PathRWKV: Enhancing Whole Slide Image Inference with Asymmetric Recurrent Modeling

Tianyi Zhang, Sicheng Chen, Borui Kang, Dankai Liao, Qiaochu Xue, Bochong Zhang, Fei Xia, Zeyu Liu, and Yueming Jin

***Abstract*— Whole Slide Imaging (WSI) has become a gold standard in cancer diagnosis, inspecting multi-scale information from cellular to tissue levels. Processing an entire WSI directly is infeasible due to GPU memory constraints; thus, Multiple Instance Learning (MIL) has emerged as the standard solution by partitioning WSIs into tiles. While recent two-stage MIL frameworks partially achieve memory efficiency by decoupling tile-level extraction from slide-level modeling, they still face four core limitations: (1) the conflict between training throughput and inference memory efficiency, (2) the high susceptibility to overfitting on small-scale WSI datasets with sparse supervision, (3) the disruption of spatial structural integrity during sampling-based training, and (4) the inadequate modeling of multi-scale feature interactions within long sequences. We therefore introduce PathRWKV, a novel State Space Model designed for efficient and robust WSI analysis. To resolve the computational trade-off, we propose an asymmetric structure utilizing max pooling aggregation, enabling parallelized training for high throughput and recurrent inference with constant ($\mathcal{O}(1)$) memory complexity. To mitigate overfitting, we employ the random sampling strategy to enhance data diversity, with a multi-task learning module to regularize feature learning on limited data. To restore spatial context, we introduce 2D sinusoidal position encoding to perceive the relative locations of tissue tiles. To capture comprehensive representations, we integrate TimeMix and ChannelMix modules, enabling dynamic multi-scale feature modeling across both temporal and spatial dimensions. Experiments on 29,073 WSIs across 11 datasets demonstrate that PathRWKV outperforms 11 state-of-the-art methods on 10 datasets, establishing it as a superior solution for clinical-grade pathological inference.**

***Index Terms*— Whole slide image analysis, multiple instance learning, multi-task learning, state space model.**

Tianyi Zhang and Sicheng Chen contributed equally to this work. Corresponding Author: Fei Xia (e-mail: fei.xia@uci.edu), Zeyu Liu (e-mail: zeyuliu@puzzlelogic.com) and Yueming Jin (e-mail: ymjin@nus.edu.sg)

Tianyi Zhang, Borui Kang, Qiaochu Xue, Bochong Zhang, and Yueming Jin are with the Department of Electrical & Computer Engineering, National University of Singapore, Singapore 117417, Singapore (e-mails: {zhangtianyi, borui.kang, e1352520, bochong}@u.nus.edu, ymjin@nus.edu.sg).

Sicheng Chen, Dankai Liao, and Zeyu Liu are with PuzzleLogic Pte Ltd, Singapore 229594, Singapore (e-mail: {sichengchen, dankailiao, zeyuliu}@puzzlelogic.com).

Yueming Jin is also with the Department of Biomedical Engineering, National University of Singapore, Singapore 117417, Singapore (e-mail: ymjin@nus.edu.sg).

Sicheng Chen and Fei Xia are with the Department of Electrical Engineering and Computer Science, University of California Irvine, 92697, Irvine, CA, USA (e-mails:{sichenc5, fei.xia}@uci.edu)

I. Introduction

PATHOLOGY diagnosis plays an essential role in clinical practice, leveraging the analysis of pathological images [1], [2] to ensure accurate cancer diagnosis and treatment planning. The process begins with tissue biopsy/grossing by a specialist, and sample preparation workflows digitize these samples. This creates gigapixel-scale Whole Slide Images (WSIs) capturing both cell-level and tissue-level morphological details [2]. While WSIs contain rich, high-dimensional, multi-scale features, their colossal sizes make manual review labor-intensive and require highly specialized expertise, leading to inconsistent diagnoses across sites that could negatively impact the quality of healthcare [3], [4]. Deep learning–driven computational pathology techniques have emerged to ease pathologists' burden and promote high-quality diagnosis by automatically identifying critical patterns in WSIs [5]–[7]. Recent studies have further advanced this by quantifying pathologists' visual patterns to integrate expert cognitive strategies into diagnostic models, thereby aiming to minimize workload while maintaining precision [8]. Nevertheless, the complex, multi-scale nature of WSIs sets challenges for deep learning models to capture and integrate features robustly across different scales [9], [10].

