mathbb{R}^{D_{\text{head}}\times M}$, initialized orthogonally, produces:

$$\mathbf{A}_{\text{logits}} = \frac{\tilde{\mathbf{x}}\mathbf{W}_{\text{cluster}}}{\tau}, \quad \mathbf{W} = \text{Softmax}(\mathbf{A}_{\text{logits}}, \text{dim} = -1), \quad \sum_{m=1}^{M} W_{b,h,n,m} = 1 \ \forall\, b, h, n,$$

where $\mathbf{A}_{\text{logits}}, \mathbf{W} \in \mathbb{R}^{B\times H\times N\times M}$, $\mathbf{W}$ is the soft assignment matrix, corresponding to the probability distribution of each embedding being mapped to each cluster. M controls the number of clusters, and $\tau \in \mathbb{R}^{H}$ is a temperature parameter we introduce as learnable, controlling the assignment entropy of each attention head. Each cluster is then aggregated into a token, corresponding to a weighted combination of input embeddings, calculated as:

$$\mathbf{S}_{b,h,m} = \frac{\sum_{n=1}^{N} W_{b,h,n,m}\, \tilde{\mathbf{f}}_{b,h,n}}{\sum_{n=1}^{N} W_{b,h,n,m} + \varepsilon}, \quad \text{with } \mathbf{S} \in \mathbb{R}^{B\times H\times M\times D_{\text{head}}}$$

Self-Attention. Given M morphology-aware tokens $\mathbf{S}$ per head, we apply Multi-Head Self-Attention. Query ($\mathbf{Q}$), Key ($\mathbf{K}$), and Value ($\mathbf{V}$) are obtained via shared linear projections of the head-wise token embeddings $\mathbf{S}$. Attention is then given by:

$$\text{Attn} = \text{Softmax}\left(\frac{\mathbf{QK}^{\top}}{\sqrt{D_{\text{head}}}}\right), \quad \mathbf{S}' = \text{Dropout}(\text{Attn} \cdot \mathbf{V})$$

with $\mathbf{S}' \in \mathbb{R}^{B\times H\times M\times D_{head}}$. Since ($M \ll N$), applying the attention operator over the context-aware tokens -instead of the N patch embeddings- reduces computational complexity and allows the model to scale linearly with the input. At the same time, since tokens aggregate global context, CAPRMIL learns meaningful correlations, beyond spatial features.

Context Broadcasting. The updated tokens $\mathbf{S}'$ are broadcast back to the input latent space using the same assignment weights $\mathbf{W}$ from the soft clustering step, reconstructing each patch representation as a weighted combination of transited tokens:

$$\mathbf{O}_{b,h,n,d} = \sum_{m=1}^{M} \mathbf{S}'_{b,h,m,d} \mathbf{W}_{b,h,n,m}, \quad \mathbf{O} \in \mathbb{R}^{B \times H \times N \times D_{\text{head}}}.$$

Head-wise representations are concatenated into $\mathbf{H}^{(T)} \in \mathbb{R}^{B \times N \times (HD_{head})}$ and linearly projected to the model dimension, yielding the final context-aware patch representations.

3.1.4. Aggregation and Prediction

After T CAPRMIL Blocks, the patch representations $\mathbf{H}^{(T)}$ are mean-pooled to form a slide-level embedding $\mathbf{z} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{H}_n^{(T)} \in \mathbb{R}^{B \times D}$. A final linear classifier projects $\mathbf{z}$ to the target class logits depending on the task.

3.2. Computational Efficiency

CAPRMIL addresses the ”curse of dimensionality” by decoupling the sequence length N from the attention mechanism. Since the attention operator displays quadratic computational complexity, attending to all N patch embeddings would yield $O(N^2)$ complexity. CAPRMIL Attention instead attends to the M context-aware tokens, achieving an overall complexity of $O(MND + M^2D)$. Given that the number of prototypes M is a constant with $M \ll N$, the model achieves linear computational complexity with respect to the input size N, making it ideal to model long sequences.

4. Experiments & Results

Datasets, Tasks and Evaluation Metrics. We evaluate our approach on four WSI benchmarks: **CAMELYON16** (Ehteshami Bejnordi et al., 2017) for tumor detection, **TCGA-NSCLC** (Cooper et al., 2018; Campbell et al., 2016) for lung cancer subtyping, **BRACS** (Brancati et al., 2022) for coarse breast lesion classification, and **PANDA** (Bulten et al., 2022) for prostate ISUP grading. Dataset-specific evaluation protocols, metrics, and implementation details are provided in the Appendix. We report slide-level classification performance and calibration using area under the curve (**AUC**) and adaptive expected calibration error (**ACE**) (Nixon et al., 2019).

4.1. Slide-level Performance and Parameter Efficiency

CAPRMIL achieves competitive AUC and calibration relative to state-of-the-art MIL methods (Table 1), with performance differences consistently within one standard deviation, while being substantially more parameter efficient. Notably, these results are obtained using simple mean aggregation, underlining the strength of the learned context-aware patch representations. Across datasets, CAPRMIL matches parameter-efficient methods such as ABMIL and CLAM on CAMELYON16 and TCGA-NSCLC, and ranks among the top-performing approaches on multiclass tasks including PANDA and BRACS. At the same

	CAMELYON16		TCGA-NSCLC		PANDA		BRACS		Params	FLOPs
	AUC	ACE	AUC	ACE	κ	ACE	AUC	ACE	(M)	(G)
ABMIL (Ilse et al., 2018)	$\mathbf{.987_{.005}}$	$.036_{.004}$	$.973_{.009}$	$.039_{.008}$	$.910_{.028}$	$.044_{.015}$	$\underline{.852_{.025}}$	$\underline{.175_{.007}}$	.660	1.31
CLAM (Lu et al., 2021b)	$\underline{.986_{.004}}$	$.044_{.027}$	$.953_{.004}$	$.056_{.016}$	$.927_{.025}$	$.031_{.018}$	$.850_{.021}$	$.183_{.011}$	.920	1.84
TransMIL (Shao et al., 2021)	$.978_{.004}$	$.044_{.012}$	$.970_{.012}$	$.046_{.019}$	$.911_{.030}$	$.043_{.021}$	$.826_{.032}$	$.186_{.012}$	2.67	85.02
DGRMIL (Zhu et al., 2025)	$.967_{.018}$	$.027_{.021}$	$.974_{.011}$	$\underline{.038_{.022}}$	$.933_{.047}$	$.036_{.025}$	$.818_{.035}$	$.186_{.023}$	4.34	79.88
BayesMIL (Cui et al., 2023)	$.975_{.006}$	$\mathbf{.023_{.006}}$	$.973_{.021}$	$\mathbf{.033_{.017}}$	$.926_{.031}$	$.031_{.016}$	$.829_{.022}$	$.183_{.028}$	1.32	2.63
SGPMIL (Lolos et al., 2025)	$\mathbf{.987_{.008}}$	$\underline{.026_{.009}}$	$.973_{.014}$	$.047_{.027}$	$\mathbf{.955_{.037}}$	$\underline{.028_{.022}}$	$\mathbf{.870_{.026}}$	$\mathbf{.142_{.032}}$	1.21	2.44
Mean	$.693_{.046}$	$.241_{.022}$	$\mathbf{.979_{.015}}$	$.041_{.019}$	$.924_{.028}$	$.035_{.013}$	$.738_{.006}$	$.223_{.015}$	**.130**	**.260**
CAPRMIL +Mean	$.975_{.006}$	$.028_{.006}$	$\underline{.978_{.016}}$	$\mathbf{.033_{.021}}$	$\underline{.944_{.053}}$	$\mathbf{.021_{.024}}$	$.850_{.031}$	$.189_{.026}$	<u>.314</u>	<u>.628</u>

Table 1: Slide-level performance comparison across datasets. Results are reported as AUC/κ, and ACE. FLOPs are measured per forward pass for a bag of 1000 patch embeddings at inference. Using a mean operator in the initial projection layer before classification (i.e., our approach without the block) leads to substantial performance degradation for large bag sizes, such as CAMELYON16 and BRACS.

