CKMIL: Cascaded Key-Instance Attention Multiple Instance Learning for Histopathology Whole Slide Image Analysis

Anonymous submission

Abstract

In computational pathology (CPath), the analysis of Whole Slide Images (WSIs) using Multiple Instance Learning (MIL) is a key technology for precision medicine. However, existing methods face a dilemma when modeling inter-instance correlations: they either overlook the correlations entirely or model them in a key-instance agnostic manner. Methods based on the independent attention weighting ignore interactions among instances, while the standard self-attention mechanism is difficult to apply to WSIs with massive numbers of instances due to its $O(n^2)$ computational complexity. Although recent linear-complexity methods have addressed the efficiency issue, they generally adopt a key-instance agnostic strategy. This can dilute the sparse yet crucial diagnostic signals in WSIs, leading to suboptimal performance. To address this challenge, we propose CKMIL, a novel Cascaded Key-Instance Attention framework. CKMIL operates via a two-stage cascaded process: first, a Subspace-Disentangled Attention (SDA) module identifies candidate key sub-instances with high discriminative scores within multiple feature subspaces. Subsequently, a Key-Instance Guided Global Attention (KGGA) module utilizes these candidates as landmarks for Nyström attention. This achieves efficient global interaction guided by key information, effectively preventing the dilution of diagnostic signals. Furthermore, postulating that local correlations exist among the components within an instance's feature vector, we introduce an Instance-Conv-Projection (ICP) module to capture this internal feature structure better. Extensive experiments for cancer subtyping and survival prediction on public datasets, including BRACS and the TCGA-BLCA/BRCA/NSCLC cohorts, demonstrate that when used with feature extractors pre-trained on the general domain, our proposed method surpasses existing mainstream methods in performance.

Introduction

Computational pathology (CPath) (Cai et al. 2021; Cifci et al. 2023), an interdisciplinary field at the intersection of pathology and computer science, has emerged as a frontier with immense potential in precision medicine (Bera et al. 2019). Unlike traditional pathology, which relies on the visual assessment of tissue slides by pathologists—a process that is costly, labor-intensive, and susceptible to inter-observer variability (Elmore et al. 2015), computational pathology leverages computational methods to analyze digitized Whole Slide Images (WSIs) (Cui and Zhang 2021; Song et al. 2023). This provides decision support for early diagnosis, prognosis prediction, and personalized treatment.

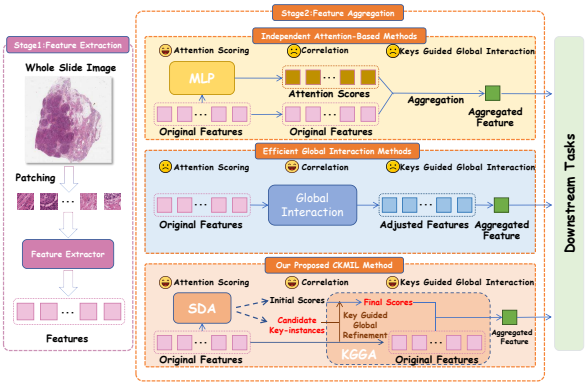

Figure 1: The two-stage paradigm of MIL and a comparison of different MIL methods. Top Methods: Generate attention scores for each instance, but ignore the correlations. Middle Methods: Model inter-instance correlations, but they cannot generate attention scores for individual instances, and their global interaction overlooks the critical role of sparse positive instances. Bottom(Our Method): Our method generates attention scores for each instance and models their correlations through a global interaction guided by key instances. This approach effectively prevents the dilution of key diagnostic signals during the correlation modeling process.

Although WSIs are considered the gold standard in computational pathology due to their ability to capture comprehensive tumor microenvironment (Cai et al. 2021), their gigapixel size (e.g.,$80,000 \times 80,000$ pixels at $40\times$ magnification) and the scarcity of fine-grained annotations present significant challenges for conventional deep learning models (Campanella et al. 2019; Jin et al. 2023).

To address these challenges, Multiple Instance Learning (MIL) has become the de facto paradigm for WSI analysis (Maron and Lozano-Pérez 1997; Amores 2013; Campanella et al. 2019; Lu et al. 2021). In this paradigm, each WSI is treated as a bag, and the patches obtained by dividing it are called instances. The prevalent MIL pipeline employs a pre-trained feature extractor to encode instances into low-dimensional features, followed by an aggregator that pools

instance features into a bag-level representation for downstream tasks such as cancer subtyping (Chen et al. 2013; Coudray et al. 2018) and survival prediction (Yu et al. 2016).

While early MIL methods used simple pooling (Yu et al. 2016), attention-based approaches such as ABMIL (Ilse, Tomczak, and Welling 2018) and CLAM (Lu et al. 2021) were introduced to weight instances by their importance. However, by treating instances as independent and identically distributed (i.i.d.), these models fundamentally ignore the crucial contextual correlations among them. To capture instance correlations, Transformer-based methods were explored, but they faced the prohibitive computational complexity of $O(n^2)$. To overcome the computational complexity, methods with linear complexity, such as MambaMIL (Yang, Wang, and Chen 2024) and TransMIL (Shao et al. 2021), were proposed. However, these approaches often failed to capture the most critical diagnostic information. Their inherent simplification strategies risked diluting the signals from sparse but vital instances within a WSI, leading to suboptimal results.

Overall, existing methods for modeling instance correlations are limited (as illustrated in Figure 1): independent attention neglects instance interplay, while efficient global methods are key-instance agnostic, diluting critical diagnostic signals.

In this paper, We propose Cascaded Key-Instance Attention Multiple Instance Learning (CKMIL), a framework built on the principle that *key instances should guide efficient global interaction.* CKMIL materializes this through a cascaded process. First, our Subspace-Disentangled Attention (SDA) module screens for candidate key instances within feature subspaces. Crucially, the subsequent Key-Instance Guided Global Attention (KGGA) module leverages these very candidates as the landmarks for Nyström attention (Xiong et al. 2021). This design anchors the efficient global interaction directly to the most salient signals. The resulting global context then refines the initial scores from SDA via a gated fusion mechanism, tightly coupling the screening and interaction stages. Additionally, we introduce an exploratory Instance-Conv-Projection (ICP) module to capture intra-feature correlations using convolutions to replace conventional linear layers for generating Q and K vectors. Our primary contributions are as follows:

- A novel cascaded attention framework, CKMIL, that efficiently models inter-instance dependencies in a key-instance-guided manner.
- A Key-Instance Guided Global Attention (KGGA) mechanism that uses key instances as landmarks to address the information dilution problem in existing linear-complexity methods.
- An Instance-Conv-Projection (ICP) module that leverages convolutional fusion to capture latent intra-feature correlations often missed by conventional linear layers.
- State-of-the-art (SOTA) performance with general-purpose feature extractors and strong competitive performance with domain-specific medical feature extractors on cancer subtyping and survival prediction tasks.