Specifically, end-to-end training on raw, high-resolution WSIs remains infeasible due to GPU memory constraints and extreme dimensionality [5], [7]. Some studies downsample WSI to a thumbnail, while reduce computational complexity, this incurs significant information loss by discarding high-resolution details (e.g., cell morphology) critical for assessing disease progression [11], [12]. Consequently, most pipelines predominantly adopt Multiple Instance Learning (MIL) as a practical two-stage solution [13]. This approach breaks WSIs into smaller tiles (e.g., 224×224 pixels from an $80,000 \times 60,000$-pixel slide) and encodes them into dense feature representations using a foundation model (e.g., Prov-GigaPath [14]). Subsequently, a slide-level backbone aggregates these compressed tile features to generate slide-level predictions [6], [15], [16]. This paradigm robustly enhances performance by enabling the processing of a larger number of tiles per iteration and leveraging the robust prior knowledge embedded in the foundation model. Accordingly, these advantages contribute to the high accuracy achieved by modern MIL frameworks [8], [17]. However, critical challenges remain unresolved as detailed below:

A primary challenge in WSI analysis arises from the drastic variation in the number of tiles per slide due to diverse tissue dimensions [17]. During training, a fixed number of tiles (e.g., 2,000) is uniformly sampled from each WSI to leverage a larger batch size. This maximizes GPU parallelization efficiency, and stabilizes gradient descent, thereby enhancing overall performance [7]. Conversely, during inference, to cover all regions for diagnostic evidences, the batch size is generally set to one to process all available tiles [5]. However, given the immense disparity in WSI dimensions (e.g., from 1,000 to over 40,000 tiles in the CAMELYON16 dataset [4] at 0.5 mpp with a 224×224 patch size), processing large-scale slides can easily exceed GPU memory limits, particularly when deploying recent effective yet complex Transformers on resource-constrained edge devices [18]. This necessitates a slide-level structure that possesses parallel computing capabilities during training to handle large batch sizes, while retaining sequential processing efficiency during inference to model entire WSIs with minimal memory overhead [19].

Another critical challenge stems from the severe data inefficiency in WSI analysis, where high-capacity models struggle to generalize under the dual constraints of data scarcity and sparse supervision [20], [21]. Specifically, obtaining annotated cohorts is prohibitive, often restricting datasets to limited sizes (e.g., fewer than 3,000 slides) [1]. This scarcity is exacerbated by the weak nature of slide-level labels, which provide supervision for only a fraction of the gigapixel-resolution tissue information [5]. Consequently, despite the improved feature representations from foundation models [7], [14], the downstream aggregators remain prone to overfitting. Notably, complex structures like TransMIL [15] frequently underperform compared to simpler baselines (e.g., CLAM [6]) in such low-data regimes [20]. This necessitates strategies that maximize the utility of limited slide-level data to enhance model generalization [22], [23].

Furthermore, the disruption of spatial structural integrity during sampling-based MIL training significantly impedes the performance of permutation-variant methods. Conventional mini-batch training requires a fixed input size, and current methods primarily address this by selecting a subset of tiles to represent the whole WSI [13]. This inevitably disrupts the global spatial context. While this issue is negligible for permutation-invariant Attention-based methods (e.g., CLAM [6]), it poses a critical challenge for the current permutation-variant state-of-the-art (SOTA) paradigms, like Transformers (e.g., TransMIL [15]) and State Space Models (SSMs) (e.g., MambaMIL [24]). All of them rely on explicit modeling of spatial relationships and contextual dependencies. Consequently, the spatial information loss induced by sampling severely compromises the modeling capabilities of these structures [25]. It is imperative to develop a mechanism that can effectively recover this missing spatial context [26].