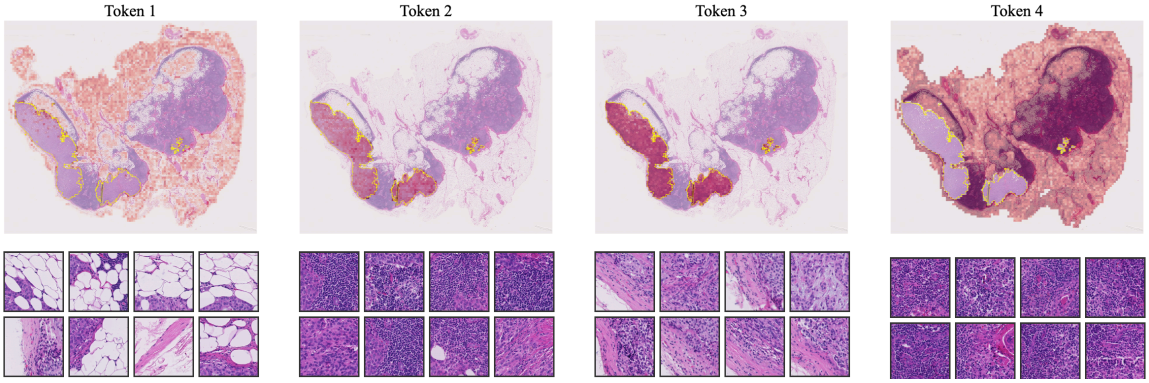


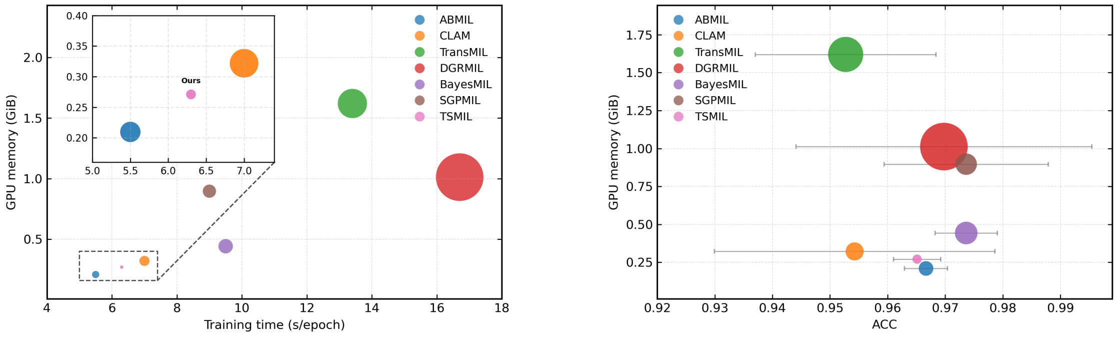
Figure 3: **Token–patch assignment heatmaps.** Test slide from CAMELYON16. *Top:* Soft assignment weights from one CAPRMIL attention head, indicating each patch's contribution to the M context-aware tokens. *Bottom:* Top-8 patches per token ranked by assignment score, highlighting dominant morphological patterns for each token.

time, CAPRMIL reduces trainable parameters by approximately 48% relative to ABMIL, and by up to 88% and 92.8% compared to transformer-based methods such as TransMIL and DGRMIL, respectively. In terms of efficiency, CAPRMIL reduces FLOPS during inference by 52% to over 99% compared to ABMIL, TransMIL, and DGRMIL.

In contrast, naive mean aggregation -using a linear projection, mean pooling, and a linear classifier- degrades substantially on large-bag tasks. The Mean baseline underperforms by 44% on CAMELYON16 and 35.5% on BRACS, where bags contain 4k–20k instances. On PANDA (average bag size ~500), mean pooling performs comparably to other methods, with a similar trend on TCGA-NSCLC. Overall, these results indicate that context-aware tokenization is critical for maintaining discriminative capacity while enabling a parameter- and computation-efficient formulation.

As seen in Figures 3 and A 5 - 7, we observe that tokens tend to aggregate patches with visually coherent histological patterns, such as adipose-rich or epithelial-dominant regions, while de-emphasizing unrelated tissue types. The top-k assigned patches per token indicate that a limited subset of instances dominates each token's construction. Specifically in Figure 3, Token 1 predominantly captures adipose-rich regions, as confirmed by their low cellular content in the top-8 assigned patches. Tokens 2 and 3 focus on tumor-related tissue, with Token 2 aggregating malignant epithelial regions, while Token 3 captures stromal or tumor-associated connective tissue. Finally, Token 4 primarily represents benign tissue, with patches exhibiting more homogeneous cellular organization.

4.2. Computational and Memory Efficiency



(a) (b)

Figure 4: **Model efficiency analysis**. (a) GPU memory footprint (peak during training, averaged over 30 epochs) vs. training time (entire training set, averaged over 30 epochs). (b) GPU memory footprint vs. ACC. Marker size denotes the number of trainable parameters.

While parameter count and FLOPs provide useful proxies for model efficiency, practical deployment at whole-slide scale additionally depends on empirical resource utilization. Figure 4 analyzes this by relating peak GPU memory usage, training time, and slide-level performance across competing MIL methods. As shown in Figure 4(a), CAPRMIL exhibits a substantially lower memory footprint and shorter training time compared to transformer-based approaches, reflecting its linear-scaling design and low-dimensional intermediate representations. CAPRMIL remains competitive with more computationally demanding models despite its resource-efficiency, by operating on rich, context-aware patch embeddings. Figure 4(b) further illustrates the trade-off between accuracy and memory consumption. CAPRMIL achieves high balanced accuracy, with performance differences consistently within one standard deviation of leading competitors , while operating under a significantly smaller GPU memory budget. In contrast, transformer-based methods such as TransMIL and DGRMIL incur large memory overheads for only marginal performance gains. While attention-based and probabilistic MIL methods offer stronger aggregation modules, they do so with increased computational or memory demands, suggesting that CAPRMIL context-aware representations provide a lightweight yet competitive alternative.

4.3. Ablation studies

Clusters and heads. Varying the number of clusters M while fixing $H = 8$ and the MLP ratio to 4 shows stable performance across a wide range of values ($M \in \{2, 4, 8, 16\}$), with no consistent gains from increasing the number of clusters beyond small to moderate values (Table A 3, top). Similarly, increasing the number of attention heads H improves performance from 2 to 8 heads but saturates thereafter, with no clear benefit on larger values (Table A 3, middle). Based on these trends, we adopt $M = 4$ and $H = 8$ as balanced choices that provide sufficient contextual capacity without unnecessary complexity.

MLP expansion ratio. Ablating the MLP expansion ratio with fixed $M = 4$ and $H = 8$ indicates that smaller ratio slightly degrades performance, while larger ratio yields more consistent results across metrics (Table A 3, bottom). We therefore use an MLP ratio of 4 in all experiments. Overall, these ablations indicate that the selected configuration ($M = 4$, $H = 8$, MLP ratio = 4) offers a robust trade-off between representational capacity and efficiency and is not sensitive to precise hyperparameter choices.

