Rethinking Transformer for Long Contextual Histopathology Whole Slide Image Analysis

Honglin Li[1,3] **Yunlong Zhang**[1,3] **Pingyi Chen**[1,3] **Zhongyi Shui**[1,3]
Chenglu Zhu[2,3*] **Lin Yang**[2,3*]
[1] Zhejiang University
[2] Research Center for Industries of the Future and [3] School of Engineering, Westlake University
{lihonglin,zhuchenglu,yanglin}@westlake.edu.cn

Abstract

Histopathology Whole Slide Image (WSI) analysis serves as the gold standard for clinical cancer diagnosis in the daily routines of doctors. To develop computer-aided diagnosis model for histopathology WSIs, previous methods typically employ Multi-Instance Learning to enable slide-level prediction given only slide-level labels. Among these models, vanilla attention mechanisms without pairwise interactions have traditionally been employed but are unable to model contextual information. More recently, self-attention models have been utilized to address this issue. To alleviate the computational complexity of long sequences in large WSIs, methods like HIPT use region-slicing, and TransMIL employs Nyströmformer as an approximation of full self-attention. Both approaches suffer from suboptimal performance due to the loss of key information. Moreover, their use of absolute positional embedding struggles to effectively handle long contextual dependencies in shape-varying WSIs. In this paper, we first analyze how the low-rank nature of the long-sequence attention matrix constrains the representation ability of WSI modelling. Then, we demonstrate that the rank of attention matrix can be improved by focusing on local interactions via a local attention mask. Our analysis shows that the local mask aligns with the attention patterns in the lower layers of the Transformer. Furthermore, the local attention mask can be implemented during chunked attention calculation, reducing the quadratic computational complexity to linear with a small local bandwidth. Additionally, this locality helps the model generalize to unseen or under-fitted positions more easily. Building on this, we propose a local-global hybrid Transformer for both computational acceleration and local-global information interactions modelling. Our method, Long-contextual MIL (LongMIL), is evaluated through extensive experiments on various WSI tasks to validate its superiority in: 1) overall performance, 2) memory usage and speed, and 3) extrapolation ability compared to previous methods. Our code will be available at https://github.com/invoker-LL/Long-MIL.

1 Introduction

Though digital pathology images have been widely used for Cancer diagnosis [50, 65, 89, 94, 17, 44] and prognosis [9, 12, 71] and gene expression [82] via automatic computer-assisted analysis, the Giga-pixels of resolution, as large as $150,000 \times 150,000$ pixels [50, 12, 60] of Whole Slide Image (WSI), still poses great challenges on both annotation labelling and efficient computation for model training [42]. Thus, previous methods [9, 12, 42, 5, 40, 92, 93, 41] focus on developing annotation-

*Corresponding Author

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-3219

& computational- efficient learning to cope with those problems by employing Multiple Instance Learning (MIL) [51, 32] with only WSI-level supervision.

Currently, there are mainly three steps (or mainstream genres) of WSI analysis framework: 1) access better instance-level patch embedding via Self-supervised Learning [30, 7, 9, 40, 77].

2) design WSI head architectures [50, 65, 89] and train the head with frozen instance embedding. 3) fine-tune patch embedding with WSI level weak label for better task-specific results [90, 42]. Here in this paper, we focus on the step-2 and uncovering that there are still some room for improvement: Firstly, the vanilla attention used in AB-MIL, DS-MIL, CLAM, etc. [32, 40, 50, 89], despite its computational efficiency (compared to self-attention), is unable to model contextual or interaction information across instances within a WSI. These interactions, which play a crucial role in prediction decision-making [11, 65] indeed, can be modelled via self-attention mechanism. However, the long sequence of WSI instances pose $O(n^2)$ computation complexity with self-attention (Fig. 1). Although this complexity can be alleviated by self-attention approximation methods like Nyströmformer [83, 75] used in TransMIL [65], this approximation only get sub-optimal performance compared to self-attention as pointed out in [20, 29]. Authors in HIPT [9] mitigates the complexity by non-overlap large region slicing, but the interactions of instances from different region slicing are highly ignored (e.g. adjacent patches may be separated into two regions).

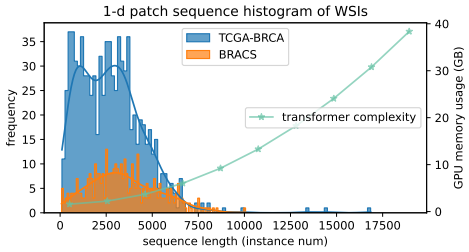

Figure 1: Handling an extremely long sequence with a magnification of $20\times$ (or quadrupling to $40\times$) poses a significant challenge. The computational complexity of transformers, denoted as $O(n^2)$, becomes prohibitive in such cases, leading to computational explosion.

The above issues highlight a strong need for an effective and efficient Transformer for WSI modelling. To begin, we discuss the performance bottleneck of basic Transformer [72] for WSI. Different to Vision Transformer [23] for natural image modelling where the number of patched tokens are smaller than patch embedding size (e.g. ViT-Small-p16 with embedding size 384 attend on 196+1 tokens), the Transformer-based WSI model suffers severe low-rank bottleneck of attention matrix [3, 22] given the long-sequence ($n >> 1024$) of WSI but limited embedding size ($d \leq 1024$). We reveal this problem in WSI theoretically, thus finding that one self-attention layer with limited embedding size can not model local contexts and global interactions at the same time. By stacking multiple self-attention layers, we notice that the low layer focus more on local context (Fig. 2a) after training while high layer focus on global. However, the rank of the attention matrix is still limited (Fig. 2a+d), resulting in constrained performance.

We assert that the low-rank bottleneck causes the attention mechanism to become confused between local and global interactions, even after training. In other words, using Q and $K^\top$ with only $2dn$ points, it's hard to model $n \times n$ interactions comprehensively in the context of WSI where $n >> d$. We believe that it would be better focusing on less interactions in one layer. Motivated by the low layers of Transformer showing highly sparse attention pattern (Fig. 2a+d) with locality, we propose a local attention mask to learn local interaction more directly. This local mask, more importantly, can highly improve the rank of the attention matrix, showing better representation ability. Furthermore, the local attention mask can be implemented during chunked attention calculation, reducing the quadratic computational complexity to linear with a small local bandwidth. In addition, this locality helps the model generalize to unseen or under-fitted positions more easily (where absolute position embedding used in methods like TransMIL may fail, see Appendix A.2 for more illustration).

Building on this, we propose a local-global hybrid Transformer for both computational acceleration and local-global information modelling.

Our main contributions can be summarized as 3 folds:

1) We firstly theoretically uncover why Transformer model for WSI-MIL analysis fails, based on the low-rank bottleneck of attention matrix for long sequence but limited embedding size. We then further analyze the sparsity and locality pattern of attention matrix empirically to hint our local attention design.

2) We convert the full self-attention into local attention which shows three advantages: higher rank for better representation ability, lighter computational complexity and extrapolation ability for shape-vary WSIs. We further combine the full self-attention for global long-range dependency after stacking layers of local attention.

3) Our WSI-analysis experiments are performed on both diagnosis and prognosis tasks on 4 WSI datasets including Breast, Stomach, Colon and Rectal Carcinoma, which show strong universality of the method and practical potential for real-world applications.

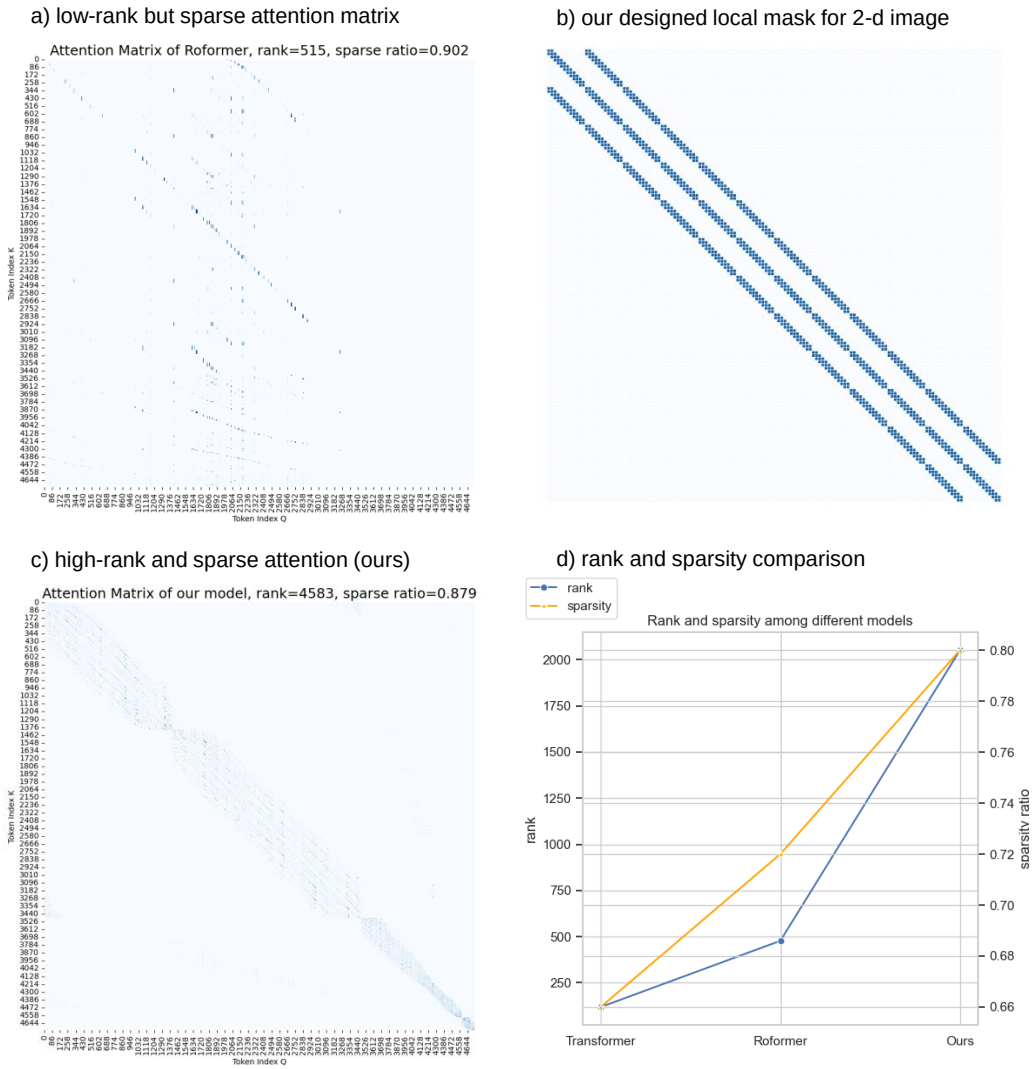

Figure 2: Rank and sparsity of attention matrix in WSI analysis.