Related Work

Multiple Instance Learning for WSI Analysis

The MIL paradigm addresses the challenge of gigapixel-scale WSIs by treating each slide as a bag of instances (Maron and Lozano-Pérez 1997), effectively leveraging bag-level labels. Under this paradigm, the primary objective of MIL becomes learning the relationships among instances within a bag. A typical two-stage MIL approach involves two steps (Lu et al. 2021). First, a feature encoder (e.g., Resnet50 (He et al. 2016)), often pre-trained on large-scale image datasets (Deng et al. 2009), transforms instances into low-dimensional feature vectors. Second, an aggregation module is designed to aggregate instance-level features into a bag-level representation for downstream tasks.

Attention as Independent Instance Weighting

To overcome the limitations of simple pooling aggregators such as Mean-pooling and Max-pooling, attention mechanisms were introduced to assign discriminative weights to instances based on their importance (Ilse, Tomczak, and Welling 2018). Foundational methods in this category, including ABMIL (Ilse, Tomczak, and Welling 2018), CLAM (Lu et al. 2021), and DSMIL (Li, Li, and Eliceiri 2021), typically employ a shared attention network to score each instance independently. However, these methods are fundamentally built on the independent and identically distributed (i.i.d.) assumption, thereby neglecting the correlations between instances. This premise contradicts core pathology principles, where interactions within the tumor microenvironment are often crucial for diagnosis. Consequently, by treating each instance in isolation, these models cannot fully model the broader tissue context and may over-focus on cytologically salient but diagnostically redundant areas.

Global Interaction in MIL with Linear Complexity

To address the context-agnostic nature of independent weighting, methods incorporating global self-attention were explored. However, the standard self-attention mechanism, with its prohibitive $O(n^2)$ computational complexity, is ill-suited for the massive number of instances in a WSI. This challenge motivated the development of several global interaction methods with linear complexity. Prominent examples include TransMIL (Shao et al. 2021), which uses Nyström's method to approximate the attention matrix; MambaMIL (Yang, Wang, and Chen 2024), which leverages the Mamba state space model (Gu and Dao 2023); and RRTMIL (Tang et al. 2024), which adapts the Swin Transformer (Liu et al. 2021). While computationally efficient, each carries its own limitations. TransMIL's Nyström approximation with pooling-based landmarks risks diluting key signals. MambaMIL is constrained by the fixed 1D sequential processing of the Mamba architecture, and RRTMIL's performance is sensitive to its window configuration and parameter count. Critically, these methods share a common flaw: they are key-instance agnostic. By treating all instances uniformly during interaction, they risk overlooking the sparse yet crucial diagnostic signals present in WSIs, leading to suboptimal outcomes.

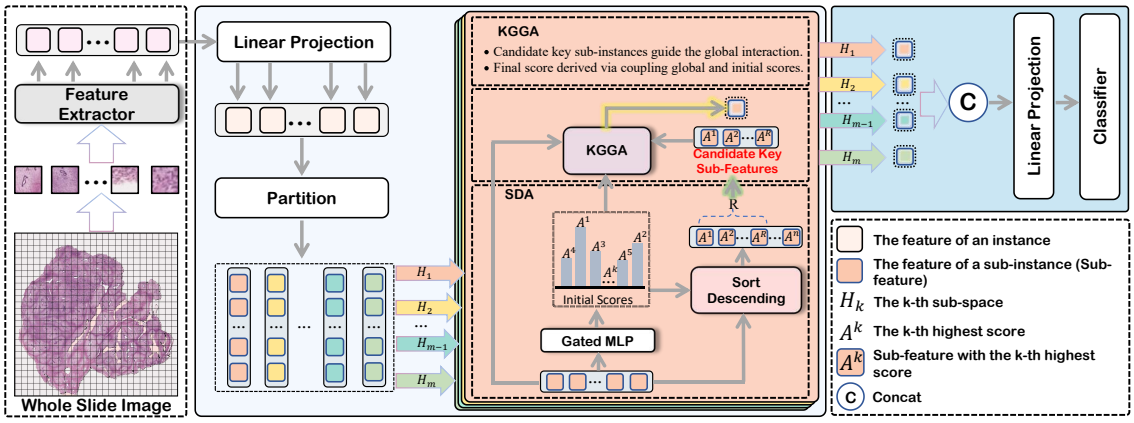

Figure 2: Overview of our proposed CKMIL. CKMIL partitions instance features into multiple sub-spaces, where a sub-space Discriminative Attention (SDA) module selects Candidate Key Sub-Features. These key candidates then drive a Global Interaction with all sub-features in their respective space to generate an aggregated Sub-Feature, achieving efficient and key-instance guided global interaction (KGGA). Finally, all aggregated sub-features are concatenated to form the final bag-level feature.

Methodology

The CKMIL framework is engineered to resolve the impasse where methods either neglect instance correlations or are key-instance agnostic. It leverages a cascaded process that uses key instances to guide global interaction, thereby achieving robust correlation modeling and preventing the dilution of critical diagnostic signals in WSIs.

Problem formulation

Taking binary classification in MIL as an example, to utilize bag-level label Y_i, for $i = 1, 2, \cdots, b, Y_i \in \{0, 1\}$, we have the corresponding instance feature set for each bag $\mathbf{X}_i \in \mathbb{R}^{n \times D} = \{x_{i,1}, \cdots, x_{i,k}, \cdots, x_{i,n}\}$, **for** $i = 1, 2, \cdots, b$. The MIL methodology can be represented as follows:

$$Y_i = \begin{cases} 0, & \text{if } \sum_{k=1}^{n} y_{i,k} = 0 \\ 1, & \text{otherwise} \end{cases} \tag{1}$$

$$\hat{Y}_i = f(\mathbf{X}_i), \tag{2}$$

where $y_{i,k} \in \{0, 1\}$ is the unknown instance-level label, $\hat{Y}_i$ in Eq.2. is the predicted value we obtain using bag $\mathbf{X}_i$, b is the number of WSIs, and n is the number of instances in each bag (the value of n can vary for different bags). The function f is what needs to be designed in MIL. Its main component is the aggregator, whose role is to aggregate instance features $x_{i,1}, \cdots, x_{i,k}, \cdots, x_{i,n}$ into a bag-level feature $\tilde{x}_i$. This feature is then fed into a classification head to obtain the prediction $\hat{Y}_i$, Unlike global interaction methods such as TransMIL (Shao et al. 2021), where the designed function f causes instance-level features to change after global interaction, our proposed CKMIL, despite having global interaction, does not alter the instance-level features themselves. As our comparative experiments show, our approach, when used in two-stage MIL with feature encoders pre-trained on general domain images (like ResNet50 on ImageNet), avoids further distortion and loss of feature information and outperforms other approaches.