The final challenge arises as most current methods face difficulties in adequately handling multi-scale feature interactions within long sequences. Effective diagnosis relies on the complementarity between fine-grained details (e.g., cell nuclei at 40×) and coarse-grained context (e.g., tissue structure at 4×), and the inherent ambiguity of cellular structures

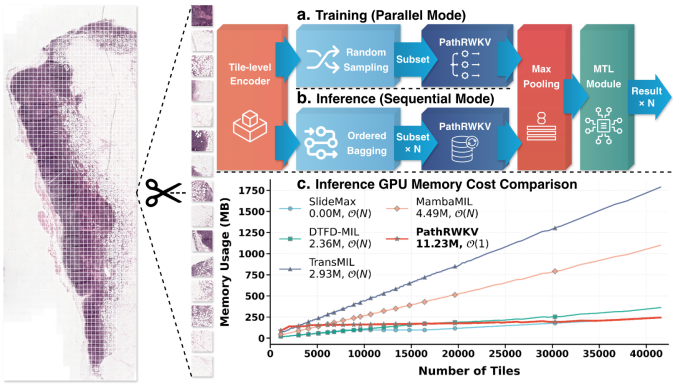

Fig. 1. The asymmetric structure of PathRWKV and its effectiveness. (a) Parallel mode: The model samples a fixed maximum number of WSI tiles for multi-sample parallel processing to maintain efficiency during training. (b) Sequential mode: The model processes all WSI tiles for precise inference. It splits tiles into equal-sized bags and processed sequentially in a single forward pass. A memorable state retains and propagates information from prior bags. (c) GPU memory usage comparison during inference. PathRWKV maintains the $\mathcal{O}(1)$ spatial complexity, showing superior memory efficiency on long-context modeling compared with previous methods.

theoretically necessitates fuzzy logic or high-order topological modeling [27]. This requires a multi-scale modeling capability to simultaneously handle the relationships between fine-grained local homogeneity and coarse-grained global heterogeneity [22]. However, existing methods demonstrate an inability to conduct effective multi-scale feature analysis for slide-level conclusions. Lightweight Attention-based methods benefit from simple, permutation-invariant structures but lack the capacity to capture intricate fine-grained relationships (e.g., inter-cell interactions). Conversely, while large Transformers excel at modeling local details, their generic structures often struggle to effectively align and fuse these heterogeneous multi-scale features into a unified slide-level representation. A promising solution is to model features from multiple perspectives, incorporating diverse indicators to construct a robust multi-scale understanding.

Due to the similar requirements of long context modeling and multi-scale understanding between natural language processing (NLP) and MIL, previous MIL approaches based on NLP structures have proven effective (e.g., TransMIL [15] from Transformer, MambaMIL [24] from Mamba). Among them, RWKV [19] stands out by uniquely combining the efficient parallelizable training of Transformers with the linear complexity inference of RNNs. This makes it exceptionally suitable for processing the massive, sequential feature representations inherent in WSI analysis. Motivated by these properties, we propose PathRWKV, a time-decayed SSM tailored for efficient and robust WSI analysis.

To resolve the asymmetric memory and efficiency constraints, we propose an asymmetric structure (Fig. 1) that integrates max pooling aggregation with linear attention. This design enables a seamless transition between Transformer-like parallelization for high-throughput training and RNN-like sequential processing for inference, achieving constant ($\mathcal{O}(1)$) memory complexity regardless of slide size. Consequently, it achieves high throughput during training while ensuring exceptional memory efficiency during inference.

To mitigate overfitting and data scarcity inherent in weak supervision, we apply the random sampling strategy with the multi-task learning (MTL) module. Random sampling acts as a dynamic data augmentation technique, counteracting the inductive bias of deterministic sampling and exposing the model to diverse subsets of tissue regions. The MTL module introduces auxiliary supervision signals to regularize feature space, preventing the model from memorizing noise in limited training samples. Together, these strategies exploit the potential of limited annotations and bolster model generalizability by capturing intrinsic inter-task dependencies.

To address the disruption of spatial structural integrity caused by random sampling, we leverage 2D sinusoidal position encoding (2D PE) to embed unique coordinate-based information into each tile feature. It is critical for the permutation-variant PathRWKV to recognize relative positions and reconstruct spatial relationships. This design effectively equips the model to preserve global spatial context, bridging the significant distributional gap between the stochastic bag-of-tiles input used during training and the ordered, sequential slide processing required for inference.