2 Related Work

2.1 Multiple Instance Learning for WSI Analysis

Whole Slide Images (WSIs) contain a rich set of visual information that can aid in pathological analysis [6, 50]. However, accurate annotation of cell-level information within WSIs is labor-intensive and time-consuming [6, 50, 9]. To address this issue, weakly-supervised methods have gained popularity in pathology WSI analysis. Attention-based Multi-Instance Learning (AB-MIL) [32] is adopted to learns instance adaptive weights, allowing the model to focus on informative regions

https://doi.org/10.52202/079017-3219

within the WSIs. This approach significantly reduces the annotation burden of pathologists while still providing valuable insights for patient-level diagnosis. In the context of weakly-supervised pathology WSI analysis, several innovative approaches, DS-MIL, CLAM, DTFD, etc. [32, 40, 50, 89, 35, 61, 18, 2] have been proposed. However, their utilized vanilla attention with light computational cost can not model WSI contextual information, which is useful in pathologist diagnosis decision making [11, 65]. The fine-grained details and global contextual information can also be captured by multi-scale modelling [9, 40]. Graph Network [11, 43, 28, 8, 25] is also useful to make model be context-aware. Similar to this, HIPT [9] and TransMIL [65] have explored the advantages of Transformer with pairwise interactions to model this contextual information. Since Transformer can be generalized to Graph Network [24], both modelling the pairwise interaction, in this paper we focus more on Transformer and try to adapt it better to fit the shape varying and long context properties of WSI. Unlike the authors in [80] who focus on the low-rank properties of pathology images, we investigate from the perspective of low-rank in the attention matrix of Transformer.

2.2 Efficient Transformer for Long Sequence

The primary goal of this area is to alleviate the computation and memory complexity of self-attention mechanism on long sequence input. A lot of modifications sparsify the attention matrix [59, 15, 1] with some fixed patterns. Extend to this, some work [73, 74, 64] using learnable patterns in a data-driven fashion, e.g. Reformer [38] introduces a hash-based similarity measure to efficiently cluster tokens into chunks. Linformer [75] technique leverage low-rank approximations of the self-attention matrix, decomposing the $N \times N$ matrix to $N \times k$. The kernels also serve as an approximation of the attention matrix, including Performers [37], Linear Transformers [16]. Another popular method of reducing computation cost is to reduce the resolution of the sequence, hence reducing computation cost by a commensurate factor, e.g. Perceiver [33], Swin Transformer [47]. The recent Nyströmformer [83] used in TransMIL [65], can also be seem as kernel-based low-rank approach. Above work mainly focus on a light approximation of self-attention or using sparse attention, which is indeed worse than the full attention [20]. Recent work like FlashAttention [20] and others [62, 34] using chunked computation scheme and IO-aware mechanism to be memory-efficient and gain full ability like self-attention. Another lines of work try to merge RNN and Transformer, e.g. Transformer-XL [19] proposed a segment-level recurrence mechanism that connects multiple segments and blocks, and now is widely used in most successful LLMs [52, 91, 69]. Recently, linear RNNs [88, 54, 53, 27] and its variants [21, 57] are also proposed, but these recurrent ability is designed for 1-d sequence with causal or auto-regressive property, not fit well for image recognition. To fit longer sequence, better positional embeddings like RoPE, ALiBi, etc. [66, 58, 14] are also proposed. Different to these work focus on NLP task, here in this paper we try to build an efficient Transformer for WSI analysis, which is a unique challenge in vision task.

3 Method

3.1 Preliminary: Attention-based WSI Analysis

Given a WSI X as input, the goal is to make slide-level prediction $\hat{Y}$ by learning a classifier $f(X;\theta)$. X is firstly patched into a long sequence of small instances $X = \{x_1, ..., x_n\}$ because of its extremely large resolution, where n is the number of instance. The slide-level supervision Y is given by doctors who consider the latent label y_i of all instance x_i. Most previous work [6, 50, 9] try to model this process by a Max-pooling operation, so initially, this annotation process is treated as:

$$Y = \max\{y_1, ..., y_n\}. \tag{1}$$

Since the end-to-end training from raw image input to WSI-level output is infeasible because of large memory cost, conventional approaches convert it into two separate stages: Firstly, convert all small patches into instance embeddings $Z = \{z_1, ..., z_n\}$ by a pre-trained backbone such as ResNet [31] or ViT [79], which refers to general features from public ImageNet, or learned on the related dataset to extract the domain-specific representations [9, 36]. Then, aggregate all patches' features within a slide and producing the slide-level prediction $\hat{Y} = g(Z;\theta)$. In this paper, we mainly focus on the latter one, where g is an vanilla attention function followed by a linear classifier head as:

$$\hat{Y} = \sigma(\sum_{i=1}^{n} a_i z_i), \tag{2}$$

where a_i is attention weights and $\sigma(\cdot)$ is a linear head.

However, above vanilla attention method assigning adaptive weight to each instance to make simple summation or pooling can not model the interactions among different instances. Thus, to handle this problem, Transformer with self-attention is employed in TransMIL [65] and HIPT [9], where the attention sublayer computes the attention scores for the i-th query $q_i \in R^{1\times d}$, $(1 \leq i \leq n)$ in each head, where d is the head dimension. In other words, each instance will compute an attention score list as interactions with all instances. These attention scores are then multiplied by the values to return the output of the attention sub-layer as:

$$o_i = \text{softmax}(q_i K^\top)V, \tag{3}$$

where the $\{Q : q_i, K : k_i, V : v_i\} \in R^{n\times d}$ are obtained through linear transform from the input embedding Z, the softmax$(q_i K^\top)$ is the attention score and $O \in R^{n\times d}$ is the output. Given O, which encodes the interactions among instances, we can further use Equation (2) and input O to replace Z for final prediction, mean-pooling and class token in ViT [79] can also be adopted. Note that here we omit dropout, FFN, residual connection and some detailed blocks in Transformer for simplicity.

Positional Embedding: Since the operation in Equation (3) is position-agnostic, Transformer [72, 23] try to model contextual interactions by incorporate position information. Absolute positional embedding assigns a positional vector p_m to each position m and adds it to the embedding vector as: $z_i = z_i + p_{m,i}$. In HIPT [9] , the absolute positional embedding [23] for 2-d is employed, while TransMIL [65] use convolutions as implicit positional embedding [78] but treat data as 1-d sequence. Relative positional embedding that model the positional difference $m-n$ has become popular. Rotary positional embedding (RoPE) [66] encodes the position with rotations: $f(q_m, m) = R_m q_m$, where R_m is a rotation matrix with angles proportional to m. With the rotation's property, the query-key product exhibits a positional difference:

$$f(q_m, m) f(k_n, n)^\top = q_m R_{n-m} k_n^\top. \tag{4}$$

The core idea of RoPE is to insert position m, n signal on q, k and reflect the relative position on the newly attention matrix. Though the RoPE is designed for 1-d language sequence, it can also be extended to 2-d paradigm for application on WSI analysis [56].

Computational Complexity: Though above Transformer with self-attention can well model the interactions among instances, its computational cost $O(n^2 d)$ is too heavy for long sequence of WSI due to the interactive attention score calculation (see Appendix A.5.5 for 40x magnification WSI modelling). Previous WSI Transformer SOTA like TransMIL [65] and HIPT [9] relieve this problem with different ways: 1) attention approximation: TransMIL [65] utilizes Nyströmformer [83], a mechanism employs kernel-based low-rank approximation to approximate full self-attention for acceleration. 2) region slicing: HIPT [9] utilizes the locality of image by slicing WSI into 4096×4096 squares without overlapping. Given the fixed window size, in each square there are fixed $16*16 = 256$ patches with shape of 256×256, thus the computational cost can also be seen as linear complexity.

3.2 Low-rank and Sparsity of Attention Matrix for Long-sequence WSI

Low-rank bottleneck: Considering all queries in $\{Q : q_i\}$, the Equation (3) can be seen as:

$$O = \text{softmax}(QK^\top)V, \tag{5}$$

where the rank of $QK^\top$ can be derived as:

$$r(Q_{n\times d}(K^\top)_{d\times n}) \leq \min(r(Q_{n\times d}), r(K_{n\times d})) = \min(n, d). \tag{6}$$

In the context of WSI analysis, the patched sequence length n of most WSIs is larger than 1024 (Fig. 1), while the embedding size d of the pre-trained patch encoder is less than 1024 (e.g. 1024 in ResNet-50, 384 in ViT-Small, 768 in ViT-Base). Thus, in Equation 6, we have $d \leq 1024 \leq n$, indicating that the rank of the attention matrix in Transformer-based WSI analysis is constrained by the embedding size d:

$$r(QK^\top) \leq \min(n, d) = d. \tag{7}$$

As a result, the representation ability of self-attention is limited by the low-rank bottleneck, thus vanilla Transformer based model in WSI analysis suffers sub-optimal performance. Though the

 https://doi.org/10.52202/079017-3219

non-linear softmax operation can change the rank, but we still observe limited rank of softmax$(QK^\top)$ after training (Fig. 2a+d). This is an extremely different problem compared to ViT modelling, e.g. ViT-Small with embedding size of 384 only need to focus on 196 patch tokens (with image size of 224 and patch size of 16), which can model both local contextual and global interactions simultaneously with full-rank attention matrix. We contend that, under this circumstance, Transformer for WSI modelling may become confused when handling local contextual and global interactions with a single layer.