Overview of CKMIL

The CKMIL framework, illustrated in Figure 2, operates through a cascaded process designed to leverage sparse diagnostic signals for efficient global attention. Initially, instance features are partitioned into multiple subspaces where the SDA module performs a screening to identify a set of candidate key sub-instances with high discriminative scores. These candidates are then utilized by the KGGA module as landmarks for Nyström-based attention (Xiong et al. 2021), facilitating an efficient global interaction explicitly guided by high-value signals. Subsequently, the global context from KGGA modulates the initial scores from SDA to obtain the global scores, then via a gated fusion mechanism to produce refined final scores. These final scores guide the weighted aggregation within each subspace, and the resulting sub-features are concatenated to form the final bag-level representation. Additionally, the framework includes the Instance-Conv-Projection (ICP) module in an attempt to capture local intra-feature correlations. This component explores using convolutional fusion instead of standard linear projections for generating Query (Q) and Key (K) vectors.

Subspace-Disentangled Attention (SDA)

To mitigate the risk of attention focusing on non-critical regions and to encourage feature diversity, inspired by the local attention within multi-head spaces in ABMILX (Tang et al. 2025), we propose SDA, the SDA module partitions instance features and screens for key signals independently within each subspace. Given a set of instances for a bag $\mathbf{X} \in \mathbb{R}^{n \times D} = \{x_1, \cdots, x_k, \cdots, x_n\}$, we first partition the features of the instances in the bag into m different low-dimensional feature subspaces, obtaining a collection of bags in different feature subspaces, denoted as $\{\mathbf{X}_1, \cdots, \mathbf{X}_h, \cdots, \mathbf{X}_m\}, \mathbf{X}_h \in \mathbb{R}^{n \times \frac{D}{m}}$, **for** $h = 1, \cdots, m$. For a given subspace $\mathbf{H}_h$, an independent gated MLP layer:

$$\mathbf{A}_h^T = \mathbf{G}_h \cdot [\mathbf{W}_h(\tanh(\mathbf{E}_h \mathbf{X}_h^T)) \odot \sigma(\mathbf{U}_h \mathbf{X}_h^T)] \in \mathbb{R}^{1 \times n}, \tag{3}$$

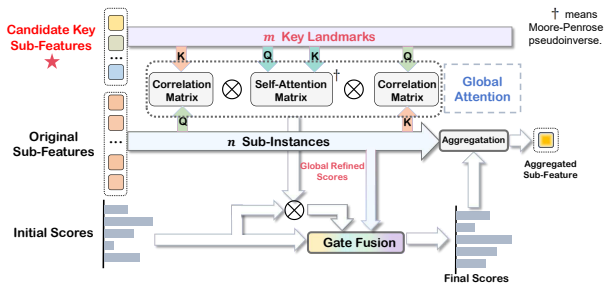

Figure 3: Our proposed KGGA refines initial weights by globally interacting with key candidate sub-instances from the SDA module to embed instance correlation.

computes initial scores $\mathbf{A}_h$ for all sub-instances, where $\mathbf{G_h} \in \mathbb{R}^{1\times\frac{D}{4m}}, \mathbf{W_h} \in \mathbb{R}^{\frac{D}{4m}\times\frac{D}{4m}}, \mathbf{E_h} \in \mathbb{R}^{\frac{D}{4m}\times\frac{D}{m}}, \mathbf{U_h} \in \mathbb{R}^{\frac{D}{4m}\times\frac{D}{m}}$ are trainable matrices, and D is the dimension of the instances. Sub-instances are then ranked by these scores, and the top-r are selected to form the candidate key set $\mathbf{L}_h \in \mathbb{R}^{r\times(D/m)}$ for that subspace:

$$(\tilde{\mathbf{x}}_{h,1}, \tilde{a}_{h,1}), \cdots, (\tilde{\mathbf{x}}_{h,r}, \tilde{a}_{h,r}), \cdots, (\tilde{\mathbf{x}}_{h,n}, \tilde{a}_{h,n}) = \text{SortDescending}((\mathbf{x}_{h,1}, a_{h,1}), \cdots, (\mathbf{x}_{h,n}, a_{h,n})), \tag{4}$$

$$\mathbf{L}_h = \{\tilde{\mathbf{x}}_{h,1}, \tilde{\mathbf{x}}_{h,2}, \ldots, \tilde{\mathbf{x}}_{h,r}\} \in \mathbb{R}^{r\times(D/m)}, \tag{5}$$

where $\mathbf{x}_{h,i}$ represents the sub-instance feature of the i-th instance in the h-th feature subspace, $a_{h,i}$ represents the independent weight score of the i-th instance in the h-th feature subspace, $\tilde{\mathbf{x}}_{h,i}$ represents the sub-instance feature with the i-th highest score in the h-th feature subspace, and $\mathbf{L}_h$ is the candidate key sub-instances in the h-th feature subspace.

Key-Instance Guided Global Attention (KGGA)

To efficiently model the correlations among the vast number of instances in a WSI, we adopt the Nyström attention mechanism (Xiong et al. 2021). This method achieves a linear $O(n)$ complexity by constructing a low-rank approximation of the full attention matrix. The mathematical foundation for this is the CUR matrix decomposition. This principle approximates a large matrix by using a subset of its actual columns ($\mathbf{C}$) and rows ($\mathbf{R}$), along with a smaller, low-dimensional core matrix ($\mathbf{U}$), to reconstruct an approximation of the original matrix (i.e., $\mathbf{A} \approx \mathbf{CUR}$). However, a critical challenge lies in the landmark selection strategy. Conventional Nyström Attention implementations typically select these landmarks using pooling-based strategies. The core matrix (approximating $\mathbf{U}$) is then formed from the self-attention matrix computed among these pooled landmarks. While this process effectively reduces computational complexity, the approach is fundamentally key-instance agnostic, which risks diluting the sparse yet crucial diagnostic signals within the WSI.