	CAMELYON16		TCGA-NSCLC		PANDA		BRACS		Params
	AUC	ACE	AUC	ACE	κ	ACE	AUC	ACE	(M)
CAPRMIL +Mean	$.975_{.006}$	$\underline{.028_{.006}}$	$\mathbf{.978_{.016}}$	$.033_{.021}$	$.944_{.053}$	$.021_{.024}$	$\underline{.850_{.031}}$	$.189_{.026}$	$\mathbf{.314}$
CAPRMIL +Attn	$\mathbf{.977_{.004}}$	$\mathbf{.027_{.003}}$	$\underline{.975_{.015}}$	$\mathbf{.031_{.023}}$	$\underline{.944_{.046}}$	$.023_{.025}$	$.834_{.032}$	$\underline{.180_{.023}}$	$\underline{.331}$
CAPRMIL +GAttn	$\underline{.976_{.009}}$	$.033_{.010}$	$.974_{.018}$	$\underline{.032_{.020}}$	$\mathbf{.952_{.043}}$	$\mathbf{.019_{.022}}$	$\mathbf{.874_{.031}}$	$\mathbf{.171_{.019}}$	$.347$

Table 2: Comparison of aggregation strategies within CAPRMIL . Results are reported as AUC/κ, and ACE; parameter counts include the aggregation module.

Modularity and aggregation robustness. Table 2 compares different MIL aggregation strategies learned together with CAPRMIL representations. In contrast to prior MIL approaches that rely heavily on sophisticated attention pooling, we observe that replacing mean aggregation with attention or gated attention leads to broadly comparable performance across datasets, within one standard deviation. Notably, on more challenging multiclass tasks such as PANDA and BRACS, attention-based aggregators yield a performance increase from 0.8% up to 2.4% respectively, suggesting that additional aggregation capacity may be beneficial in more complex settings. Overall, these results indicate that the CAPRMIL block already encodes most of the relevant contextual and discriminative information at the patch level, rendering the choice of final aggregation largely non-critical for performance. While attention-based aggregators introduce increased parameterization, they do not provide consistent gains across all tasks, highlighting diminishing returns once strong instance representations are learned. These findings underline the modularity of CAPRMIL and demonstrate that competitive performance can be achieved with simple, parameter-efficient aggregation, while still allowing the flexibility to incorporate more expressive MIL heads when task complexity demands it.

5. Conclusions

We present a parameter-efficient and scalable MIL framework that learns context-aware patch representations, substantially reducing reliance on complex aggregation mechanisms. Experimental results show that once rich contextual features are learned, simple pooling performs on par with more elaborate MIL heads, underscoring the robustness and modularity of the proposed approach. A current limitation is the focus on unimodal visual inputs; evaluating scalability and robustness in larger multimodal pipelines remains and interesting direction for future work.

Acknowledgments

This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program. Hardware resources were granted with the support of GRNET.

References

Abdulmohsen Khalaf Ali Alkhalaf, Sulaiman Ali Sulaiman Alkhateeb, and Maha Mohammed Alshammari. Integration of artificial intelligence in histopathological and radiological image analysis: Enhancements in diagnostic workflow. *International journal of health sciences*, 8(S1):938–953, January 2024. ISSN 2550-696X. doi: 10.53730/ijhs.v8nS1.15010. URL https://sciencescholar.us/journal/index.php/ijhs/article/view/15010.

Mohsin Bilal, Robert Jewsbury, Ruoyu Wang, Hammam M. AlGhamdi, Amina Asif, Mark Eastwood, and Nasir Rajpoot. An aggregation of aggregation methods in computational pathology. *Medical Image Analysis*, 88:102885, August 2023. ISSN 1361-8415. doi: 10.1016/j.media.2023.102885. URL https://www.sciencedirect.com/science/article/pii/S1361841523001457.

Nadia Brancati, Anna Maria Anniciello, Pushpak Pati, Daniel Riccio, Giosuè Scognamiglio, Guillaume Jaume, Giuseppe De Pietro, Maurizio Di Bonito, Antonio Foncubierta, Gerardo Botti, Maria Gabrani, Florinda Feroce, and Maria Frucci. BRACS: A Dataset for BReAst Carcinoma Subtyping in H&E Histology Images. *Database*, 2022:baac093, January 2022. ISSN 1758-0463. doi: 10.1093/database/baac093. URL https://doi.org/10.1093/database/baac093.

Romain Brixtel, Sebastien Bougleux, Olivier Lezoray, Yann Caillot, Benoit Lemoine, Mathieu Fontaine, Dalal Nebati, and Arnaud Renouf. Whole Slide Image Quality in Digital Pathology: Review and Perspectives. *IEEE Access*, 10:131005–131035, 2022. ISSN 2169-3536. doi: 10.1109/ACCESS.2022.3227437. URL https://ieeexplore.ieee.org/document/9973240/.

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

Joshua D. Campbell, Anton Alexandrov, Jaegil Kim, Jeremiah Wala, Alice H. Berger, Chandra Sekhar Pedamallu, Sachet A. Shukla, Guangwu Guo, Angela N. Brooks, Bradley A. Murray, Marcin Imielinski, Xin Hu, Shiyun Ling, Rehan Akbani, Mara Rosenberg, Carrie Cibulskis, Aruna Ramachandran, Eric A. Collisson, David J. Kwiatkowski, Michael S. Lawrence, John N. Weinstein, Roel G. W. Verhaak, Catherine J. Wu, Peter S. Hammerman, Andrew D. Cherniack, Gad Getz, Maxim N. Artyomov, Robert Schreiber, Ramaswamy Govindan, and Matthew Meyerson. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Na-*

ture Genetics, 48(6):607–616, June 2016. ISSN 1546-1718. doi: 10.1038/ng.3564. URL https://www.nature.com/articles/ng.3564. Publisher: Nature Publishing Group.

Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, March 2024. ISSN 1546-170X. doi: 10.1038/s41591-024-02857-3. URL https://www.nature.com/articles/s41591-024-02857-3. Publisher: Nature Publishing Group.

Lee AD Cooper, Elizabeth G Demicco, Joel H Saltz, Reid T Powell, Arvind Rao, and Alexander J Lazar. PanCancer insights from The Cancer Genome Atlas: the pathologist's perspective. *The Journal of Pathology*, 244(5):512–524, 2018. ISSN 1096-9896. doi: 10.1002/path.5028. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/path.5028. _eprint: https://pathsocjournals.onlinelibrary.wiley.com/doi/pdf/10.1002/path.5028.

Yufei Cui, Ziquan Liu, Xiangyu Liu, Xue Liu, Cong Wang, Tei-Wei Kuo, Chun Jason Xue, and Antoni B. Chan. Bayes-MIL: A New Probabilistic Perspective on Attention-based Multiple Instance Learning for Whole Slide Images. September 2022. URL https://openreview.net/forum?id=_geIwiOyUhZ.

Yufei Cui, Ziquan Liu, Xiangyu Liu, Xue Liu, Cong Wang, Tei-Wei Kuo, Chun Jason Xue, and Antoni B. Chan. Bayes-MIL: 11th International Conference on Learning Representations (ICLR 2023). *The Eleventh International Conference on Learning Representations*, May 2023. URL http://www.scopus.com/inward/record.url?scp=85192551962&partnerID=8YFLogxK.

Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen A. W. M. van der Laak, the CAMELYON16 Consortium, Meyke Hermsen, Quirine F. Manson, Maschenka Balkenhol, Oscar Geessink, Nikolaos Stathonikos, Marcory Crf van Dijk, Peter Bult, Francisco Beca, Andrew H. Beck, Dayong Wang, Aditya Khosla, Rishab Gargeya, Humayun Irshad, Aoxiao Zhong, Qi Dou, Quanzheng Li, Hao Chen, Huang-Jing Lin, Pheng-Ann Heng, Christian Haß, Elia Bruni, Quincy Wong, Ugur Halici, Mustafa Ümit Öner, Rengul Cetin-Atalay, Matt Berseth, Vitali Khvatkov, Alexei Vylegzhanin, Oren Kraus, Muhammad Shaban, Nasir Rajpoot, Ruqayya Awan, Korsuk Sirinukunwattana, Talha Qaiser, Yee-Wah Tsang, David Tellez, Jonas Annuscheit, Peter Hufnagl, Mira Valkonen, Kimmo Kartasalo, Leena Latonen, Pekka Ruusuvuori, Kaisa Liimatainen, Shadi Albarqouni, Bharti Mungal, Ami George, Stefanie Demirci, Nassir Navab, Seiryo Watanabe, Shigeto Seno, Yoichi Takenaka, Hideo Matsuda, Hady Ahmady Phoulady, Vassili Kovalev, Alexander Kalinovsky, Vitali Liauchuk, Gloria Bueno, M. Milagro Fernandez-Carrobles, Ismael Serrano, Oscar Deniz, Daniel Racoceanu, and Rui Venâncio. Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer. *JAMA*, 318(22):2199–2210, December 2017. ISSN 1538-3598. doi: 10.1001/jama.2017.14585.

Zijie Fang, Yifeng Wang, Ye Zhang, Zhi Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. MamMIL: Multiple Instance Learning for Whole Slide Images with State Space Models. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 3200–3205, December 2024. doi: 10.1109/BIBM62325.2024.10822552. URL https://ieeexplore.ieee.org/document/10822552. ISSN: 2156-1133.

Michael Gadermayr and Maximilian Tschuchnig. Multiple instance learning for digital pathology: A review of the state-of-the-art, limitations & future potential. *Computerized Medical Imaging and Graphics*, 112:102337, March 2024. ISSN 0895-6111. doi: 10.1016/j.compmedimag.2024.102337. URL https://www.sciencedirect.com/science/article/pii/S0895611124000144.

Tamara G Grossmann, Urszula Julia Komorowska, Jonas Latz, and Carola-Bibiane Schönlieb. Can physics-informed neural networks beat the finite element method? *IMA Journal of Applied Mathematics*, 89(1):143–174, January 2024. ISSN 0272-4960. doi: 10.1093/imamat/hxae011. URL https://doi.org/10.1093/imamat/hxae011.

Zhongkai Hao, Zhengyi Wang, Hang Su, Chengyang Ying, Yinpeng Dong, Songming Liu, Ze Cheng, Jian Song, and Jun Zhu. GNOT: A General Neural Operator Transformer for Operator Learning. In *Proceedings of the 40th International Conference on Machine Learning*, pages 12556–12569. PMLR, July 2023. URL https://proceedings.mlr.press/v202/hao23c.html. ISSN: 2640-3498.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based Deep Multiple Instance Learning. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2127–2136. PMLR, July 2018. URL https://proceedings.mlr.press/v80/ilse18a.html. ISSN: 2640-3498.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention. In *Proceedings of the 37th International Conference on Machine Learning*, pages 5156–5165. PMLR, November 2020. URL https://proceedings.mlr.press/v119/katharopoulos20a.html. ISSN: 2640-3498.

Bin Li, Yin Li, and Kevin W. Eliceiri. Dual-Stream Multiple Instance Learning Network for Whole Slide Image Classification With Self-Supervised Contrastive Learning. pages 14318–14328, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Li_Dual-Stream_Multiple_Instance_Learning_Network_for_Whole_Slide_Image_Classification_CVPR_2021_paper.html.

Zijie Li, Kazem Meidani, and Amir Barati Farimani. Transformer for Partial Differential Equations' Operator Learning. 2022. doi: 10.48550/ARXIV.2205.13671. URL https://arxiv.org/abs/2205.13671. Publisher: arXiv Version Number: 3.

Zong-Yi Li, Nikola B. Kovachki, K. Azizzadenesheli, Burigede Liu, K. Bhattacharya, Andrew M. Stuart, and Anima Anandkumar. Fourier Neural Operator for Parametric Partial Differential Equations. *ArXiv*, October 2020. URL https://www.semanticscholar.org/paper/Fourier-Neural-Operator-for-Parametric-Partial-Li-Kovachki/2f7dc1ee85e9f6a97810c66016e09ffeed684f03.

Andreas Lolos, Stergios Christodoulidis, Maria Vakalopoulou, Jose Dolz, and Aris Moustakas. SGPMIL: Sparse Gaussian Process Multiple Instance Learning. 2025. doi: 10.48550/ARXIV.2507.08711. URL https://arxiv.org/abs/2507.08711. Publisher: arXiv Version Number: 1.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, March 2021a. ISSN 2522-5839. doi: 10.1038/s42256-021-00302-5. URL https://www.nature.com/articles/s42256-021-00302-5. Publisher: Nature Publishing Group.

Ming Y. Lu, Drew F. K. 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, June 2021b. ISSN 2157-846X. doi: 10.1038/s41551-020-00682-w. URL https://www.nature.com/articles/s41551-020-00682-w. Publisher: Nature Publishing Group.

Huakun Luo, Haixu Wu, Hang Zhou, Lanxiang Xing, Yichen Di, Jianmin Wang, and Mingsheng Long. Transolver++: An Accurate Neural Solver for PDEs on Million-Scale Geometries. June 2025. URL https://openreview.net/forum?id=AM7iAh0krx.

Ali Mammadov, Loïc Le Folgoc, Julien Adam, Anne Buronfosse, Gilles Hayem, Guillaume Hocquet, and Pietro Gori. Self-supervision enhances instance-based multiple instance learning methods in digital pathology: a benchmark study. *Journal of Medical Imaging*, 12(6):061404, June 2025. ISSN 2329-4302, 2329-4310. doi: 10.1117/1.JMI.12.6.061404. URL https://www.spiedigitallibrary.org/journals/journal-of-medical-imaging/volume-12/issue-6/061404/Self-supervision-enhances-instance-based-multiple-instance-learning-methods-in/10.1117/1.JMI.12.6.061404.full. Publisher: SPIE.

Jeremy Nixon, Michael W. Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. Measuring Calibration in Deep Learning. pages 38–41, 2019. URL https://openaccess.thecvf.com/content_CVPRW_2019/html/Uncertainty_and_Robustness_in_Deep_Visual_Learning/Nixon_Measuring_Calibration_in_Deep_Learning_CVPRW_2019_paper.html.

Daniel Shao, Richard J Chen, Andrew H Song, Joel Runevic, Ming Y. Lu, Tong Ding, , and Faisal Mahmood. Do multiple instance learning models transfer? In *International conference on machine learning*, 2025.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and yongbing zhang. TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classification. In *Advances in Neural Information Processing Systems*, volume 34, pages 2136–2147. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/10c272d06794d3e5785d5e7c5356e9ff-Abstract.html.

Andrew H. Song, Guillaume Jaume, Drew F. K. Williamson, Ming Y. Lu, Anurag Vaidya, Tiffany R. Miller, and Faisal Mahmood. Artificial intelligence for digital and computational pathology. *Nature Reviews Bioengineering*, 1(12):930–949, December 2023. ISSN 2731-6092. doi: 10.1038/s44222-023-00096-8. URL https://www.nature.com/articles/s44222-023-00096-8. Publisher: Nature Publishing Group.

Andrew H. Song, Richard J. Chen, Tong Ding, Drew F. K. Williamson, Guillaume Jaume, and Faisal Mahmood. Morphological Prototyping for Unsupervised Slide Representation Learning in Computational Pathology. pages 11566–11578, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Song_Morphological_Prototyping_for_Unsupervised_Slide_Representation_Learning_in_Computational_Pathology_CVPR_2024_paper.html.