Under this assumption, it is also more easy to understand the limitation of previous SOTA transformer for WSI: TransMIL [65] with Nyströmformer [83] employs kernel-based low-rank approximation to approximate full self-attention. However, it is worth noting that the approximation may produce lower rank of attention matrix compared to basic self-attention, thus resulting lower performance in TransMIL. Similar problems also happens in other models like softmax-free linear attention [54, 67, 85]. We show a lot of experiment of these linear attention model in Appendix A.5.6.

An intuitive modification to handle the low-rank problem is to set a larger embedding size d, but this makes computational complexity $O(n^2d)$ more severe, let alone most pathology patch pre-trained foundation models [36, 49, 10] carry fixed embedding size. Noting that it is infeasible to fully represent a matrix with a shape of $n \times n$ if $n >> d$ given totally $2nd$ feature points from $Q_{n\times d}$ and $K_{n\times d}$, why not focus the attention to more important interactions? Thus in the contrast, we alleviate the low-rank bottleneck together with the problem of computational cost by focusing on locality, motivated by the sparse and local pattern in low layers of attention (Fig. 2a).

Sparsity with locality: Here we define the the selection index for retained sparse attention matrix when the softmax probability is greater than a threshold:

$$I = \text{where}\{\text{softmax}(QK^\top) > \tau\}, \tag{8}$$

where the threshold τ is normalized by sequence number n, e.g. $\tau = 0.0001/n$. Then, the sparsity ratio can be denoted as:

$$r = 1 - \frac{\text{len}(I)}{n^2}. \tag{9}$$

We note that there is an obvious sparsity and local pattern in lower Transformer layers (Fig. 2a) under this protocol. Given the learned locality and sparsity, we introduce to learn local contextual information with a addable local attention mask for $A = \text{softmax}(QK^\top)$ (without loss of generality, here we mainly derive on 1-d sequence for simplicity):

$$\text{mask}_{i,j} = \begin{cases} -inf & \text{if } |i-j| > b, \\ 0 & \text{otherwise}, \end{cases} \tag{10}$$

where b is the local band width. Then the attention matrix is converted as:

$$A_{i,j} = \text{softmax}(QK^\top + \text{mask}) = \begin{cases} 0 & \text{if } |i-j| > b, \\ p & \text{otherwise}, \end{cases} \tag{11}$$

where $0 < p < 1$ is the softmax probability.

We claim that our proposed sparse local attention capture both higher rank and reduced computational complexity, which will work better for WSI modelling:

1) Higher rank: It is easy to prove that the band matrix A in Equation 11 is of a lower-bound of rank as $n - b$. Here we give a intuitive verification with a small band matrix:

$$A^{<n=9,b=3>} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & 0 & 0 & 0 & 0 & 0 \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} & 0 & 0 & 0 & 0 \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} & a_{36} & 0 & 0 & 0 \\ \underline{a_{41}} & a_{42} & a_{43} & a_{44} & a_{45} & a_{46} & a_{47} & 0 & 0 \\ 0 & a_{52} & a_{53} & a_{54} & a_{55} & a_{56} & a_{57} & a_{58} & 0 \\ 0 & 0 & a_{63} & a_{64} & a_{65} & a_{66} & a_{67} & a_{68} & a_{69} \\ 0 & 0 & 0 & a_{74} & a_{75} & a_{76} & a_{77} & a_{78} & a_{79} \\ 0 & 0 & 0 & 0 & a_{85} & a_{86} & a_{87} & a_{88} & a_{89} \\ 0 & 0 & 0 & 0 & 0 & \underline{a_{96}} & a_{97} & a_{98} & a_{99} \end{bmatrix}. \tag{12}$$

Then, let's consider the lower-left sub-matrix ranging from $(b+1, 1)$ to $(n, n-b)$:

$$A_{sub} = \begin{bmatrix} a_{(b+1)1} & a_{(b+1)2} & \cdots & a_{(b+1)(n-b)} \\ 0 & a_{(b+2)2} & \cdots & a_{(b+2)(n-b)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{n(n-b)} \end{bmatrix}, \tag{13}$$

which is apparently a upper triangular matrix with a full rank of $n-b$. Thus, we have:

$$rank(A) \geq rank(A_{sub}) = n - b. \tag{14}$$

Since $n > 1024 >> b$ practically, our model with higher rank carry stronger representation ability which can focus more on local contextual information.

2) Reduced computational complexity: Given a larger n but fixed small b in Equation 12, there will be a lot of zeros in upper-right and lower-left areas. We note that these zeros can be omitted during attention matrix calculation by a chunking method [20]. Since most operations in softmax$(QK^\top)$ can be skipped, the modified complexity $O(bnd)$ will linearly related to sequence length and band-width, which is heavily reduced compared to $O(n^2 d)$.

3) Extrapolation ability: Despite above advantages, here we further show that our model can tackle WSIs with varying input shape better, in other words, extrapolation ability. RoPE need to be well trained or fine-tuned on unseen or seldom seen longer length [45, 13, 81]. Another strategy Attention with Linear Bias (ALiBi) adds pre-defined bias term after the query-key dot-product attention matrix before softmax. For the original 1-d ALiBi [58], the bias is a static, non-learned matrix softmax$(q_i k_j^\top - \rho\,|i-j|$ computed by the distance between tokens from different positions. We also introduce 2-d ALiBi by 2-d Euclidean distance among token positions, and we find that it shows similar pattern (Appendix A.3) compared to our 2-d local attention but it needs to focus on all instances.

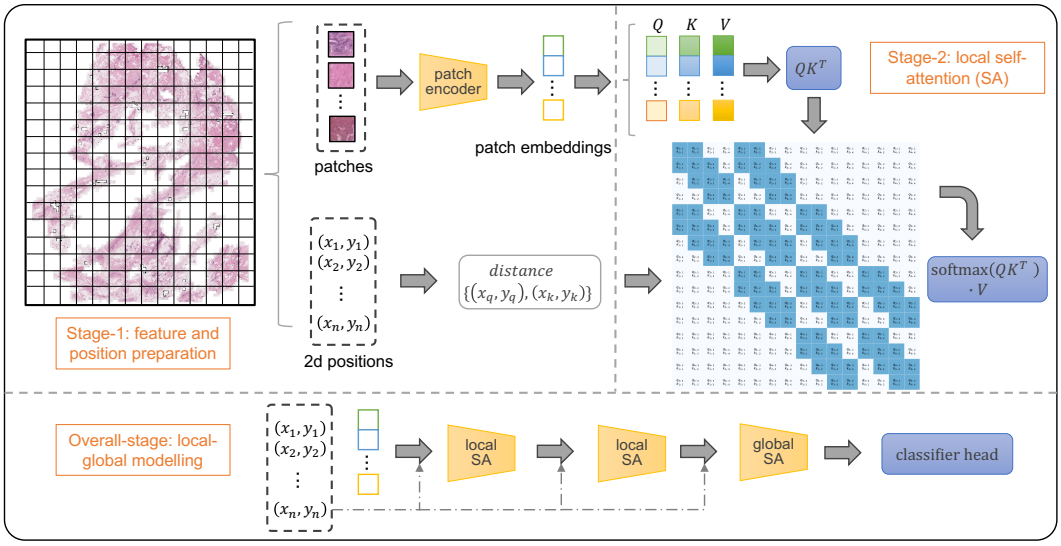

Figure 3: LongMIL framework for WSI local-global spatial contextual information interaction and fusion. 1) Preparing patch feature embedding and 2-d positions of WSIs. 2) Performing pairwise computations among all positions within a WSI by local masking as acceleration. 3) Overall local-global forward of the model, where position information need to be feed to both local (local masking) and global (positional embedding).

3.3 LongMIL framework and implementation

To realize long contextual MIL modelling and better WSI analysis performance, the overall framework (as depicted in Fig. 3) of our method includes 3 stages:

1) Segment and patch WSI into instances, then save its corresponding foreground patch feature embedding and 2-d positions for preparation.

https://doi.org/10.52202/079017-3219

2) Calculate the local self-attention matrix by the local window mask given position distance. This process is finished by trunk method like FlashAttention [20] to omit non-masking areas.

3) After multi layers (two as default) of local attention focusing on local contextual information interactions, a pooling function with window size 2×2 is employed to reduce token number by 4 times. Then a basic self-attention focus on global interactions is computed to get final feature and prediction.

4 Experiments

In this section, we present the performance of the proposed method and compare it with various baselines. Ablation experiments are performed to further study the proposed method, for paper length, more experimental results are presented in the Appendix A.5.

Datasets and Tasks. We use four datasets to evaluate our method for both **tumor subtyping** and **survival prediction**. For data details and pre-processing, please see Appendix A.4.

Table 1: **Slide-Level Tumor Subtyping** on BRACS by using two pre-trained embeddings. **Top Rows.** Various WSI-MIL architectures with vanilla attention (no interaction among different instances). **Bottom Rows.** TransMIL (using Nyströmformer and learnable absolute position embedding), full attention (+RoPE) and our LongMIL.

	BRACS tumor subtyping			
	ViT-S Lunit [36]		ViT-S DINO (our pre-train)	
Method	F1	AUC	F1	AUC
KNN (Mean)	0.503±0.011	0.691±0.007	0.430±0.029	0.649±0.008
KNN (Max)	0.472±0.009	0.771±0.018	0.416±0.019	0.645±0.007
Mean-pooling	0.534±0.026	0.741±0.017	0.487±0.034	0.717±0.020
Max-pooling	0.649±0.032	0.843±0.018	0.598±0.032	0.818±0.006
AB-MIL [32]	0.668±0.032	0.866±0.016	0.621±0.048	0.837±0.035
DS-MIL [40]	0.607±0.044	0.824±0.028	0.622±0.063	0.808±0.033
CLAM-SB [50]	0.647±0.020	0.836±0.021	0.627±0.032	0.836±0.009
DTFD-MIL MaxS [89]	0.597±0.025	0.874±0.026	0.521±0.059	0.807±0.016
DTFD-MIL AFS [89]	0.608±0.083	0.869±0.018	0.538±0.053	0.824±0.011
TransMIL [65]	0.648±0.054	0.835±0.031	0.591±0.049	0.798±0.029
Full Attention	0.689±0.036	0.870±0.010	0.648±0.028	0.839±0.018
LongMIL (ours)	**0.706**±**0.025**	**0.888**±**0.019**	**0.657**±**0.026**	**0.848**±**0.004**

Pre-training Patch Encoders. Our work mainly focus on the WSI-head results based on some good pre-trained encoders for histopathology including HIPT [9], Lunit [36] and newly foundation models like UNI [10] and GigaPath [84]. We also include ResNet-50 pretrained in ImangeNet-1k and ViT-small pretrained in BRACS patch data by ourself with DINO [7].