To tackle complex multi-scale feature interactions, we incorporate TimeMix and ChannelMix modules. The TimeMix module focuses on capturing long-range spatial dependencies and local homogeneity across the sequence of tiles, while the ChannelMix module focus on high-level abstract semantic patterns. By jointly modeling these dimensions, the structure ensures a robust representation that encompasses both fine-grained cellular details and coarse-grained global tissue heterogeneity across the entire slide.

This work makes the following contributions:

- We propose *PathRWKV*, a novel SSM for efficient and robust slide-level modeling in computational pathology.
- We design an asymmetric slide-level structure that combines max pooling aggregation, enabling efficient parallelized training and recurrent inference with constant ($\mathcal{O}(1)$) memory complexity. Built upon this structure, we further introduce random sampling strategy and MTL module to mitigate overfitting under weak supervision and improve data efficiency and generalization.
- We restore spatial context by incorporating 2D PE, and enhance multi-scale representation learning via TimeMix and ChannelMix modules, enabling dynamic interaction between fine-grained cellular features and coarse-grained tissue structures.
- We conduct extensive experiments on 29,073 WSIs across 11 public datasets, demonstrating SOTA performance on 10 datasets. Beyond accuracy, these results validate PathRWKV as a scalable and trustworthy framework for clinical-grade inference, establishing a new perspective where the asymmetry between training and inference serves as a powerful inductive principle for computational pathology. The code is publicly available at https://github.com/Puzzle-Logic/PathRWKV.

II. Related Works

The two-stage MIL paradigm becomes the mainstream recently, with the first stage extracts tile-level features, and the second aggregates them for slide-level predictions. Initial attempts within this paradigm employed simple aggregation strategies, such as average pooling (SlideAve) and max pooling (SlideMax) from MINNs [28], to combine tile features into a slide-level representation. While computationally efficient, these methods simply treat all tiles embeddings equally or focus exclusively on the most salient one, often failing to capture the complex, fine-grained information required for accurate diagnosis. To address this limitation, ABMIL [13] introduced a gated attention mechanism that adaptively weights tiles to enable instance-level interpretability. This simple yet effective mechanism has been widely adopted by most modern methods. Building on this, CLAM [6] imposes instance clustering constraints to encourage diverse and discriminative feature learning, thereby improving model generalization. DSMIL [29] further incorporates contrastive learning by combining instance- and bag-level supervision to better distinguish informative tiles. To address the challenge of limited data scale, DTFD-MIL [16] proposes a double-tier feature distillation framework that utilizes pseudo-bags to virtually expand the training set, enhancing robustness in small-sample scenarios. Despite their widespread adoption, these attention-based methods often struggle to fully capture complex spatial dependencies within WSIs due to their inherent permutation invariance.

Transformers have been introduced to alleviate the permutation invariance of attention-based methods and improve global context modeling. A representative method, TransMIL [15], employs a Transformer-based structure that explicitly encodes positional and structural relationships among tiles, enabling more effective global spatial reasoning. Furthermore, Prov-GigaPath [14] enhances global information flow through a dilated attention mechanism, leveraging efficient sequence modeling structures like LongNet [18] to improve scalability. Despite these advancements, Transformers still face challenges regarding overfitting and high memory consumption, especially when trained on small-scale datasets.

State Space Models (SSMs) have emerged as a compelling alternative, balancing the efficiency of conventional attention methods with the long-range modeling capabilities of Transformers. S4MIL [30] introduces the Structured State Space Sequence (S4) model to capture long-range dependencies across tiles. By imposing a structured state representation, it mitigates the overfitting risks associated with data-hungry Transformers. Building on this, MambaMIL [24] integrates the selective state space model, Mamba, into the MIL pipeline, enabling linear scaling and selective information flow across tiles. Similarly, MamMIL [31] adapts the Mamba structure to model WSIs as long sequences, effectively capturing bidirectional contextual dependencies with minimal computational overhead. These SSM-based approaches not only offer superior scalability but also enhance generalization through their inherent inductive biases, making them particularly well-suited for WSI analysis on limited datasets. However, most current SSM-based approaches retain attention-based pooling mechanisms for final aggregation. This design necessitates storing features from all tiles, reintroducing an $\mathcal{O}(N)$ memory bottleneck during inference that undermines the inherent linear efficiency of the SSM backbone.