To address the key-agnostic nature, the KGGA module is designed (as illustrated in Figure 3). In contrast to the method based on average pooling, it leverages the candidate key sub-instances $\mathbf{L}_h$ from SDA as landmarks for Nyström attention, ensuring that global interaction is anchored by diagnostically relevant signals. The computation of the approximated global attention matrix $\hat{\mathbf{S}}_h$ is described as:

$$\hat{\mathbf{S}}_h = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}_{\mathbf{L_h}}^T}{\sqrt{D/m}}\right)(\mathbf{M})^\dagger \,\text{softmax}\left(\frac{\mathbf{Q}_{\mathbf{L_h}}\mathbf{K}^T}{\sqrt{D/m}}\right), \tag{6}$$

$$\mathbf{M} = \text{softmax}\left(\frac{\mathbf{Q}_{\mathbf{L_h}}\mathbf{K}_{\mathbf{L_h}}^T}{\sqrt{D/m}}\right), \tag{7}$$

where $\mathbf{Q}_{\mathbf{L_h}}$ and $\mathbf{K}_{\mathbf{L_h}}$ are the query and key matrices corresponding to the $\mathbf{L}_h$ landmarks, and $\mathbf{M}^\dagger$ denotes the Moore-Penrose pseudoinverse of $\mathbf{M}$.

The initial scores $\mathbf{A}_h$ obtained from the SDA module fail to adequately consider the correlations among instances. Therefore, to generate the global-aware scores $\mathbf{B}_h$ while maintaining a computational complexity of $O(n)$, we apply the associative law of multiplication to left-multiply $\hat{\mathbf{S}}_h$ by the initial scores $\mathbf{A}_h$, resulting in the following expression:

$$\mathbf{\Phi_1} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}_{\mathbf{L_h}}^T}{\sqrt{D/m}}\right), \mathbf{\Phi_2} = \text{softmax}\left(\frac{\mathbf{Q}_{\mathbf{L_h}}\mathbf{K}^T}{\sqrt{D/m}}\right), \tag{8}$$

$$\mathbf{B}_h = \hat{\mathbf{S}}_h \cdot \mathbf{A}_h = [\mathbf{\Phi_1}(\mathbf{M})^\dagger]_{n\times m}[\mathbf{\Phi_2}\mathbf{A}_h]_{m\times 1}. \tag{9}$$

To create a synergistic coupling between the screening (SDA) and interaction (KGGA) stages, a gated mechanism fuses the initial scores $\mathbf{A}_h$ and the global refined scores $\mathbf{B}_h$ into final scores $\mathbf{C}_h$:

$$\mathbf{g} = \sigma(\mathbf{X}_h\mathbf{W}_g) \in \mathbb{R}^{n\times 1}, \tag{10}$$

$$\mathbf{C}_h = (1-\mathbf{g}) \odot \mathbf{A}_h + \mathbf{g} \odot \mathbf{B}_h, \tag{11}$$

where $\mathbf{W}_g$ is a trainable matrix and σ means the Sigma function. These final scores guide the weighted aggregation of sub-features into a subspace representation $\mathbf{Z}_h$ for downstream task analysis:

$$\mathbf{Z}_h = \text{softmax}\left(\mathbf{C}_h^T\right)\mathbf{X}_h. \tag{12}$$

Finally, all subspace representations are concatenated to form the bag-level feature $\mathbf{Z}$:

$$\mathbf{Z} = \text{concat}(\mathbf{Z}_1, \ldots, \mathbf{Z}_h, \ldots, \mathbf{Z}_m). \tag{13}$$

Instance-Conv-Projection (ICP)

Conventional attention mechanisms generate Query (Q) and Key (K) vectors using linear projections, which have weak capabilities in modeling the local, intra-feature correlations crucial in pathology. To address this, the ICP module integrates the local fusion capabilities of convolutions.

As shown in Figure 4, ICP implements a Reshape-Convolution-Reshape-Projection pipeline. An input 1D instance feature $x_i \in \mathbb{R}^{1\times D}$ is first reshaped ($\mathbf{R}$) into a 2D pseudo-image. A lightweight convolutional layer then processes this tensor, capturing local structural patterns imperceptible to a standard linear layer. The tensor is then flattened back ($\mathbf{F}$) to a 1D vector and projected to generate the final Q_i or K_i vector:

$$Q_i(K_i) = \textbf{Linear}(\mathbf{F}(\textbf{Conv}(\mathbf{R}(x_i)))). \tag{14}$$