Implementation Details. We train our model with PyTorch on a RTX-3090 GPU, with a WSI-level batchsize of 1, learning rate of 1e-4, and weight decay of 1e-2. We add positional encoding into the framework, please check our code for details.

4.1 Slide-level Tumor Subtyping

Evaluation Metrics. For all the experiments, we report the macro-AUC and macro-F1 scores since all these dataset suffering class imbalance. For TCGA-BRCA, we perform 10-fold cross-validation with the same data split adopted in HIPT [9]. Besides, the dataset BRACS is officially split into training, validation and testing, thus the experiment is conducted 5-times with different random seeds. The mean and standard variance values of performance metrics are reported for multi-runs or cross-validation runs.

Baselines for Comparison. We first show the results of Mean-/Max- pooling and KNN for traditional evaluation. Then we directly evaluate several classical WSI-MIL methods, including AB-MIL [32],

DS-MIL [40], CLAM [50], DTFD-MIL [89]. Then we compare our method with Full Attention (RoFormer) and TransMIL [65]. We omit HIPT [9] for BRACS since it need WSI larger than a threshold and should based on their pre-trained backbone.

Results Analysis: For BRACS 3-categories tumor subtyping, the results are reported in Table 1. We can first observe that both Full Attention and our LongMIL show improvement respectively. For Full Attention, attributing to its full self-attention for pairwise interaction ability, it shows better performance compared to all vanilla attention modules [32, 40, 50] and especially TransMIL [65] which use attention approximation, but it is not quite stable to beat DTFD [89].

For TCGA-BRCA 2-categories tumor subtyping, we show the results in the Appendix A.5.1.

4.2 Slide-level Survival Prediction

Table 2: **Slide-Level Survival Prediction** based on HIPT [9] pre-trained embedding with various WSI-MIL architectures including vanilla attention, GCN, TransMIL, self-attention (HIPT with region slicing and absolute embedding), full self-attention and our LongMIL.

Method	COADREAD	STAD	BRCA
AB-MIL [32]	0.566±0.075	0.562±0.049	0.549±0.057
AMISL [86]	0.561±0.088	0.563±0.067	0.545±0.071
DS-MIL [40]	0.470±0.053	0.546±0.047	0.548±0.058
GCN-MIL [43]	0.538±0.049	0.513±0.069	-
HIPT [9]	0.608±0.088	0.570±0.081	-
TransMIL [65]	0.597±0.134	0.564±0.080	0.587±0.063
Full Attention	0.603±0.048	0.568±0.074	0.601±0.047
LongMIL (ours)	**0.624±0.057**	**0.589±0.066**	**0.619±0.053**

Evaluation Metrics. For all the experiments, C-Index scores are reported for the 3 datasets. We follow the data splits and pre-trained patch embedding provided in HIPT [9] for fair comparison. The performance results are also reported via the mean and standard variance values of performance metrics by multiple folder cross-validation with the same running setting to HIPT [9].

Baselines for Comparison. For this task, we use the survival cross-entropy loss proposed by Zadeh et al. [87]. The results are summarized in Table 2, where we directly evaluate several survival prediction WSI-MIL methods, including AB-MIL [32], AMISL [86], DS-MIL [40], GCN-MIL [43]. Then we compare our method with some state-of-the-art combining position embedding on Transformer: TransMIL [65] and HIPT [9]. Though our method show some improvement, the C-index score is still too low to daily clinical usage depending on only WSI information. In the near future, we would like to investigate more on this task, e.g. combining multi-modality features as used in [12], since Transformer also born with great ability on multi-modality fusion [39, 70, 12, 63].

4.3 Evaluation on Pathology Foundation Models

Since recent Pathology Foundation Model (PathFMs) [10, 48, 84] have been emerging as strong patch encoders, we here further provide evaluations based on PathFMS including UNI [10] and GigaPath [84]. The pre-processing procedure is the same to previous sections. Since the WSI params are pre-trained in GigaPath, we also experiment it using random initialization for fair comparison. For the mismatch of UNI patch encoder and GigaPath WSI head, we add a nn.Linear layer as a feature projector. The results is shown in Table 3, we find that our method also show consistency improvement with PathFMs. Furthermore, we find that the pre-training plays a key role to the success of Prov-GigaPath WSI-head, since transformers are much more over-parameterized than previous simple attention-based MIL. However, the WSI-level pretrained model relies on patch encoder (e.g. UNI patch encoder + GigaPath WSI do not show competing result). We also provide more difference analysis of efficient attention mechanism compared to GigaPath WSI head in Appendix A.6. In table 4, we also include survival prediction based on PathFMs.

 https://doi.org/10.52202/079017-3219

Table 3: Slide-Level Tumor Subtyping on BRACS based on Pathology Visual Foundation Models.

Method	BRACS tumor subtyping			
	UNI [10]		GigaPath [84]	
	F1	AUC	F1	AUC
AB-MIL [32]	0.692±0.033	0.875±0.020	0.640±0.022	0.837±0.010
CLAM-SB [50]	0.640±0.057	0.844±0.025	0.624±0.023	0.826±0.014
DTFD-MIL [89]	0.655±0.031	0.878±0.022	0.610±0.032	0.843±0.017
TransMIL [65]	0.592±0.036	0.859±0.023	0.599±0.058	0.838±0.048
Full Attention	0.715±0.043	0.884±0.017	0.663±0.023	0.850±0.018
GigaPath-random init	0.648±0.041	0.837±0.033	0.627±0.038	0.808±0.038
GigaPath-pretrained	0.668±0.026	0.861±0.030	**0.677±0.033**	**0.862±0.034**
LongMIL (ours)	**0.728±0.045**	**0.887±0.008**	0.673±0.023	0.856±0.015

Table 4: TCGA-BRCA Survival Prediction based on Pathology Visual Foundation Models.

Method	UNI [10]	GigaPath [84]
AB-MIL [32]	0.630±0.054	0.635±0.033
AMISL [86]	0.627±0.080	0.620±0.040
DS-MIL [40]	0.616±0.034	0.612±0.086
TransMIL [65]	0.598±0.059	0.599±0.064
Full Attention	0.638±0.056	0.617±0.069
LongMIL (ours)	**0.656±0.061**	**0.645±0.055**

4.4 Further Experiments and Ablations

We also provide abundant ablations in Appendix A.5.3 to select best setting including: Transformer blocks and multi-head number, dropout ratio, weight decay, learning rate and the local window size. For linear attention result, please check Appendix A.5.6 with main findings that the linear attentions show low performance like TransMIL since it get low-rank problem more easily. For linear RNN method like Mamba, the result is also relatively lower than Transformer since 2-d WSIs do not hold causal property like 1-d data. For extrapolation validation, please check Appendix A.2 where our method show significant performance improvement (p-value ≈ 0.1). We also find that our method show consistency improvement when equipped on different magnification (e.g. 40x) and patch size (224 -> 448) as shown in Appendix A.5.4.

5 Conclusions and Limitations

In conclusion, our work introduces advancements in computer-aided diagnosis for histopathology WSI analysis. By analyze the low-rank bottleneck and sparsity property and proposing a local-global hybrid Transformer model, our method, Long-contextual MIL (LongMIL), demonstrates superior performance in handling large and shape-varying WSIs. The evaluations across various tasks highlight its accuracy, extrapolation ability, and efficiency compared to previous methods. Our contributions enhance WSI analysis and provide valuable insights for future research. The LongMIL has two limitations: First, its application is restricted to very large image via Transformer modelling. Second, limited embedding size is adopted for practical aim and fair comparison, which is the key to stronger performance based on the low-rank assumption. Future work will aim to address these limitations.

6 Acknowledgements

This study was partially supported by Zhejiang Provincial Natural Science Foundation of China (Grant no. XHD23F0201), the National Natural Science Foundation of China (Grant no. 92270108), and the Research Center for Industries of the Future (RCIF) at Westlake University.

References

[1] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *Proceedings of EMNLP*, 2020.

[2] Benjamin Bergner, Christoph Lippert, and Aravindh Mahendran. Iterative patch selection for high-resolution image recognition. In *The Eleventh International Conference on Learning Representations*, 2023.

[3] Srinadh Bhojanapalli, Chulhee Yun, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. Low-rank bottleneck in multi-head attention models. In *International conference on machine learning*, pages 864–873. PMLR, 2020.

[4] 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, 2021.

[5] 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.

[6] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019.

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *CoRR*, abs/2104.14294, 2021.

[8] Tsai Hor Chan, Fernando Julio Cendra, Lan Ma, Guosheng Yin, and Lequan Yu. Histopathology whole slide image analysis with heterogeneous graph representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15661–15670, 2023.

[9] Richard J. Chen and et al. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *CVPR*, pages 16144–16155, June 2022.

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

[11] Richard J. Chen, Ming Y. Lu, Muhammad Shaban, Chengkuan Chen, Tiffany Y. Chen, Drew F. K. Williamson, and Faisal Mahmood. Whole slide images are 2d point clouds: Context-aware survival prediction using patch-based graph convolutional networks, 2021.

[12] Richard J. Chen, Ming Y. Lu, Wei-Hung Weng, Tiffany Y. Chen, Drew F.K. Williamson, Trevor Manz, Maha Shady, and Faisal Mahmood. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4025, October 2021.