III. METHODS

A. Overview of the MIL Pipeline

Fig. 2a illustrates the overall PathRWKV pipeline. Following existing works [32], each WSI S is first loaded at a target resolution (e.g., 0.5 microns per pixel (mpp)) and partitioned into a non-overlapping grid of tiles $\{T_{i,j}\}$ of size T_{size}. To ensure data quality, a two-stage filtering protocol is employed. First, tiles with tissue coverage below a predefined threshold (e.g., $< 50\%$ of the tile area) are discarded. Second, tiles with pixel variance falling below a quantitative cutoff (e.g., $Var(I) < 0.01$, where I denotes the normalized pixel intensity in [0,1]) are removed. This ensures that only informative tiles are retained for downstream tasks. After pre-processing and filtering, each tile is embedded into a dense semantic feature vector (Fig. 2a) using a pathological foundation model (e.g., Prov-GigaPath [14]). This embedding process enhances both the training efficiency and performance of the slide-level MIL. During training, PathRWKV processes a randomly shuffled subset of tiles to facilitate efficient multi-slide learning; conversely, during inference, it utilizes the complete sequential tile sequence from the WSI. Finally, the slide-level output features from PathRWKV are passed to the MTL module to generate predictions for each task (e.g., cancer subtyping, tumor grading, overall survival).

B. The PathRWKV Slide-level Backbone

PathRWKV serves as the backbone for slide-level feature modeling, consisting of 2 blocks with a hidden dimension of 768, and 12 heads. The input tile features are first combined with 2D sinusoidal Position Encoding (2D PE) to restore spatial relationships disrupted during the sampling process. Subsequently, these features are processed by the PathRWKV blocks (Fig. 2b). Each block comprises a TimeMix module and a ChannelMix module, integrated with layer normalization and residual connections. The TimeMix module dynamically captures multi-scale inter-tile dependencies via the temporal dimension, while the ChannelMix module focuses on intra-tile feature interactions across the channel dimension. Output features from the final PathRWKV block are aggregated via max pooling to produce the final slide-level representation.

Fig. 2c illustrates the key mathematical operations within the TimeMix and ChannelMix modules. The TimeMix module is specifically designed to capture multi-scale temporal dependencies among tiles. It effectively integrates fine-grained local interactions via token shifting and interpolation with coarse-grained global context via time-decayed linear attention, combining the strengths of Transformers [33] and RNNs [34].

To capture short-range, fine-grained dependencies between adjacent tiles, the module first employs a token-shift mechanism, TimeShift. A shifted version of the input x, denoted as x_{last}, is generated using zero-padding:

$$x_{last,t} = x_{t-1}, \quad x_{last,0} = \mathbf{0} \tag{1}$$

The current input x is then mixed with x_{last} via data-dependent linear interpolation (ddlerp). Unlike static interpolation, ddlerp dynamically computes the mixing coefficient μ using a Low-Rank Adaptation (LoRA) mechanism:

$$\begin{aligned} \delta x_t &= x_{last,t} - x_t \\ x'_t &= x_t + \delta x_t \odot \mu_x \\ \mu_t &= \lambda + \tanh(x'_t W_A) W_B \\ x_{ddlerp,t} &= x_t + \delta x_t \odot \mu_t \end{aligned} \tag{2}$$

This mechanism allows the model to adaptively aggregate local information from the immediate predecessor based on the current context, ensuring that high-frequency local variations are preserved before global processing.