Susu Sun, Dominique van Midden, Geert Litjens, and Christian F. Baumgartner. Prototype-Based Multiple Instance Learning for Gigapixel Whole Slide Image Classification. In James C. Gee, Daniel C. Alexander, Jaesung Hong, Juan Eugenio Iglesias, Carole H. Sudre, Archana Venkataraman, Polina Golland, Jong Hyo Kim, and Jinah Park, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, pages 507–517, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-05185-1. doi: 10.1007/978-3-032-05185-1_49.

Anurag Vaidya, Andrew Zhang, Guillaume Jaume, Andrew H. Song, Tong Ding, Sophia J. Wagner, Ming Y. Lu, Paul Doucet, Harry Robertson, Cristina Almagro-Perez, Richard J. Chen, Dina ElHarouni, Georges Ayoub, Connor Bossi, Keith L. Ligon, Georg Gerber, Long Phi Le, and Faisal Mahmood. Molecular-driven Foundation Model for Oncologic Pathology. 2025. doi: 10.48550/ARXIV.2501.16652. URL https://arxiv.org/abs/2501.16652. Publisher: arXiv Version Number: 1.

Jun Wang, Yu Mao, Nan Guan, and Chun Jason Xue. Advances in Multiple Instance Learning for Whole Slide Image Analysis: Techniques, Challenges, and Future Directions. 2024. doi: 10.48550/ARXIV.2408.09476. URL https://arxiv.org/abs/2408.09476. Publisher: arXiv Version Number: 1.

Haixu Wu, Jialong Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Flowformer: Linearizing Transformers with Conservation Flows. February 2022. URL https://www.semanticscholar.org/paper/Flowformer%3A-Linearizing-Transformers-with-Flows-Wu-Wu/9b61adb6f0d1e8831ab2f5481a12e2125b13c50a.

Haixu Wu, Huakun Luo, Haowen Wang, Jianmin Wang, and Mingsheng Long. Transolver: A Fast Transformer Solver for PDEs on General Geometries. In *Proceedings of the 41st International Conference on Machine Learning*, pages 53681–53705. PMLR, July 2024. URL https://proceedings.mlr.press/v235/wu24r.html. ISSN: 2640-3498.

Conghao Xiong, Hao Chen, and Joseph J. Y. Sung. A Survey of Pathology Foundation Model: Progress and Future Directions. 2025. doi: 10.48550/ARXIV.2504.04045. URL https://arxiv.org/abs/2504.04045. Publisher: arXiv Version Number: 2.

Andrew Zhang, Guillaume Jaume, Anurag Vaidya, Tong Ding, and Faisal Mahmood. Accelerating Data Processing and Benchmarking of AI Models for Pathology. 2025a. doi: 10.48550/ARXIV.2502.06750. URL https://arxiv.org/abs/2502.06750. Publisher: arXiv Version Number: 1.

Yunlong Zhang, Honglin Li, Yunxuan Sun, Sunyi Zheng, Chenglu Zhu, and Lin Yang. Attention-Challenging Multiple Instance Learning for Whole Slide Image Classification. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 125–143, Cham, 2025b. Springer Nature Switzerland. ISBN 978-3-031-73668-1. doi: 10.1007/978-3-031-73668-1_8.

Wenhui Zhu, Xiwen Chen, Peijie Qiu, Aristeidis Sotiras, Abolfazl Razi, and Yalin Wang. DGR-MIL: Exploring Diverse Global Representation in Multiple Instance Learning for Whole Slide Image Classification. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 333–351, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72920-1. doi: 10.1007/978-3-031-72920-1_19.

Appendix A. Experiments

A.1. Datasets

CAMELYON16 (Ehteshami Bejnordi et al., 2017) consists of 399 WSIs of sentinel lymph node tissue sections derived from women with breast cancer. The dataset is split into a training set of 270 images and a test set of 129 images. Collected from two medical centers in the Netherlands, it includes exhaustive pixel-level annotations of metastatic regions (both macrometastases and micrometastases) verified by expert pathologists. We use TRIDENT (Vaidya et al., 2025; Zhang et al., 2025a) to segment and patch the WSIs at 10× magnification (Mammadov et al., 2025) into 224x224 non-overlapping patches and utilize the UNIv1 (Chen et al., 2024) encoder for feature extraction (Chen et al., 2024) . Similarly to Lu et. al (Lu et al., 2021b), we follow a 10-fold cross-validation protocol and report mean bag-level performance.

TCGA-NSCLC We use the dataset from The Cancer Genome Atlas (TCGA) program for the non-small cell lung carcinoma (NSCLC) subtyping task (Cooper et al., 2018; Campbell et al., 2016). The dataset consists of Hematoxylin and Eosin (H&E) stained WSIs in 2 distinct cohorts: Lung Adenocarcinoma (TCGA-LUAD) and Lung Squamous Cell Carcinoma (TCGA-LUSC) (Campbell et al., 2016; Cooper et al., 2018). Specifically, we use 494 LUAD and 512 LUSC cases for a total of 1,006 slides, segment and patch at 10× magnification (Mammadov et al., 2025) into 224x224 non-overlapping patches and use the UNIv1 (Chen et al., 2024) encoder for feature extraction. Performance is reported over 4 folds.

PANDA (Bulten et al., 2022) is derived from the MICCAI 2020 Prostate Cancer Grade Assessment challenge and comprises 10,609 WSIs from prostate core needle biopsies annotated, providing slide-level Gleason scores and ISUP grades alongside expert tissue annotations. We address ISUP grading (0-5) as a 6-class classification task and follow a 5-fold cross-validation protocol using stratified splits, with each fold containing approximately 80 splits for training, 5 for validation and 15 for testing. We segment the WSIs into non-overlapping patches of size 224×224 pixels at 20× magnification (Song et al., 2024) and use the UNIv1 (Chen et al., 2024) encoder for feature extraction.

BRACS (Brancati et al., 2022) The BReAst Carcinoma Subtyping (BRACS) dataset comprises 547 H&E stained WSIs and over 4,500 annotated regions of interest derived from 189 patients, designed to advance the automatic detection of challenging "atypical" (precancerous) lesions that are often underrepresented in other public datasets. It is annotated into seven histological subtypes, grouped into three main categories: Benign (Normal, Pathological Benign, Usual Ductal Hyperplasia), Atypical (Flat Epithelial Atypia, Atypical Ductal Hyperplasia), and Malignant (Ductal Carcinoma in Situ, Invasive Carcinoma). We specifically focus on coarse classification into the three main categories (3-class classification), using the train/validation/test split provided with the dataset. We segment the WSIs into non-overlapping patches of size 224×224 pixels at 20× magnification (Song et al., 2024), and use the UNIv1 (Chen et al., 2024) encoder for feature extraction. Performance is reported over 5 seeds.

A.2. Implementation Details

All models are trained and evaluated in Python with PyTorch, using the same PyTorch Lightning training pipeline with identical data loading, batching, and hardware configurations. For CAPRMIL , training is performed using the standard cross-entropy loss on slide-level labels, while competing methods are trained using the loss functions specified in their original works.

The models are trained for a maximum of 30 epochs on a single A100 GPU, using full-precision (FP32) arithmetic, except MeanMIL which is trained for a maximum of 50 epochs to obtain convergence. CAPRMIL optimization employs AdamW with a base learning rate of 2×10^{-4}, weight decay of 1×10^{-5}, and momentum parameter 0.9. We use a cosine annealing learning-rate schedule, with a 6-epoch warm-up phase starting at 1×10^{-5}, and minimum learning rate of 1×10^{-7}. Early stopping was governed by a patience of 20 epochs and a performance threshold of 10^{-4}. Learning-rate dynamics were logged every epoch, with explicit tracking of weight-decay values to enable fine-grained monitoring of the training process.