Methods	BRACS-3		BRCA-2		NSCLC-2	
	AUC	ACC	AUC	ACC	AUC	ACC
ResNet-50						
Mean-Pooling	0.8051±0.0319	0.6444±0.0337	0.9068±0.0276	0.8410±0.0262	0.8914±0.0203	0.8209±0.0282
Max-Pooling	0.8064±0.0359	0.6907±0.0356	0.8372±0.0239	0.8152±0.0260	0.9163±0.0314	0.8342±0.0340
ABMIL (Ilse, Tomczak, and Welling 2018)	0.8004±0.0382	0.6981±0.0368	0.8883±0.0190	0.8139±0.0401	0.9359±0.0276	0.8685±0.0463
CLAM-MB (Lu et al. 2021)	0.8134±0.0287	0.6833±0.0280	0.8929±0.0177	0.8210±0.0232	0.9407±0.0207	0.8685±0.0266
DSMIL (Li, Li, and Eliceiri 2021)	0.7950±0.0365	0.6481±0.0476	0.8196±0.0766	0.7809±0.0540	0.8491±0.0779	0.7561±0.0701
TransMIL (Shao et al. 2021)	0.8160±0.0406	0.7111±0.0200	0.8774±0.0386	0.8145±0.0445	0.9348±0.0192	0.8495±0.0415
MambaMIL (Yang, Wang, and Chen 2024)	0.8305±0.0427	0.7111±0.0553	0.8949±0.0375	0.8632±0.0273	0.9374±0.0190	0.8743±0.0302
RRTMIL (Tang et al. 2024)	0.8160±0.0257	0.7129±0.0185	0.9163±0.0290	0.8484±0.0386	0.9421±0.0146	0.8723±0.0136
CKMIL-Base (ours)	0.8483±0.0260	0.7130±0.0515	**0.9269±0.0358**	**0.8716±0.0274**	0.9439±0.0225	0.8752±0.0317
CKMIL (ours)	**0.8583±0.0297**	**0.7370±0.0427**	0.9255±0.0261	0.8648±0.0252	**0.9549±0.0148**	**0.8838±0.0253**
UNI						
Mean-Pooling	0.8771±0.0259	0.7203±0.0411	0.9552±0.0258	0.8943±0.0237	0.9746±0.0122	0.9257±0.0219
Max-Pooling	0.8596±0.0285	0.7503±0.0101	0.9627±0.0190	0.9136±0.0109	0.9816±0.0109	0.9361±0.0246
ABMIL (Ilse, Tomczak, and Welling 2018)	0.8901±0.0426	0.7635±0.0567	**0.9671±0.0240**	0.9187±0.0106	0.9796±0.0118	**0.9485±0.0197**
CLAM-MB (Lu et al. 2021)	0.8862±0.0343	0.7629±0.0456	0.9625±0.0176	**0.9291±0.0186**	0.9825±0.0117	0.9409±0.0183
DSMIL (Li, Li, and Eliceiri 2021)	0.8399±0.0169	0.7185±0.0266	0.9533±0.0124	0.8900±0.0129	0.9739±0.0129	0.9200±0.0278
TransMIL (Shao et al. 2021)	0.8549±0.0226	0.7407±0.0340	0.9488±0.0293	0.9195±0.0172	0.9766±0.0124	0.9190±0.0187
MambaMIL (Yang, Wang, and Chen 2024)	0.8842±0.0234	0.7645±0.0292	0.9568±0.0234	0.9099±0.0221	0.9791±0.0120	0.9352±0.0204
RRTMIL (Tang et al. 2024)	0.8754±0.0284	0.7574±0.0583	0.9586±0.0221	0.9178±0.0153	0.9818±0.0115	0.9323±0.0182
CKMIL-Base (ours)	**0.8967±0.0275**	**0.7648±0.0274**	0.9579±0.0192	0.9160±0.0253	0.9756±0.0086	0.9342±0.0169
CKMIL (ours)	0.8952±0.0203	**0.7648±0.0258**	0.9556±0.0208	0.9125±0.0274	**0.9836±0.0103**	0.9361±0.0234

Table 1: Performance comparison on cancer subtyping tasks. Best results are in **bold**, and second-best results are underlined.

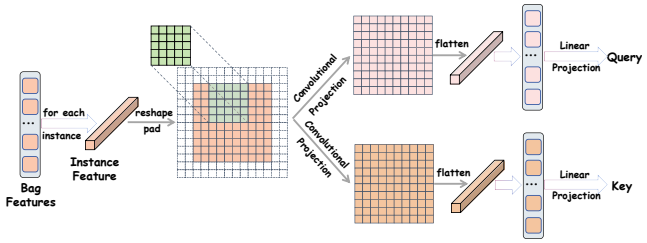

Figure 4: The framework of ICP following a Reshape-Convolution-Reshape-Projection pipeline.

Experiments

Datasets and Evaluation Metrics

To validate the efficacy of our proposed CKMIL, we conducted extensive experiments on two representative downstream tasks across four public datasets.

Survival Prediction We selected three public cancer datasets from The Cancer Genome Atlas (TCGA) (Weinstein et al. 2013): BLCA, BRCA, and LUAD, which contain Whole Slide Images (WSIs) with corresponding survival time annotations. Following the experimental setup of GPFM (Ma et al. 2024), we employ a 5-fold cross-validation methodology to mitigate the impact of data partitioning on model evaluation, splitting the data into training and validation sets at a 4:1 ratio. We utilize the cross-validated Concordance Index (C-Index), with its standard deviation (std).

Cancer Subtyping We also conduct experiments on three challenging public datasets: BRACS (Brancati et al. 2022), and the NSCLC and BRCA cohorts from the TCGA database (Weinstein et al. 2013). For dataset partitioning, we follow the protocol from GPFM (Ma et al. 2024), splitting the data into training, validation, and testing sets at a 7:1:2 ratio. To ensure a robust evaluation, we generate 5 different random splits with this ratio for our experiments. For evaluation, we adopt the Area Under the Curve (AUC) and Accuracy (ACC) metrics, reporting their mean and standard deviation (std). Supplementary Material offers more details.

Comparison Methods and Training Details

Comparison Methods We compare our proposed methods CKMIL-base and CKMIL, against several categories of methods:(1) Simple Pooling Methods: Mean-Pooling and Max-Pooling; (2) Attention-Based Methods: ABMIL (Ilse, Tomczak, and Welling 2018), and its variants CLAM-MB (Lu et al. 2021) and DSMIL (Li, Li, and Eliceiri 2021); (3) Global Interaction Methods with Linear Complexity: TransMIL (Shao et al. 2021), MambaMIL (Yang, Wang, and Chen 2024), and RRTMIL (Tang et al. 2024).

Comparison of CKMIL-Base and CKMIL The distinction between our two models lies solely in the Q/K vector generation: CKMIL-Base uses conventional linear layers, while CKMIL incorporates our exploratory ICP module.

Training Details Patches of size 256×256 were cropped at 20× magnification WSIs without overlap. To extract patch features, we utilized two offline encoders: a ResNet50 (He et al. 2016) pre-trained on ImageNet (Deng et al. 2009) for general visual representations, and the UNI (Chen et al. 2024) model, which was self-supervised on a pancancer cohort to learn domain-specific pathology features. Supplementary Material offers more training details.

Results and Analysis

Tables 1 and 2 provide a comprehensive performance comparison of various MIL methods on cancer subtyping and survival prediction, utilizing two distinct feature extractors: the general-purpose ResNet50 (He et al. 2016) and the domain-specific UNI (Chen et al. 2024).