[13] Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. Longlora: Efficient fine-tuning of long-context large language models. *arXiv preprint arXiv:2309.12307*, 2023.

[14] Ta-Chung Chi, Ting-Han Fan, Peter J Ramadge, and Alexander Rudnicky. Kerple: Kernelized relative positional embedding for length extrapolation. *Advances in Neural Information Processing Systems*, 35:8386–8399, 2022.

[15] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *arXiv:1904.10509*, 2019.

 https://doi.org/10.52202/079017-3219

[16] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Jared Davis, Tamas Sarlos, David Belanger, Lucy Colwell, and Adrian Weller. Masked language modeling for proteins via linearly scalable long-context transformers. *Proceedings of ICLR*, 2020.

[17] Yufei CUI, Ziquan Liu, Yixin CHEN, Yuchen Lu, Xinyue Yu, Xue Liu, Tei-Wei Kuo, Miguel R. D. Rodrigues, Chun Jason Xue, and Antoni B. Chan. Retrieval-augmented multiple instance learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[18] 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. In *The Eleventh International Conference on Learning Representations*, 2023.

[19] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. In *ACL*, 2019.

[20] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*, 2022.

[21] Tri Dao, Daniel Y Fu, Khaled K Saab, Armin W Thomas, Atri Rudra, and Christopher Ré. Hungry hungry hippos: Towards language modeling with state space models. *arXiv preprint arXiv:2212.14052*, 2022.

[22] Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth, 2023.

[23] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[24] Vijay Prakash Dwivedi and Xavier Bresson. A generalization of transformer networks to graphs. *arXiv preprint arXiv:2012.09699*, 2020.

[25] Olga Fourkioti, Matt De Vries, and Chris Bakal. CAMIL: Context-aware multiple instance learning for cancer detection and subtyping in whole slide images. In *The Twelfth International Conference on Learning Representations*, 2024.

[26] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[27] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *The International Conference on Learning Representations (ICLR)*, 2022.

[28] Yonghang Guan, Jun Zhang, Kuan Tian, Sen Yang, Pei Dong, Jinxi Xiang, Wei Yang, Junzhou Huang, Yuyao Zhang, and Xiao Han. Node-aligned graph convolutional network for whole-slide image representation and classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18813–18823, 2022.

[29] Dongchen Han, Xuran Pan, Yizeng Han, Shiji Song, and Gao Huang. Flatten transformer: Vision transformer using focused linear attention, 2023.

[30] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *CoRR*, abs/2111.06377, 2021.

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[32] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2127–2136. PMLR, 10–15 Jul 2018.

[33] Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver: General perception with iterative attention. *arXiv preprint arXiv:2103.03206*, 2021.

[34] Hanhwi Jang, Joonsung Kim, Jae-Eon Jo, Jaewon Lee, and Jangwoo Kim. Mnnfast: A fast and scalable system architecture for memory-augmented neural networks. In *Proceedings of the 46th International Symposium on Computer Architecture*, pages 250–263, 2019.

[35] Syed Ashar Javed, Dinkar Juyal, Harshith Padigela, Amaro Taylor-Weiner, Limin Yu, and aaditya prakash. Additive MIL: Intrinsically interpretable multiple instance learning for pathology. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[36] Mingu Kang, Heon Song, Seonwook Park, Donggeun Yoo, and Sérgio Pereira. Benchmarking self-supervised learning on diverse pathology datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3344–3354, June 2023.

[37] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. *arXiv preprint arXiv:2006.16236*, 2020.

[38] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020.

[39] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, pages 201–216, 2018.

[40] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, 2021.

[41] Honglin Li, Yusuan Sun, Chenglu Zhu, Yunlong Zhang, Shichuan Zhang, Zhongyi Shui, Pingyi Chen, Jingxiong Li, Sunyi Zheng, Can Cui, et al. Large-scale cervical precancerous screening via ai-assisted cytology whole slide image analysis. *arXiv preprint arXiv:2407.19512*, 2024.

[42] Honglin Li, Chenglu Zhu, Yunlong Zhang, Yuxuan Sun, Zhongyi Shui, Wenwei Kuang, Sunyi Zheng, and Lin Yang. Task-specific fine-tuning via variational information bottleneck for weakly-supervised pathology whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7454–7463, June 2023.

[43] Ruoyu Li, Jiawen Yao, Xinliang Zhu, Yeqing Li, and Junzhou Huang. Graph cnn for survival analysis on whole slide pathological images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 174–182. Springer, 2018.

[44] Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard J. Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, Ruslan Salakhutdinov, and Louis-Philippe Morency. Quantifying & modeling multimodal interactions: An information decomposition framework. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[45] Xiaoran Liu, Hang Yan, Shuo Zhang, Chenxin An, Xipeng Qiu, and Dahua Lin. Scaling laws of rope-based extrapolation. *arXiv preprint arXiv:2310.05209*, 2023.

[46] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*, 2024.

[47] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

 https://doi.org/10.52202/079017-3219

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

[49] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Andrew Zhang, Long Phi Le, et al. Towards a visual-language foundation model for computational pathology. *arXiv preprint arXiv:2307.12914*, 2023.

[50] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, 2021.

[51] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *Advances in neural information processing systems*, 10, 1997.

[52] OpenAI. Gpt-4 technical report. *arXiv*, 2023.

[53] Antonio Orvieto, Samuel L Smith, Albert Gu, Anushan Fernando, Caglar Gulcehre, Razvan Pascanu, and Soham De. Resurrecting recurrent neural networks for long sequences. *arXiv preprint arXiv:2303.06349*, 2023.

[54] Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Huanqi Cao, Xin Cheng, Michael Chung, Matteo Grella, Kranthi Kiran GV, et al. Rwkv: Reinventing rnns for the transformer era. *arXiv preprint arXiv:2305.13048*, 2023.

[55] Nicholas A Petrick, Shazia Akbar, Kenny HH Cha, Sharon Nofech-Mozes, Berkman Sahiner, Marios A Gavrielides, Jayashree Kalpathy-Cramer, Karen Drukker, Anne LL Martel, et al. Spie-aapm-nci breastpathq challenge: an image analysis challenge for quantitative tumor cellularity assessment in breast cancer histology images following neoadjuvant treatment. *Journal of Medical Imaging*, 8(3):034501, 2021.

[56] Etienne Pochet, Rami Maroun, and Roger Trullo. Roformer for position aware multiple instance learning in whole slide image classification, 2023.

[57] Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. Hyena hierarchy: Towards larger convolutional language models. *arXiv preprint arXiv:2302.10866*, 2023.

[58] Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*, 2021.

[59] Jiezhong Qiu, Hao Ma, Omer Levy, Scott Wen-tau Yih, Sinong Wang, and Jie Tang. Blockwise self-attention for long document understanding. *arXiv preprint arXiv:1911.02972*, 2019.

[60] Linhao Qu, xiaoyuan Luo, Kexue Fu, Manning Wang, and Zhijian Song. The rise of AI language pathologists: Exploring two-level prompt learning for few-shot weakly-supervised whole slide image classification. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[61] Linhao Qu, xiaoyuan Luo, Manning Wang, and Zhijian Song. Bi-directional weakly supervised knowledge distillation for whole slide image classification. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[62] Markus N. Rabe and Charles Staats. Self-attention does not need $o(n^2)$ memory, 2022.

[63] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[64] Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier. Efficient content-based sparse attention with routing transformers. *Proceedings of TACL*, 2020.

[65] 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 M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 2136–2147. Curran Associates, Inc., 2021.

[66] Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2021.

[67] Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive network: A successor to transformer for large language models. *arXiv preprint arXiv:2307.08621*, 2023.

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

[69] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[70] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019.

[71] CHAO TU, YU ZHANG, and Zhenyuan Ning. Dual-curriculum contrastive multi-instance learning for cancer prognosis analysis with whole slide images. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[72] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[73] Apoorv Vyas, Angelos Katharopoulos, and François Fleuret. Fast transformers with clustered attention. *Proceedings of NeurIPS*, 2020.

[74] Shuohang Wang, Luowei Zhou, Zhe Gan, Yen-Chun Chen, Yuwei Fang, Siqi Sun, Yu Cheng, and Jingjing Liu. Cluster-former: Clustering-based sparse transformer for long-range dependency encoding. *Proceedings of ACL-IJCNLP (Findings)*, 2020.

[75] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

[76] Wenhui Wang, Shuming Ma, Hanwen Xu, Naoto Usuyama, Jiayu Ding, Hoifung Poon, and Furu Wei. When an image is worth 1,024 x 1,024 words: A case study in computational pathology. *arXiv preprint arXiv:2312.03558*, 2023.

[77] Xiyue Wang, Jinxi Xiang, Jun Zhang, Sen Yang, Zhongyi Yang, Ming-Hui Wang, Jing Zhang, Yang Wei, Junzhou Huang, and Xiao Han. SCL-WC: Cross-slide contrastive learning for weakly-supervised whole-slide image classification. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[78] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22–31, 2021.

[79] Kan Wu, Houwen Peng, Minghao Chen, Jianlong Fu, and Hongyang Chao. Rethinking and improving relative position encoding for vision transformer. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10013–10021, 2021.

 https://doi.org/10.52202/079017-3219

[80] Jinxi Xiang and Jun Zhang. Exploring low-rank property in multiple instance learning for whole slide image classification. In *The Eleventh International Conference on Learning Representations*, 2023.

[81] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023.

[82] Ronald Xie, Kuan Pang, Sai W Chung, Catia Perciani, Sonya MacParland, BO WANG, and Gary Bader. Spatially resolved gene expression prediction from histology images via bi-modal contrastive learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[83] Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. Nyströmformer: A nyström-based algorithm for approximating self-attention. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[84] Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, pages 1–8, 2024.

[85] Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. *arXiv preprint arXiv:2312.06635*, 2023.

[86] Jiawen Yao, Xinliang Zhu, Jitendra Jonnagaddala, Nicholas Hawkins, and Junzhou Huang. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical Image Analysis*, 65:101789, 2020.