Following local aggregation, the interpolated features are projected into five vectors: receptance r, key k, value v, time-decay w, and gate g. To capture long-range, coarse-grained dependencies across the entire slide sequence, we employ a time-decayed linear attention mechanism. Notably, the decay rate w is modulated by a LoRA projection, allowing for data-dependent decay speeds that can adaptively focus on relevant historical context. The global linear attention is computed efficiently as:

$$y_t = r_t \odot \left(\sum_{i=1}^{t-1} \left(\prod_{j=i+1}^{t} w_j \right) \odot k_i v_i^\top + u \odot k_t v_t^\top \right) \tag{3}$$

Crucially, this formulation can be switched to a recurrent structure, which underpins our asymmetric design. By maintaining a recurrent state S_t, the model propagates global context sequentially:

$$\begin{aligned} S_t &= w_t \odot S_{t-1} + k_t v_t^\top \\ y_t &= r_t \odot (S_{t-1} + u \odot k_t v_t^\top) \end{aligned} \tag{4}$$

Finally, the output is gated by g and projected by W_o:

$$y = W_o(\text{GroupNorm}(\mathbf{y}) \odot g) \tag{5}$$

The ChannelMix module focuses on intra-tile feature interactions. Specifically, it employs learnable linear projections (r, k, v) to blend information across the channel dimension D for each tile independently. This is coupled with a Squared ReLU activation:

$$\sigma(x) = \max(0, x)^2 \tag{6}$$

which induce robust non-linear transformations, enabling the extraction of complex morphological features within each tile.

C. Asymmetric Structure and Max Pooling Aggregation

As illustrated in Fig. 1, distinct from the standard token-by-token processing in original RWKV, we propose a novel hybrid set-by-set recurrent architecture tailored for high-resolution WSI analysis. While leveraging the mathematical efficiency of linear attention, our core innovation lies in the asymmetric formulation of slide-level modeling to resolve the memory bottlenecks inherent in existing SSMs.

During the training phase, we adopt a set-based parallel strategy to maximize throughput. Instead of processing tiles sequentially, the model ingests a fixed number of sampled tiles as a dense batch. By utilizing a parallel CUDA kernel, we compute cumulative states and gradients simultaneously.