Model-specific hyperparameters and optimizer choices for competitors follow the respective original papers and are selected to ensure stable convergence based on observed loss curves. FLOPs are reported for a single forward pass during evaluation and a dummy input bag of 1000 patch embeddings and serve as a proxy for algorithmic complexity. Wall-clock training and inference times measure end-to-end execution. We additionally report peak GPU utilization as an implementation-level efficiency metric reflecting how effectively each model translates computation into hardware usage under identical experimental conditions. Complete code and instructions are publicly available at https://github.com/mandlos/CAPRMIL.

A.3. Evaluation of context-aware tokens

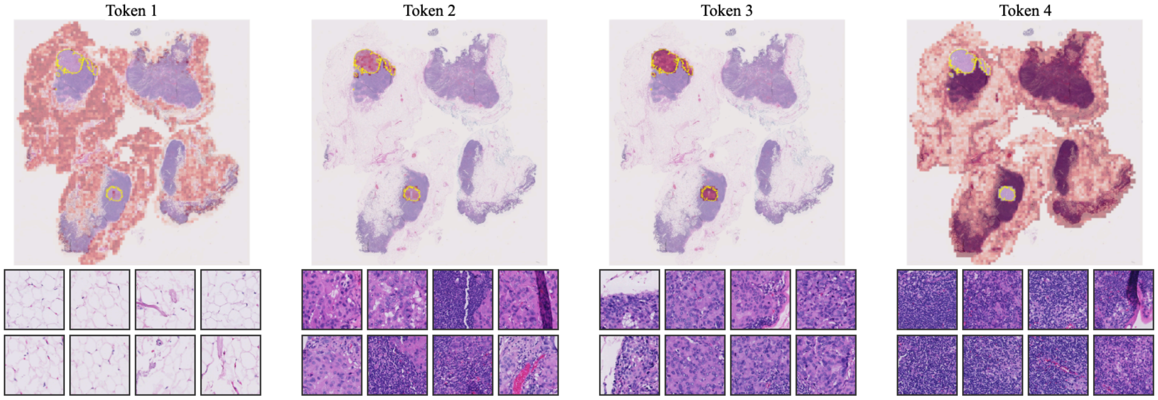


Figure 5: **Token–patch assignment heatmaps.** Test slide from CAMELYON16. *Top:* Soft assignment weights from one TSMIL attention head, indicating each patch's contribution to the M context-aware tokens. *Bottom:* Top-8 patches per token ranked by assignment score, highlighting the dominant morphological patterns contributing to each token.

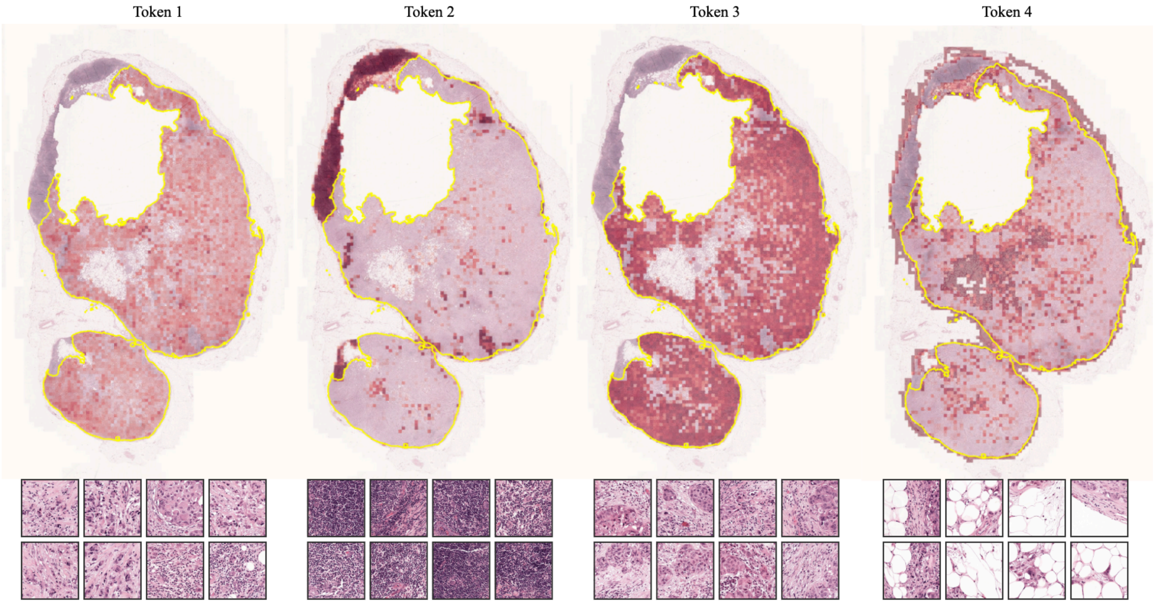


Figure 6: **Token–patch assignment heatmaps.** Test slide from CAMELYON16. *Top:* Soft assignment weights from one TSMIL attention head, indicating each patch's contribution to the M context-aware tokens. *Bottom:* Top-8 patches per token ranked by assignment score, highlighting the dominant morphological patterns contributing to each token.

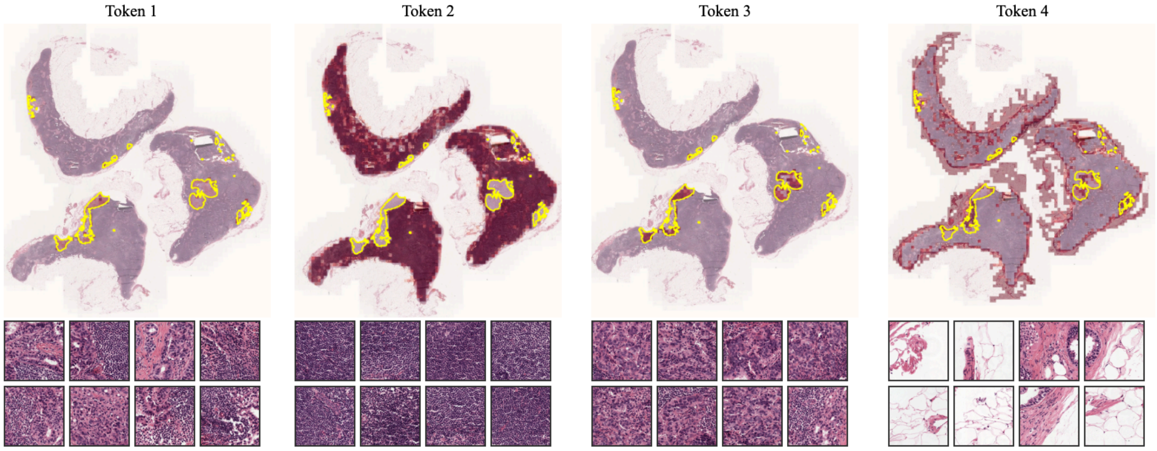


Figure 7: **Token–patch assignment heatmaps.** Test slide from CAMELYON16. *Top:* Soft assignment weights from one TSMIL attention head, indicating each patch's contribution to the M context-aware tokens. *Bottom:* Top-8 patches per token ranked by assignment score, highlighting the dominant morphological patterns contributing to each token.