[87] Shekoufeh Gorgi Zadeh and Matthias Schmid. Bias in cross-entropy-based training of deep survival networks. *IEEE transactions on pattern analysis and machine intelligence*, 43(9):3126–3137, 2020.

[88] Shuangfei Zhai, Walter Talbott, Nitish Srivastava, Chen Huang, Hanlin Goh, Ruixiang Zhang, and Josh Susskind. An attention free transformer, 2021.

[89] Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E. Coupland, and Yalin Zheng. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. *ArXiv*, abs/2203.12081, 2022.

[90] Jingwei Zhang, Saarthak Kapse, Ke Ma, Prateek Prasanna, Joel Saltz, Maria Vakalopoulou, and Dimitris Samaras. Prompt-mil: Boosting multi-instance learning schemes via task-specific prompt tuning, 2023.

[91] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

[92] Yunlong Zhang, Honglin Li, Yuxuan Sun, Sunyi Zheng, Chenglu Zhu, and Lin Yang. Attention-challenging multiple instance learning for whole slide image classification. *arXiv preprint arXiv:2311.07125*, 2023.

[93] Yunlong Zhang, Zhongyi Shui, Yunxuan Sun, Honglin Li, Jingxiong Li, Chenglu Zhu, Sunyi Zheng, and Lin Yang. Adr: Attention diversification regularization for mitigating overfitting in multiple instance learning based whole slide image classification. *arXiv preprint arXiv:2406.15303*, 2024.

[94] Yunlong Zhang, Yuxuan Sun, Honglin Li, Sunyi Zheng, Chenglu Zhu, and Lin Yang. Benchmarking the robustness of deep neural networks to common corruptions in digital pathology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 242–252. Springer, 2022.

A Appendix

A.1 Why linear complexity is important

We find that FlashAttention [20] using memory-efficient trucking and hardware IO-aware operations is good enough in both memory and speed to cope with 20x magnification (about twice slower) WSI model (as a result we use Full Attention as the last global attention layer of our hybrid LongMIL model for 20x magnification, we also provide a replacement version using linear attention as last layer as shown in Table 9 signed as LongMIL+V-Mamba). However it is unacceptable in dealing with 40x magnification (about 30-times slower in BRACS), which takes us 2 days or more to train 5-fold runs models on BRACS, and it takes longer if stacking more layers and on larger WSI with more rounds (e.g.average over 50000 instances of TCGA-BRCA dataset with double WSI num and traditionally using 10 fold-cross validation). This hinders the improvement of 40x magnification (which includes more useful details) for both development and deployment. Thus we use LongMIL+V-Mamba (hybrid transformer as local+local+linear global attention) in 40x and also get a strong performance as shown in Table 7 left-column, which is faster in speed and comparable in performance compared to FlashAttention.

A.2 Extrapolation: Train small but test large

We first split BRACS dataset into training (small images) and validation+testing (large image) by sorting them via instance number and then use train-val-test ratio as 6:2:2. The experimental results are plot in lower-right area of Fig. 4.

A.3 HIPT Region Slicing, Local-mask Matrix and 2-d ALiBi

In Fig. 5, we show the difference and similarity between HIPT region slicing, local-mask matrix and 2-d ALiBi, where our local mask can be seen as a generalization of HIPT and 2-d ALiBi.

A.4 Details of Datasets

For the slide-level **tumor subtyping** performance, our method is evaluated on two datasets:

BReAst Carcinoma Subtyping (BRACS) [4] collect H&E stained Histology Images, containing 547 WSIs for three lesion types, i.e., benign, malignant and atypical, which are further subtyped into seven categories. Here, since the WSIs number is limited, we only perform three class subtyping. The WSIs are segmented in $20\times$ magnification and non-overlapping patching with 224×224 size. The Cancer Genome Atlas Breast Cancer (TCGA-BRCA) [68, 55] is a public dataset for breast invasive carcinoma cohort for Invasive Ductal Carcinoma versus Invasive Lobular Carcinoma subtyping. The WSIs are segmented into non-overlapping tissue-containing patches with a size of 256×256 (keep consistency to previous work [9]) at $20\times$ magnification patches were curated from 1038 WSIs. For the slide-level **survival prediction**, despite TCGA-BRCA, we further includes 2 TCGA histology datasets: 1) A combination dataset of the Colon adenocarcinoma and Rectum adenocarcinoma Esophageal carcinoma (TCGA-COADREAD), which includes 316 WSIs as used in HIPT [9]. 2) Stomach adenocarcinoma (TCGA-STAD) dataset including 321 WSIs. For pre-processing, we using the implementation of CLAM [50] which mainly includes HSV, Blur, Thresholding, and Contours methods to localize the tissue regions in each WSI.

A.5 Further Experiments and Ablations

A.5.1 TCGA-BRCA 2-categories tumor subtyping

The results are reported in Table 5 and right column of Table 6. We could observe a significant improvement in our method when using HIPT pre-trained patch embeddings, but only a slight improvement with Lunit. This could be because this task reaches an upper bound with the high-quality Lunit embeddings. Given that it only predicts binary categories, even simple max-pooling can outperform almost all previous MIL methods.

 https://doi.org/10.52202/079017-3219

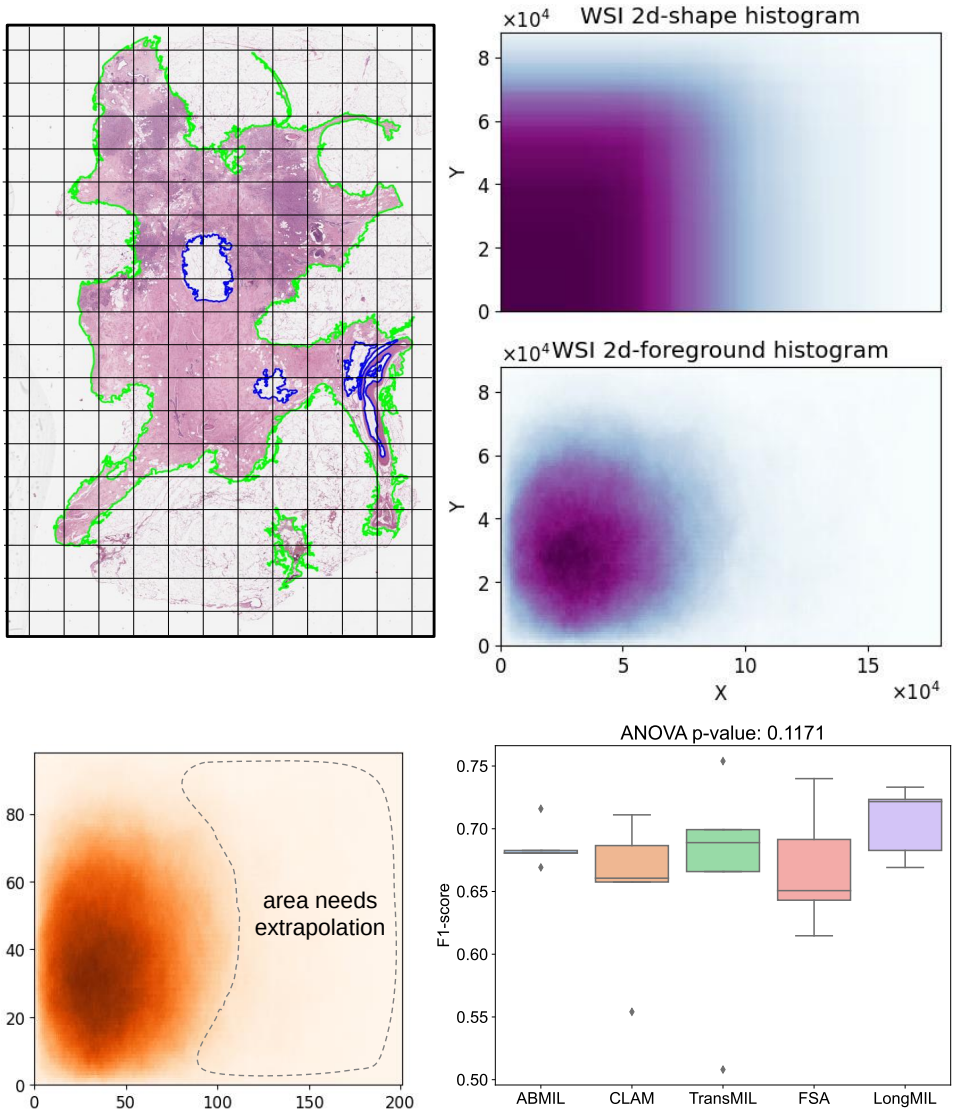

Figure 4: **upper left:** The WSI fore-ground shows irregularity (inner the green line). **upper right and lower left:** The 2-d position index of WSI foreground patches mainly scattered within index$<$100, thus area enclosed by the dashedline suffers under-fitting with previous method. **lower right:** TransMIL and full self-attention (FSA) get a relatively low performance during testing on unseen larger WSI. Assisted by our method, this case show significant performance improvement (p-value near 0.1).