Methods	BLCA (C-index)		BRCA (C-index)		LUAD (C-index)	
	ResNet-50	UNI	ResNet-50	UNI	ResNet-50	UNI
Mean-Pooling	0.5870±0.0583	0.5989±0.0129	0.6135±0.0631	0.6777±0.0602	0.6095±0.0820	0.6276±0.0623
Max-Pooling	0.5589±0.0593	0.5742±0.0476	0.5754±0.0382	0.6119±0.0522	0.6063±0.0396	0.5951±0.0069
ABMIL (Ilse, Tomczak, and Welling 2018)	0.5503±0.0986	0.6035±0.0491	0.6103±0.0739	0.6688±0.0534	0.6015±0.0767	0.6240±0.0762
CLAM-MB (Lu et al. 2021)	0.5695±0.0951	0.5975±0.0445	0.5887±0.0592	0.6701±0.0413	0.6165±0.0761	0.6265±0.0490
DSMIL (Li, Li, and Eliceiri 2021)	0.5774±0.0588	0.5885±0.0536	0.6199±0.0297	0.6460±0.0346	0.6147±0.0250	0.5496±0.0594
TransMIL (Shao et al. 2021)	0.6055±0.0485	0.6119±0.0312	0.6158±0.0559	0.6163±0.0360	0.6335±0.0347	0.6222±0.0615
MambaMIL (Yang, Wang, and Chen 2024)	OOM	OOM	0.6524±0.0494	0.6480±0.0399	0.6452 ± 0.0168	0.6142±0.0580
RRTMIL (Tang et al. 2024)	OOM	OOM	0.6445±0.0604	0.6500±0.0503	0.6231±0.0490	0.6303±0.0687
CKMIL-Base (ours)	**0.6287±0.0429**	0.6038±0.0349	0.6440±0.0794	**0.6920±0.0717**	**0.6820±0.0267**	0.6300±0.0267
CKMIL (ours)	0.6185±0.0406	**0.6155±0.0429**	**0.6825±0.0887**	0.6869±0.0661	0.6467±0.0402	**0.6380±0.0640**

Table 2: Performance comparison on survival prediction tasks. Best results are in **bold**, and second-best results are underlined. OOM denotes out of memory in the experiment settings.

When benchmarked with the ResNet50 feature extractor, our CKMIL models achieve state-of-the-art (SOTA) performance across all tasks and datasets. Notably, the full CKMIL model consistently outperforms all competing methods, with CKMIL-Base being the only exception. For instance, as shown in Table 1, our CKMIL model demonstrates significant improvements on the BRACS-3 subtyping task, outperforming the strong baseline RRTMIL with a 2.78% improvement in AUC and 2.01% in ACC. This superiority extends to survival prediction tasks. On the LUAD cohort, our CKMIL-Base model sets a new SOTA with a C-Index of 0.6820. Meanwhile, on the BRCA survival task, CKMIL achieves a C-Index of 0.6825, a substantial 3.81% improvement over the next-best comparable method. This finding is particularly significant, as it validates our core hypothesis that an effective aggregation mechanism can overcome the limitations of non-domain-specific features by effectively modeling instance correlations.

When using the pathology-specific UNI feature extractor, our models achieve new SOTA results across all survival prediction tasks. However, in certain subtyping tasks, such as on the BRCA dataset (for both AUC and ACC) and the NSCLC dataset (for ACC), the performance of methods that model inter-instance correlations, including ours, was surpassed by simpler approaches like ABMIL and CLAM. We hypothesize that this phenomenon occurs because UNI generates features that are already highly discriminative. For such strong features, explicitly modeling correlations might introduce noise from redundant instances, which inadvertently dilutes the weights or the features themselves of sparse, critical instances, and thus degrades performance. Conversely, our models' SOTA performance with the generic ResNet50 extractor corroborates the effectiveness of our correlation modeling, demonstrating its ability to adapt general-purpose features for specialized medical analysis through guided interaction.

Ablation Study and Sensitivity Analysis

To rigorously validate the effectiveness of our proposed CKMIL framework, we conduct a series of ablation studies on its core components: the Subspace-Disentangled Attention (SDA), the Key-Instance Guided Global Attention (KGGA), and the Instance-Conv-Projection (ICP) module. We perform quantitative evaluations on the BRACS-3 cancer subtyping task (reporting mean AUC and ACC) and the TCGA-BRCA survival prediction task (reporting mean C-Index), using ResNet50 as the feature extractor and following the same experimental protocol as in the main experiments.

Effectiveness of Subspace-Disentangled Attention (SDA) The SDA module is designed to screen for key instances within multiple disentangled feature subspaces. To isolate its contribution, we conduct two sets of experiments:

- **CKMIL vs. CKMIL ($m = 1$):** We reduce the number of subspaces in the SDA module to one (i.e., $m = 1$). This variant, denoted as CKMIL ($m = 1$) or ABMIL+KGGA, replaces SDA with a single, shared attention layer akin to ABMIL, while keeping the KGGA module.
- **ABMIL (Ilse, Tomczak, and Welling 2018) vs. ABMIL+SDA:** To demonstrate that the multi-subspace scoring mechanism is inherently superior to a single attention layer, we integrate the SDA module into the standard ABMIL framework, creating ABMIL+SDA.

As presented in Table 3, ABMIL+SDA consistently surpasses ABMIL across all metrics. Similarly, CKMIL outperforms the original CKMIL ($m = 1$) across all metrics, further validating that the multi-subspace scoring design is a more effective strategy than a single shared attention layer.

Model	**BRACS-3 (AUC ↑)**	**BRACS-3 (ACC ↑)**	**BRCA (C-Index ↑)**
ABMIL	0.8004	0.6981	0.6103
ABMIL+SDA	0.8423 **(+4.19%)**	0.7074 **(+0.93%)**	0.6131 **(+0.28%)**
CKMIL ($m = 1$)	0.8454	0.7240	0.6687
CKMIL (ours)	0.8583 **(+1.29%)**	0.7370 **(+1.30%)**	0.6825 **(+1.38%)**

Table 3: Ablation study on the effectiveness of SDA.

Effectiveness of Key-Instance Guided Global Attention (KGGA) The KGGA module is premised on the principle that global interaction should be guided by key instances. We validate its efficacy through the following experiments:

- **CKMIL vs. CKMIL (Pooling):** We replace the key-instance-guided landmark selection in KGGA with a conventional mean pooling strategy to select landmarks, a method similar to that used in TransMIL.
- **ABMIL (Ilse, Tomczak, and Welling 2018) vs. ABMIL+KGGA:** To demonstrate the importance of the global interaction mechanism itself, we augment the

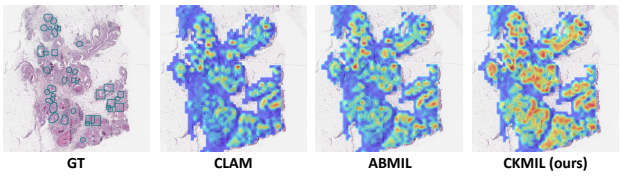

Figure 5: Global attention heatmap comparison on a WSI from the BRACS dataset.

baseline ABMIL with our KGGA module which is equivalent to the CKMIL (m=1) variant.