Appendix B. Ablation Studies

We ablate key architectural choices of the Transolver block along three orthogonal axes: the number of clusters used for tokenization (M), the number of attention heads (H), and

Clusters (M)	Heads (H)	MLP ratio	AUC	ACE	Params (M)
2	8	4	$.971_{.009}$	$\mathbf{.027_{.007}}$	0.314
4	8	4	$\mathbf{.975_{.006}}$	$.028_{.006}$	0.314
8	8	4	$.974_{.009}$	$.030_{.008}$	0.314
16	8	4	$.973_{.008}$	$\mathbf{.027_{.008}}$	0.314
4	2	4	$.971_{.010}$	$.031_{.011}$	0.326
4	4	4	$.972_{.009}$	$.033_{.011}$	0.316
4	8	4	$\mathbf{.975_{.006}}$	$\mathbf{.028_{.006}}$	0.314
4	12	4	$.972_{.009}$	$.032_{.006}$	**0.310**
4	8	1	$.973_{.008}$	$\mathbf{.027_{.006}}$	**0.215**
4	8	2	$.969_{.011}$	$.030_{.008}$	0.248
4	8	4	$\mathbf{.975_{.006}}$	$.028_{.006}$	0.314

Table 3: Ablation of the number of clusters M, attention heads H, and the MLP expansion ratio in the Transolver block. Exhaustive ablation results for the number of clusters, attention heads, and input embedding dimensionality are provided in Figures A 8, 9, 10, 11, and 12.

the MLP expansion ratio. Experiments are conducted on CAMELYON16, with full sweeps reported in Figures A 8, 9, 10, 11 and 12.

Model	Training (s)	Inference (s)	Params (M)	FLOPs (G)
ABMIL (Ilse et al., 2018)	**5.5**	**0.8**	0.660	1.31
CLAM (Lu et al., 2021b)	7.0	0.9	0.920	1.84
TransMIL (Shao et al., 2021)	13.4	1.2	2.67	85.02
DGRMIL (Zhu et al., 2025)	16.7	1.5	4.34	79.88
BayesMIL (Cui et al., 2023)	9.5	1.1	1.32	2.63
SGPMIL (Lolos et al., 2025)	9.0	1.0	1.21	2.43
CAPRMIL	6.3	**0.8**	**0.314**	**0.628**

Table 4: Training and inference times (in seconds) and model sizes (number of trainable parameters in millions, M). Training times are averaged over 30 epochs, while inference times correspond to processing the full test set of 129 slides.

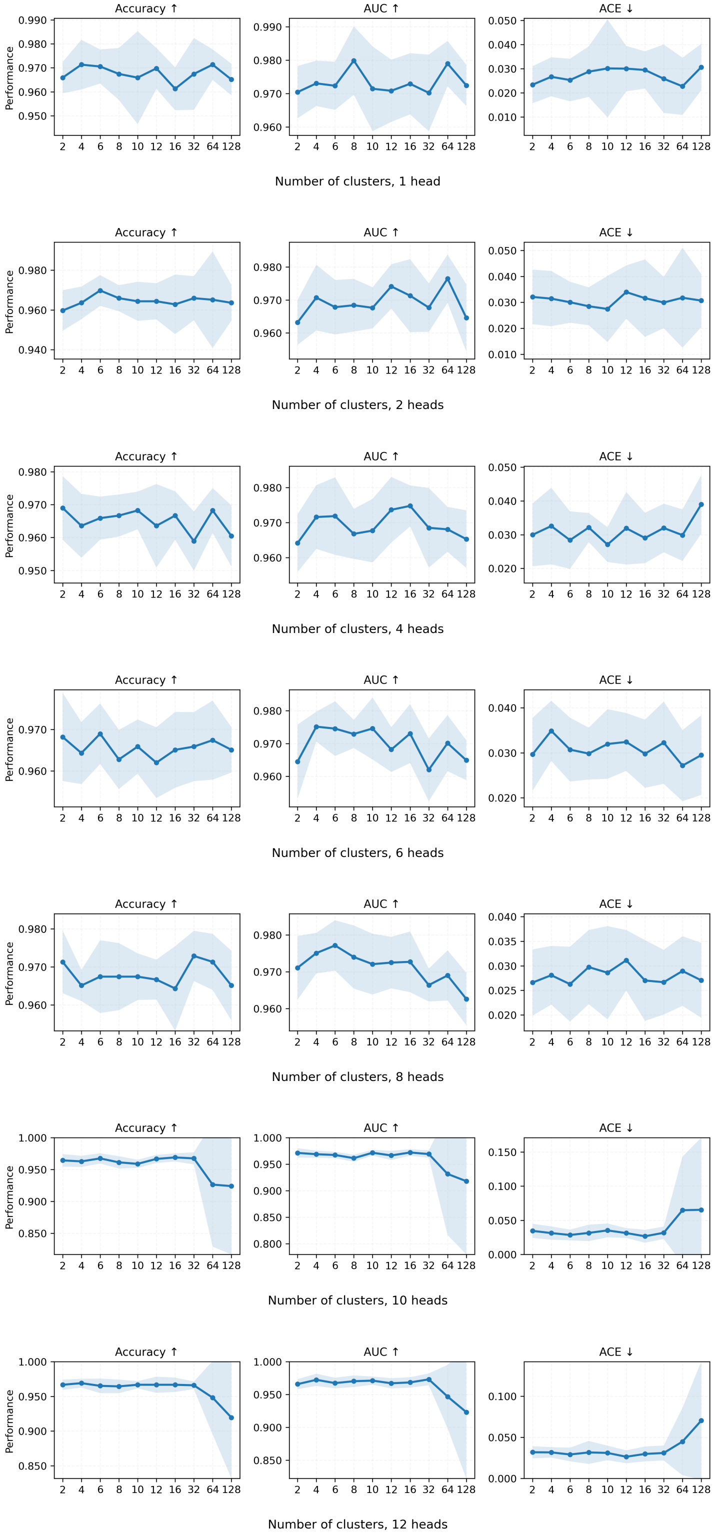


Figure 8: Ablation on the number of clusters for CAMELYON16. For each row, the number of attention heads is fixed while the number of clusters is varied. From top to bottom: 1, 2, 4, 6, 8, 10 and 12 heads.