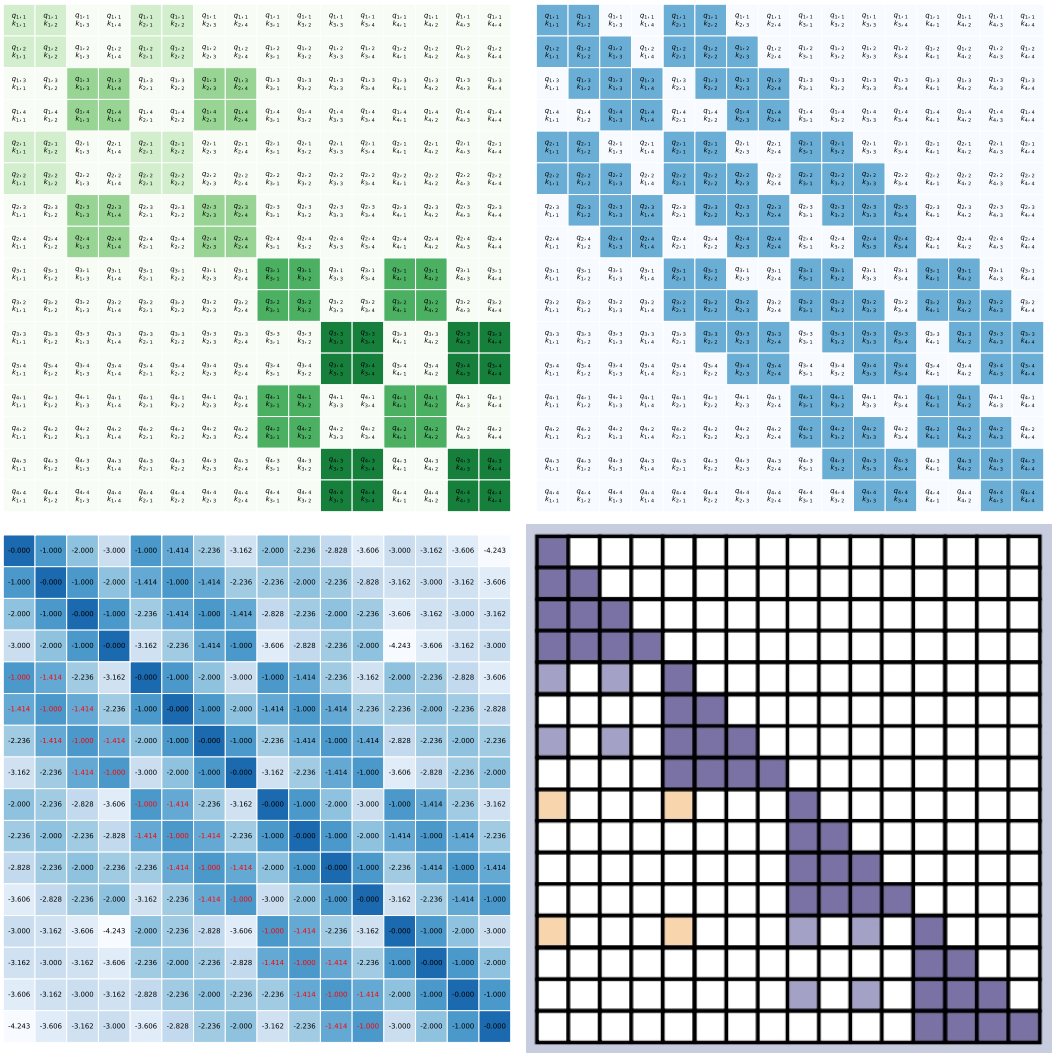

Figure 5: Difference and similarity between various methods. **upper left:** HIPT slicing with extremely hard pattern, **upper right:** our proposed local mask, **lower left:** 2-d ALiBi, or 2-d Euclid distance, **lower right:** attention mask of Prov-GigaPath from their paper (their causal attention, only focus on lower triangular matrix, may be a drawing problem). Apparently the local mask of Prov-GigaPath mainly focus on 1-d interactions (weigh x-axis of WSI more than y-axis), e.g. the interactions when distance less than 2.0 are almost missed, as depicted in the red text areas of the lower-left 2-d Euclid distance subfigure. We have checked their code implementation, which directly apply 1-d LongNet to the serialized (via z-scan) patch sequence.

https://doi.org/10.52202/079017-3219

Method	TCGA-BRCA tumor subtyping			
	ViT-S Lunit [36]		ViT-S HIPT [9]	
	F1	AUC	F1	AUC
KNN (Mean)	0.669±0.088	0.821±0.038	0.585±0.048	0.742±0.016
KNN (Max)	0.657±0.069	0.799±0.036	0.516±0.033	0.691±0.016
Mean-pooling	0.841±0.050	0.934±0.024	0.731±0.049	0.867±0.037
Max-pooling	0.849±0.051	0.949±0.022	0.688±0.074	0.826±0.058
AB-MIL [32]	0.820±0.037	0.928±0.023	0.757±0.069	0.873±0.036
DS-MIL [40]	0.841±0.047	0.925±0.024	0.723±0.068	0.854±0.036
CLAM-SB [50]	**0.850±0.039**	0.942±0.020	0.733±0.057	0.861±0.041
DTFD-MIL MaxS [89]	0.812±0.044	0.911±0.031	0.678±0.082	0.781±0.067
DTFD-MIL AFS [89]	0.843±0.035	0.931±0.015	0.704±0.075	0.851±0.056
TransMIL [65]	0.824±0.026	0.933±0.019	0.715±0.061	0.840±0.053
HIPT [9]	-	-	0.752±0.042	0.874±0.060
Full Attention	0.843±0.060	0.944±0.024	0.758±0.046	0.852±0.046
LongMIL (ours)	0.845±0.046	**0.950±0.023**	**0.762±0.064**	**0.880±0.045**

Table 5: Slide-Level Tumor Subtyping on TCGA-BRCA.

A.5.2 ResNet-50 ImageNet pre-trained embedding results of tumor subtyping

The results experimented in Table 6.

Method	ResNet-50 ImageNet pre-trained embedding			
	BRACS		TCGA-BRCA	
	F1	AUC	F1	AUC
Mean-pooling	0.483±0.018	0.710±0.004	0.751±0.049	0.861±0.026
Max-pooling	0.495±0.018	0.763±0.005	0.780±0.027	0.886±0.301
AB-MIL [32]	0.553±0.033	0.752±0.005	0.760±0.046	0.851±0.057
DS-MIL [40]	0.564±0.037	0.779±0.032	0.797±0.036	0.894±0.029
CLAM-SB [50]	0.548±0.026	0.769±0.007	0.779±0.035	0.878±0.027
TransMIL [65]	0.500±0.054	0.734±0.019	0.741±0.126	0.854±0.051
Full Attention	0.544±0.037	0.775±0.018	**0.800±0.014**	0.901±0.014
LongMIL (ours)	**0.591±0.084**	**0.810±0.038**	0.781±0.047	**0.919±0.008**

Table 6: Slide-Level Tumor Subtyping on BRACS and TCGA-BRCA based on ResNet-50 embedding pre-trained via ImageNet supervised learning.

A.5.3 Hyper-parameters of Transformer training

Number of Transformer blocks and multi-head, bias slope coefficient and local window size, weight decay and dropout ratio: Here we include following hyper-parameters for our results on BRACS with ViT-S patch embedding pre-trained by [36]: Transformer blocks and multi-head number, dropout ratio, weight decay, and learning rate, finally the local window size. Due to time-consumption, we fixed other hyper-parameters when ablation on selected variant (The default setting is Transformer local blocks number = 2, multi-head number = 1, dropout ratio = 0.0, weight decay = 1e-2, and learning rate = 1e-4, local-window size = 10 (radius)), the details can be found in Fig. 6.

A.5.4 Multi-scale and magnification

There are large differences in speed and performance for 20x and 40x magnification, since FlashAttention [20] will be quite slow if given over 20k instances compared to linear attention our local attention. For performance and speed please check Table 7 and Fig. 7b, respectively.

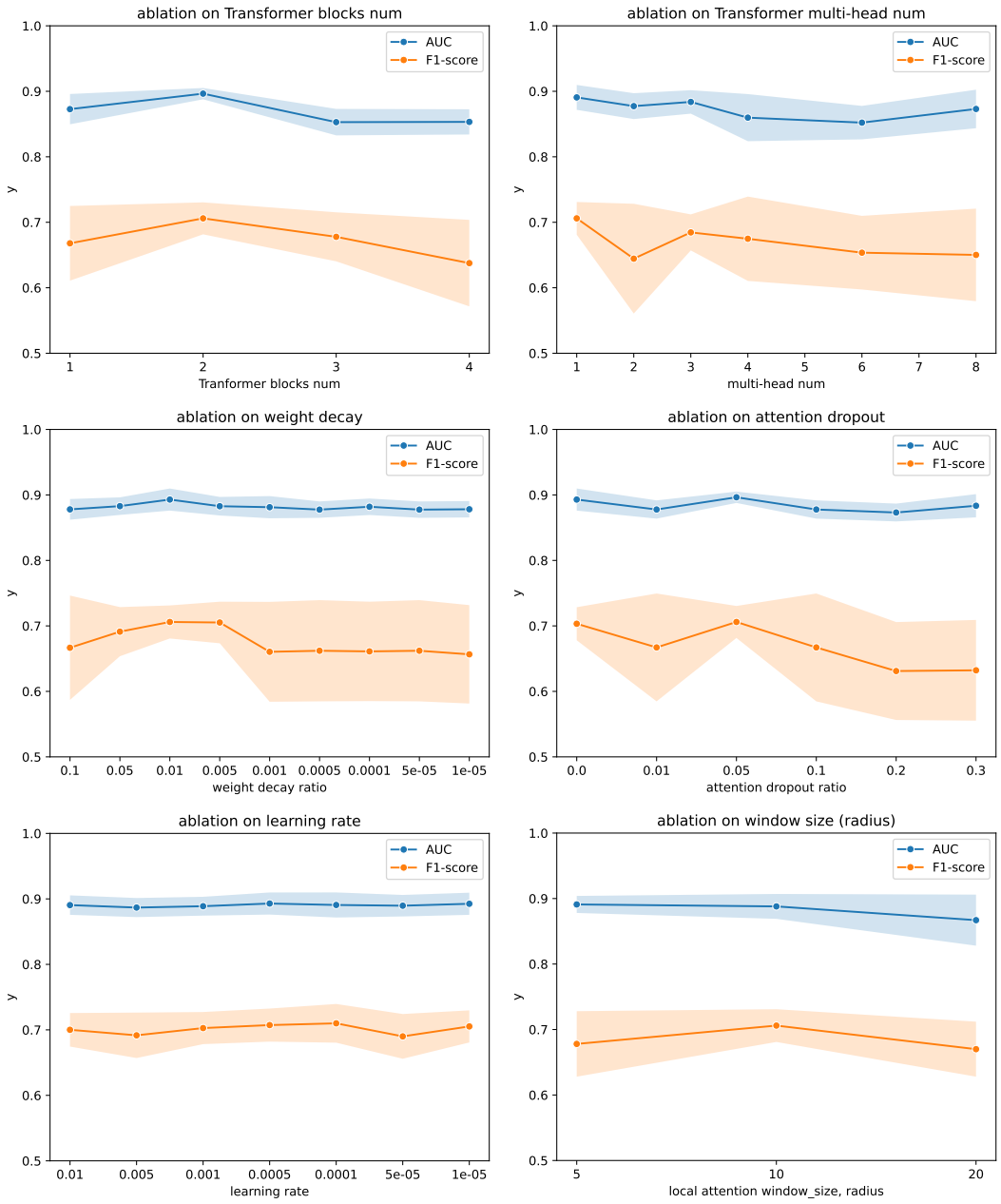

Figure 6: Ablations results on BRACS with ViT-S Lunit [36] patch embedding.