- **TransMIL (Shao et al. 2021) vs. TransMIL+KGGA:** To show that our key-instance-guided approach is superior, we modify TransMIL by first adding an attention layer to score instances and then using the top-scoring instances as landmarks for its global interaction. We term this variant TransMIL+KGGA.

As shown in Table 4, CKMIL significantly outperforms CKMIL (Pooling), confirming our hypothesis that using candidate key instances as landmarks is more effective than using landmarks derived from pooling. The comparison between ABMIL and ABMIL+KGGA shows that incorporating our KGGA module brings substantial performance gains across all tasks, underscoring the necessity of modeling global inter-instance correlations. Finally, TransMIL+KGGA surpasses the original TransMIL, further proving that a key-instance-guided strategy is a more powerful approach for global attention in MIL.

Model	BRACS-3 (AUC ↑)	BRACS-3 (ACC ↑)	BRCA (C-Index ↑)
ABMIL	0.8004	0.6981	0.6103
ABMIL+KGGA	0.8454 (**+4.50%**)	0.7240 (**+2.59%**)	0.6687 (**+5.84%**)
TransMIL	0.8160	0.7111	0.6158
TransMIL+KGGA	0.8297 (**+1.37%**)	0.7278 (**+1.67%**)	0.6281 (**+1.23%**)
CKMIL (Pooling)	0.8477	0.7185	0.6445
CKMIL (ours)	0.8583 (**+1.06%**)	0.7370 (**+1.85%**)	0.6825 (**+3.80%**)

Table 4: Ablation study on the effectiveness of KGGA.

Effectiveness of Instance-Conv-Projection (ICP) The ICP module is designed based on the hypothesis that local correlations exist among the components of an instance's feature vector. To investigate the feasibility of this exploratory module, we conducted a comprehensive comparison between the full CKMIL model and the CKMIL-Base model which uses standard linear projections across all tasks and datasets. The detailed results, presented in Table 1 and 2, reveal that the ICP module offers clear benefits in specific contexts. For instance, when using ResNet50 features for the BRACS subtyping task, CKMIL shows a 2.4% improvement in ACC and a 1.0% improvement in AUC over CKMIL-Base. Similarly, for survival prediction on the LUAD cohort with UNI features, CKMIL yields a 0.8% improvement in C-Index. On the TCGA-BRCA survival prediction task, the benefit is even more pronounced, with CKMIL delivering a 3.85% higher C-Index than CKMIL-Base. However, on other datasets, the impact of ICP is more varied and appears to be influenced by the choice of the upstream feature extractor. This suggests that while ICP can effectively capture latent intra-feature correlations, the prominence and utility of these correlations may depend on the specific dataset and the nature of the features generated by the encoder.

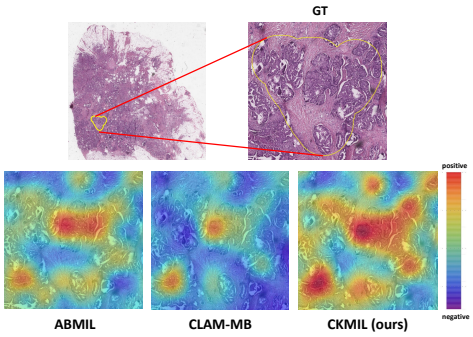

Figure 6: Local attention heatmap comparison on a WSI from the BRACS dataset.

The sensitivity analysis for key hyperparameters and the attention heatmaps of the ablation study are provided in the Supplementary Material.

Visualization Results

Fundamentally, our proposed CKMIL is an Attention-Based method. To evaluate CKMIL's interpretability and localization capability, we visualize its attention heatmaps against baseline methods ABMIL and CLAM-MB, comparing them to ground truth (GT) annotations provided by pathologists.

As shown in the global view (Figure 5), the attention from ABMIL and CLAM-MB is diffuse and highlights non-diagnostic areas, failing to localize the scattered tumor regions indicated by the GT. In contrast, CKMIL produces precise, concentrated heatmaps that show high concordance with GT annotations, successfully identifying multiple key tumor clusters. This is due to the synergy between our SDA and KGGA modules, which suppresses non-critical regions.

The superiority of CKMIL is apparent in the local view (Figure 6), while baseline methods fail to focus on core pathological cell structures, CKMIL's high-attention areas precisely cover the dense, diagnostically relevant cell regions, as confirmed by GT. This demonstrates the effectiveness of our key-instance-guided mechanism in identifying the most informative regions within a WSI.

Conclusion

In this work, we proposed CKMIL, a novel cascaded attention framework for WSI analysis that addresses the key information dilution problem in existing MIL methods. By first identifying key instances with a SDA module and then using them to guide an efficient global interaction via our KGGA module, CKMIL achieves a more focused and effective aggregation. Extensive experiments demonstrate that our approach sets a new state-of-the-art in cancer subtyping and survival prediction, proving the effectiveness of a key-instance-aware mechanism in computational pathology.

References

Amores, J. 2013. Multiple instance classification: Review, taxonomy and comparative study. *Artificial intelligence*, 201: 81–105.

Bera, K.; Schalper, K. A.; Rimm, D. L.; Velcheti, V.; and Madabhushi, A. 2019. Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. *Nature reviews Clinical oncology*, 16(11): 703–715.

Brancati, N.; Anniciello, A. M.; Pati, P.; Riccio, D.; Scognamiglio, G.; Jaume, G.; De Pietro, G.; Di Bonito, M.; Foncubierta, A.; Botti, G.; et al. 2022. Bracs: A dataset for breast carcinoma subtyping in h&e histology images. *Database*, 2022: baac093.

Cai, Z.; Song, H.; Fingerhut, A.; Sun, J.; Ma, J.; Zhang, L.; Li, S.; Yu, C.; Zheng, M.; and Zang, L. 2021. A greater lymph node yield is required during pathological examination in microsatellite instability-high gastric cancer. *BMC cancer*, 21(1): 319.

Campanella, G.; Hanna, M. G.; Geneslaw, L.; Miraflor, A.; Werneck Krauss Silva, V.; Busam, K. J.; Brogi, E.; Reuter, V. E.; Klimstra, D. S.; and Fuchs, T. J. 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8): 1301–1309.

Chen, R. J.; Ding, T.; Lu, M. Y.; Williamson, D. F.; Jaume, G.; Chen, B.; Zhang, A.; Shao, D.; Song, A. H.; Shaban, M.; et al. 2024. Towards a General-Purpose Foundation Model for Computational Pathology. *Nature Medicine*.