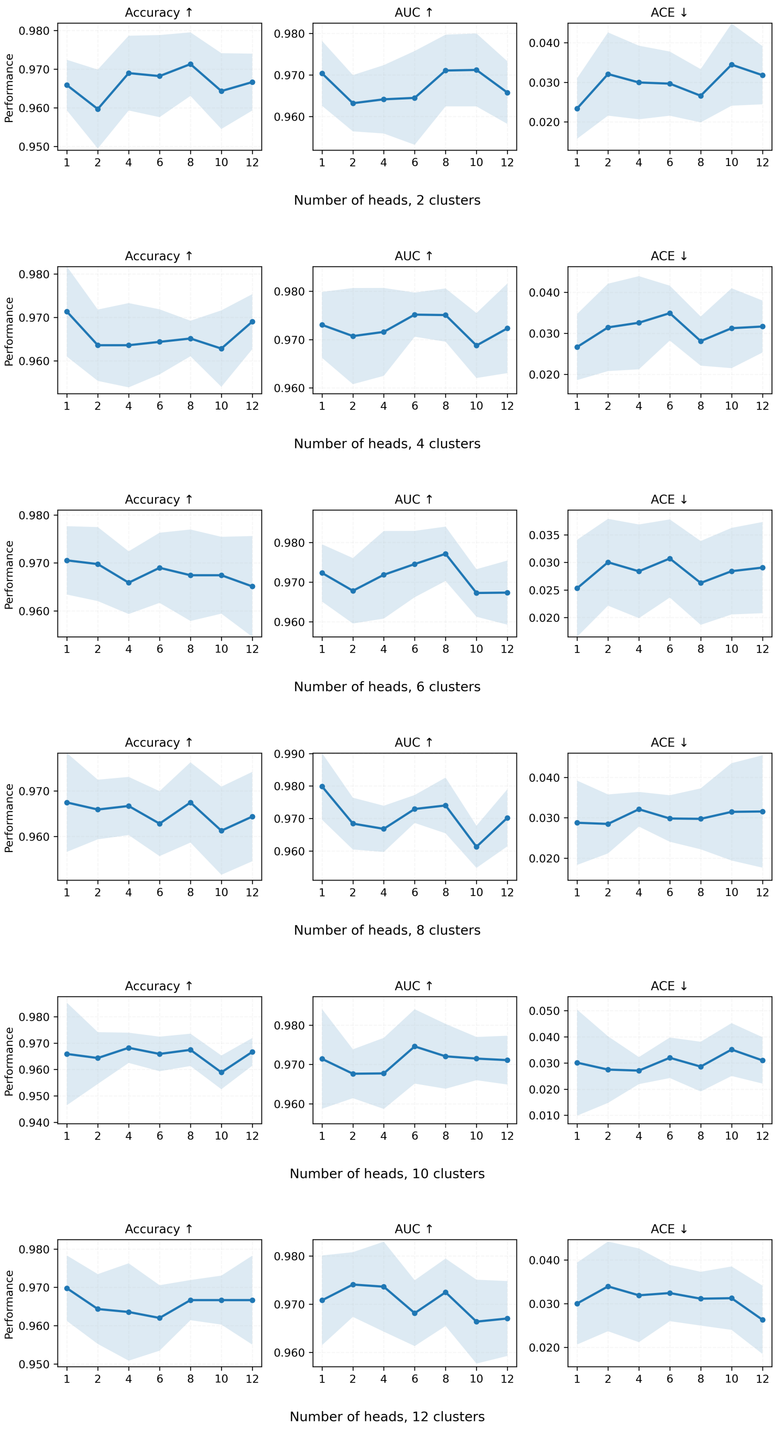


Figure 9: Ablation on the number of attention heads for a single CAPRMIL block on CAMELYON16. For each row, the number of clusters is fixed while the number of heads is varied. From top to bottom: 2, 4, 6, 8, 10, and 12 clusters.

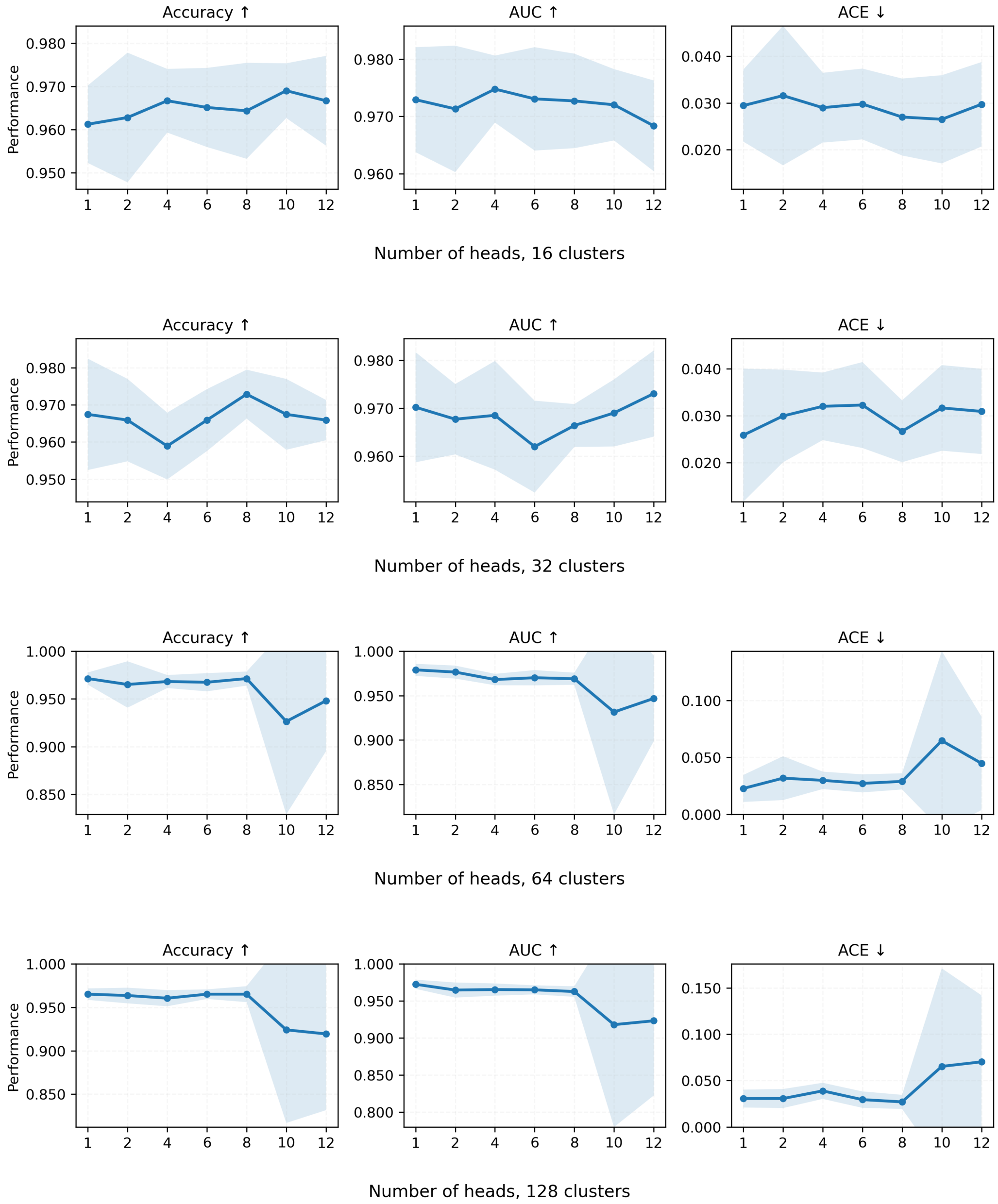


Figure 10: Ablation on the number of attention heads for a single CAPRMIL block on CAMELYON16. For each row, the number of clusters is fixed while the number of heads is varied. From top to bottom: 16, 32, 64, and 128 clusters.

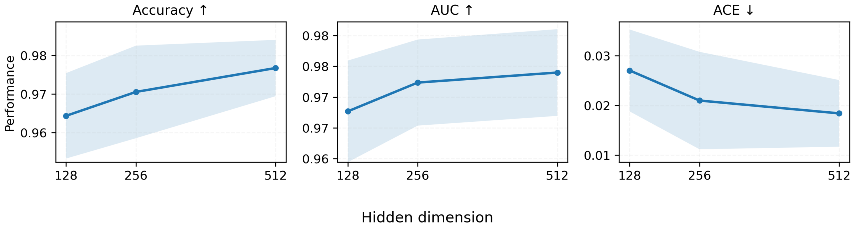


Figure 11: Ablation on the dimensionality of the input projection layer on CAMELYON16. We vary the number of hidden units while keeping the number of clusters (16) and attention heads (8) fixed.

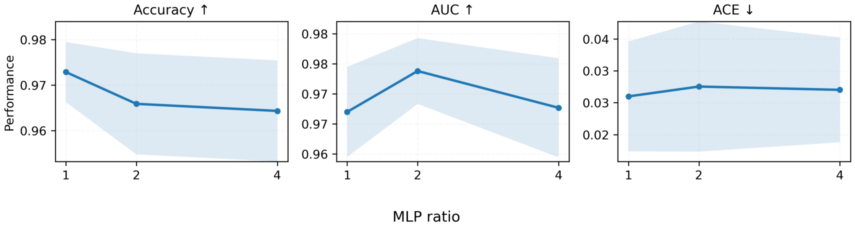


Figure 12: Ablation on the MLP expansion ratio in the CAPRMIL block on CAMELYON16. We vary the expansion factor of the feed-forward network (MLP ratio $\in \{1, 2, 4\}$) while keeping all other components fixed (16 clusters, 8 heads).