Method	BRACS tumor subtyping			
	40x		20x	
	F1	AUC	F1	AUC
AB-MIL [32]	0.610±0.034	0.811±0.013	0.668±0.032	0.866±0.016
TransMIL [65]	0.576±0.059	0.777±0.019	0.648±0.054	0.835±0.031
Full Attention	0.618±0.042	0.831±0.014	0.689±0.036	0.870±0.010
LongMIL (full global)	**0.624±0.060**	**0.842±0.022**	**0.706±0.025**	**0.888±0.019**
LongMIL (linear global)	0.622±0.055	0.835±0.026	0.693±0.024	0.870±0.016

Table 7: **Ablations on magnification** (40x and 20x) in BRACS tumor subtyping.

https://doi.org/10.52202/079017-3219

We also experiment on larger patch size (from 224 to 448, Table 8) to decrease the overall token number, but find that our method still shows stronger performance.

1. Simple attentions (ABMIL, CLAM without pair wise interactions) gain improvement, and we speculate that the larger image-size can modelling the local context better.

2. DTFD try to split the whole bag into 3 sub-bags, but smaller bag size may result in larger label noise of sub-bags which may answer its performance drop.

3. The gap between LongMIL and TransMIL decreases given closer n and d. Full attention and LongMIL show small drops, since less interactions can be modelled with less patches.

4. LongMIL still out-performs full attention. We speculate that local attention also works better when dealing with the shape-varying WSI even with less n.

	BRACS tumor subtyping			
	224		448	
Method	F1	AUC	F1	AUC
AB-MIL [32]	0.692±0.03	0.875±0.02	0.695±0.01	0.875±0.01
CLAM-SB [50]	0.640±0.06	0.844±0.03	0.654±0.03	0.851±0.02
DTFD-MIL [89]	0.655±0.03	0.878±0.02	0.625±0.03	0.839±0.01
TransMIL [65]	0.592±0.04	0.859±0.02	0.646±0.07	0.855±0.02
Full Attention	0.715±0.04	0.884±0.02	0.700±0.04	0.874±0.02
LongMIL	**0.728±0.05**	**0.887±0.01**	**0.722±0.04**	**0.883±0.01**

Table 8: **Ablations on patch size** (224 and 448) in BRACS tumor subtyping based on UNI feature.

A.5.5 Memory efficiency and speed

We show the memory efficiency and speed of various transformer structures in Fig. 7.

A.5.6 Linear Attention

We provide ablation on different linear attention e.g. RetNet, GLA [67, 85] and linear RNN structure like Mamba [26] to uncover their advantages and limitations in WSI analysis. As shown in Table 9, we first show the results of these vanilla Linear attention or RNN directly as MIL model (first row), but none of these methods can compete with Full Attention in performance. Then, we combine these modules into our LongMIL to replace its last global attention layer and we observe that this can provide us strong performance as well as linear complexity in total (2 layers of local attention + 1 layer of linear attention, better than two layers of Full Attention in both speed and performance).

A.6 Detailed comparison to Prov-GigaPath

1. The motivation /contribution: our paper not only focus on proposing an efficient self-attention mechanism for WSI, but also showing analysis on why some previous work like Roformer and TransMIL fail for WSI from the low-rank perspective, which we believe to be insightful to the digital pathology community. However, both the Prov-GigaPath [84] and LongViT [76] focus on scaling up to a large-scale of data with pre-training, which is more empirical. We believe that our analysis may also work for Prov-GigaPath and could be one potential explain on why Prov-GigaPath success and how to improve further.

2. The method details: Prov-GigaPath does not treat interactions inside x-axis and y-axis equally, though the 2-d positional embedding is applied. By putting all patches into a 1-d sequence in a 'z-scan' manner like ViT, their 1-d local attention focus more on x-axis but less on y-axis, as depicted in Fig. 5. Although this can be alleviated by their higher-level dilated attention term, the x-y inequality still exists. Whereas, our local-attention is designed for 2-d (based on 2d Euclid distance), thus treat them equally.

3. The pretrained Prov-GigaPath WSI-head seems relying heavily on their own patch-pretrained encoder, which may be a potential barrier to wide usage, e.g. there are still some cases when GigaPath

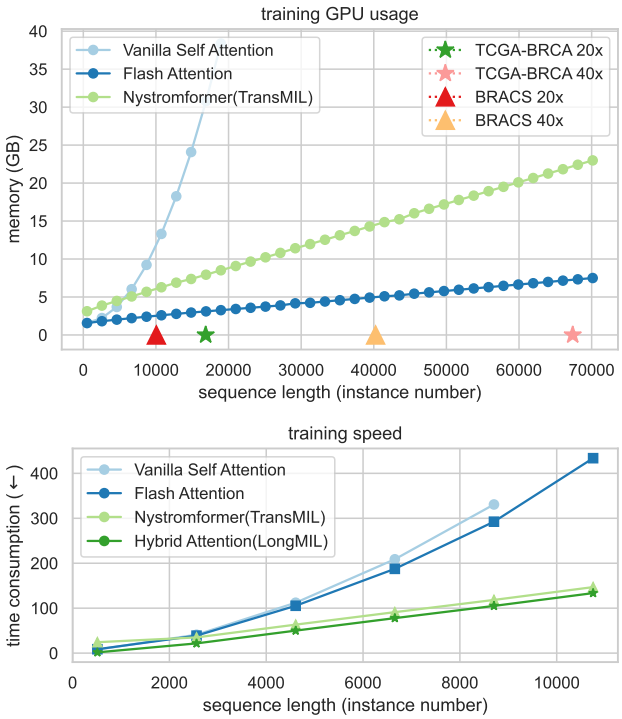

Figure 7: Training memory usage and speed using different Attentions. **Upper**: The chunk method (depicted as Flash Attention) for self-attention calculation convert memory complexity to linear, even better than Nyströmformer. **Lower**: lash Attention with chunk method still suffers quadratic complexity in speed even with hardware-aware accelerated operations. Our introduced LongMIL for WSI analysis can convert it as linearity with local-window mask. We also show markers about the max instance number of WSI used in this paper to show potentials on higher (e.g. 40x) magnification learning. We omit more instances (e.g. over 50k) speed test since it takes a long time, but based on its quadratic complexity of full self-attention, it will be about 25 to 35 times slower than linear attention.

patch features weaker than UNI or Conch, as posted in the github repo of UNI. The WSI pretraining is indeed useful as the key to their superior performance, which covers their problem of spatial inequality on x and y. When dealing with the case 'BRACS', as shown in the following table, our method (even AB-MIL) with better UNI feature can outperform their 'worse patch feature with stronger pretrained slide encoder'.

https://doi.org/10.52202/079017-3219

Method	BRACS tumor subtyping	
	F1	AUC
AB-MIL [32]	0.668±0.032	0.866±0.016
TransMIL [65]	0.648±0.054	0.835±0.031
RetNet [67]	0.628±0.034	0.805±0.009
GLA [85]	0.589±0.032	0.794±0.013
Mamba [26](random)	0.650±0.024	0.816±0.028
Mamba (single)	0.633±0.094	0.834±0.037
V-Mamba [46] (cross)	0.642±0.060	0.821±0.028
Full Attention	0.689±0.036	<u>0.870±0.010</u>
LongMIL (ours)	**0.706±0.025**	**0.888±0.019**
+ RetNet	0.690±0.051	0.848±0.013
+ GLA	0.667±0.037	0.860±0.014
+ Mamba (random)	0.678±0.044	0.856±0.030
+ Mamba (single)	0.650±0.052	0.838±0.027
+ V-Mamba	<u>0.693±0.024</u>	<u>0.870±0.016</u>

Table 9: **Experiment on Linear Attention** and combine it into our LongMIL as hybrid local-local-linear-attention Transformer model.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our main claim is clear that the low-rank with sparsity of attention matrix could be the key to improve effectiveness and efficiency for Transformer based WSI analysis.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discussed the two main limitations including application scene and current employed limited embedding size considering speed and fair comparison, in the last section of main paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

https://doi.org/10.52202/079017-3219

Answer: [Yes]

Justification: We provide full set of assumptions and complete proof for both low-rank bottleneck and high-rank of local band matrix. For the latter one, the proof by sub-matrix is clear and we also provide a full local band matrix (n = 9 in main paper) in supplemental material.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide comprehensive details of dataset selection and pre-processing, together with training details. We will release full code implementation.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All our data is public and use 'CLAM' for pre-processing. We will release full code implementation.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide model backbone selection and implementation details in core paper and further provide data split and pre-processing in supplemental material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: For each experiments we have reported the mean and std via multi-runs or cross-folds. We also show some p-value significance in train small test large experiment in supplemental material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

https://doi.org/10.52202/079017-3219

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Our memory efficiency method is run on RTX3090 GPU (24g), as detailed in core paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have read the NeurIPS Code of Ethics and understood the policies, and we believe that neither the manuscript nor the study violates any of these.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: There is no negative societal impact of our paper. Our work focus on computer aided diagnosis for potential medical use, which is discussed in conclusions of core paper.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our work mainly focus on recognition for real-world medical pathology image, without any generative problem.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All the data and tools are public accessible and well referenced.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

https://doi.org/10.52202/079017-3219

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: There is no new assets right now.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-3219