Chen, Z.; Chi, Z.; Fu, H.; and Feng, D. 2013. Multi-instance multi-label image classification: A neural approach. *Neurocomputing*, 99: 298–306.

Cifci, D.; Veldhuizen, G. P.; Foersch, S.; and Kather, J. N. 2023. AI in computational pathology of cancer: improving diagnostic workflows and clinical outcomes? *Annual Review of Cancer Biology*, 7(1): 57–71.

Coudray, N.; Ocampo, P. S.; Sakellaropoulos, T.; Narula, N.; Snuderl, M.; Fenyö, D.; Moreira, A. L.; Razavian, N.; and Tsirigos, A. 2018. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature medicine*, 24(10): 1559–1567.

Cui, M.; and Zhang, D. Y. 2021. Artificial intelligence and computational pathology. *Laboratory Investigation*, 101(4): 412–422.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. IEEE.

Elmore, J. G.; Longton, G. M.; Carney, P. A.; Geller, B. M.; Onega, T.; Tosteson, A. N.; Nelson, H. D.; Pepe, M. S.; Allison, K. H.; Schnitt, S. J.; et al. 2015. Diagnostic concordance among pathologists interpreting breast biopsy specimens. *Jama*, 313(11): 1122–1132.

Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Ilse, M.; Tomczak, J.; and Welling, M. 2018. Attention-based deep multiple instance learning. In *International conference on machine learning*, 2127–2136. PMLR.

Jin, C.; Guo, Z.; Lin, Y.; Luo, L.; and Chen, H. 2023. Label-efficient deep learning in medical image analysis: Challenges and future directions. *arXiv preprint arXiv:2303.12484*.

Li, B.; Li, Y.; and Eliceiri, K. W. 2021. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14318–14328.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Lu, M. Y.; Williamson, D. F.; Chen, T. Y.; Chen, R. J.; Barbieri, M.; and Mahmood, F. 2021. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6): 555–570.

Ma, J.; Guo, Z.; Zhou, F.; Wang, Y.; Xu, Y.; Li, J.; Yan, F.; Cai, Y.; Zhu, Z.; Jin, C.; et al. 2024. Towards a generalizable pathology foundation model via unified knowledge distillation. *arXiv preprint arXiv:2407.18449*.

Maron, O.; and Lozano-Pérez, T. 1997. A framework for multiple-instance learning. *Advances in neural information processing systems*, 10.

Shao, Z.; Bian, H.; Chen, Y.; Wang, Y.; Zhang, J.; Ji, X.; et al. 2021. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34: 2136–2147.

Song, A. H.; Jaume, G.; Williamson, D. F.; Lu, M. Y.; Vaidya, A.; Miller, T. R.; and Mahmood, F. 2023. Artificial intelligence for digital and computational pathology. *Nature Reviews Bioengineering*, 1(12): 930–949.

Tang, W.; Qin, R.; Fang, H.; Zhou, F.; Chen, H.; Li, X.; and Cheng, M.-M. 2025. Revisiting End-to-End Learning with Slide-level Supervision in Computational Pathology. *arXiv preprint arXiv:2506.02408*.

Tang, W.; Zhou, F.; Huang, S.; Zhu, X.; Zhang, Y.; and Liu, B. 2024. Feature re-embedding: Towards foundation model-level performance in computational pathology. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11343–11352.

Weinstein, J. N.; Collisson, E. A.; Mills, G. B.; Shaw, K. R.; Ozenberger, B. A.; Ellrott, K.; Shmulevich, I.; Sander, C.; and Stuart, J. M. 2013. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10): 1113–1120.

Xiong, Y.; Zeng, Z.; Chakraborty, R.; Tan, M.; Fung, G.; Li, Y.; and Singh, V. 2021. Nyströmformer: A nyström-based

algorithm for approximating self-attention. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 14138–14148.

Yang, S.; Wang, Y.; and Chen, H. 2024. Mambamil: Enhancing long sequence modeling with sequence reordering in computational pathology. In *International conference on medical image computing and computer-assisted intervention*, 296–306. Springer.

Yu, K.-H.; Zhang, C.; Berry, G. J.; Altman, R. B.; Ré, C.; Rubin, D. L.; and Snyder, M. 2016. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. *Nature communications*, 7(1): 12474.

Reproducibility Checklist

1. General Paper Structure

1.1. Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA) yes

1.2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no) yes

1.3. Provides well-marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes/no) yes

2. Theoretical Contributions

2.1. Does this paper make theoretical contributions? (yes/no) no

If yes, please address the following points:

2.2. All assumptions and restrictions are stated clearly and formally (yes/partial/no)

2.3. All novel claims are stated formally (e.g., in theorem statements) (yes/partial/no)

2.4. Proofs of all novel claims are included (yes/partial/no)

2.5. Proof sketches or intuitions are given for complex and/or novel results (yes/partial/no)

2.6. Appropriate citations to theoretical tools used are given (yes/partial/no)

2.7. All theoretical claims are demonstrated empirically to hold (yes/partial/no/NA)

2.8. All experimental code used to eliminate or disprove claims is included (yes/no/NA)

3. Dataset Usage

3.1. Does this paper rely on one or more datasets? (yes/no) yes

If yes, please address the following points:

3.2. A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA) yes

3.3. All novel datasets introduced in this paper are included in a data appendix (yes/partial/no/NA) NA

3.4. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no/NA) NA

3.5. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations (yes/no/NA) yes

3.6. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available (yes/partial/no/NA) yes

3.7. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing (yes/partial/no/NA) NA

4. Computational Experiments

4.1. Does this paper include computational experiments? (yes/no) yes

If yes, please address the following points:

4.2. This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting (yes/partial/no/NA) partial

4.3. Any code required for pre-processing data is included in the appendix (yes/partial/no) no

4.4. All source code required for conducting and analyzing the experiments is included in a code appendix (yes/partial/no) no

4.5. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no) yes

4.6. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes/partial/no) partial

4.7. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results (yes/par-

tial/no/NA) yes

4.8. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks (yes/partial/no) partial

4.9. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics (yes/partial/no) yes

4.10. This paper states the number of algorithm runs used to compute each reported result (yes/no) yes

4.11. Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information (yes/no) yes

4.12. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank) (yes/partial/no) no

4.13. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments (yes/partial/no/NA) yes