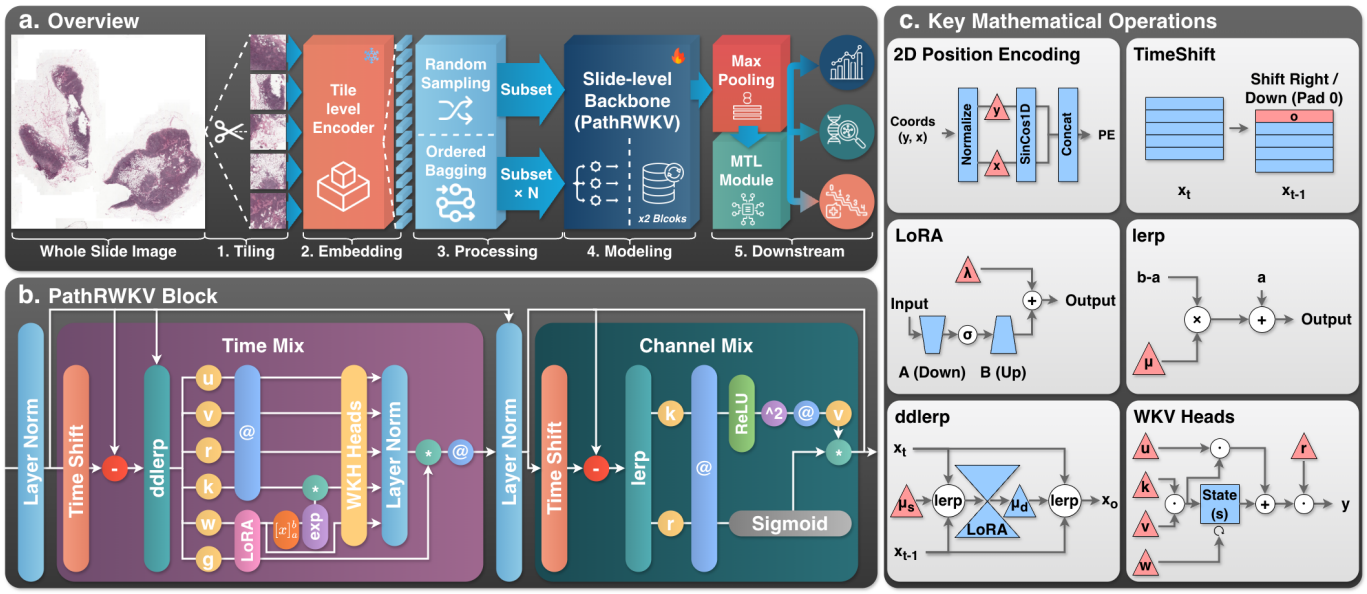

Fig. 2. Overview of PathRWKV. a) The pipeline begins with WSI tiling and tile-level feature embedding via Prov-GigaPath, followed by the slide-level backbone via PathRWKV, which enables multi-task learning for different downstream tasks. b) The PathRWKV block consists of the TimeMix module, which integrates tile features with previous states from multi-scale using linear attention, and the ChannelMix module, which blends features via spatial aspect. c) Details of the specific mechanisms employed, including 2D Position Encoding, TimeShift, LoRA, lerp, ddlerp, and WKV Heads.

This design fully exploits the massive parallelism of modern GPUs, facilitating rapid backpropagation and stable convergence compared to pure recurrent training.

During the inference phase, we introduce a streaming recurrent mechanism to achieve constant $\mathcal{O}(1)$ spatial complexity. A critical limitation in previous SSM-MIL methods (e.g., S4MIL, MambaMIL) is their reliance on Attention-based aggregation, which necessitates storing all tile features ($\mathcal{O}(N)$) for the final calculation, causing memory overflows on gigapixel slides. To overcome this, our architecture decomposes the WSI into sequential chunks. A latent state S (of size HeadSize$\times$HeadSize) acts as a memory carrier, propagating context across chunks via a recurrent kernel.

Furthermore, a critical bottleneck remains in the aggregation mechanism of existing frameworks, including recent SSMs. While SSM backbones theoretically allow constant spatial complexity ($\mathcal{O}(1)$) during inference, they typically adapt the Gated Attention mechanism from ABMIL [13] for slide-level aggregation. This approach necessitates computing a global Softmax normalization term across all tiles:

$$\alpha_k = \frac{\exp(w^\top h_k)}{\sum_{i=1}^{N} \exp(w^\top h_i)} \tag{7}$$

Consequently, the feature vectors of all N tiles must be retained in GPU memory to calculate the denominator until the entire slide is processed, forcing the inference memory complexity to scale linearly with the slide size ($\mathcal{O}(N)$). This disrupts the memory efficiency gained by the SSM backbone.

To achieve a fully memory-efficient pipeline, we integrate the backbone's sequential processing with a streaming aggregation strategy. We decompose the WSI $\mathcal{X}$ into bags $\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_M$ aligned with the inference chunks, and identify the *feature-wise max operation* as the optimal choice due to its recursive update property. Given the tile encoder ϕ_θ, we define the local summary z_i and the combination rule as:

$$z_i = g(\mathcal{B}_i) := \max_{x \in \mathcal{B}_i} \phi_\theta(x), \quad \mathrm{Comb}(a, b) = \max(a, b) \tag{8}$$

We then update the slide-level representation sequentially:

$$h_i = \mathrm{Comb}(h_{i-1}, z_i), \quad h_0 = \varnothing \tag{9}$$

where h_i represents the summary of the first i bags. Unlike attention mechanisms, this design decouples memory usage from sequence length, achieving true $\mathcal{O}(1)$ space complexity during inference. This enables PathRWKV to process arbitrarily large slides on edge devices while preserving consistency with the ideal global computation. Comprehensive theoretical analysis, including proofs for the unbiased nature of the gradients and memory complexity comparisons, are provided in the supplementary materials.

D. Random Sampling and 2D Position Encoding

Given the high dimensionality of WSIs and the sparse supervision signal, models are prone to overfitting. To mitigate this, we employ the random sampling strategy during training. Instead of processing the entire slide or a fixed region, we randomly sample a subset of tiles from the WSI in each iteration. This approach acts as a strong data augmentation technique, preventing the model from memorizing specific tile sequences and enhancing its generalization capabilities.

However, a critical side effect of random sampling is the disruption of the intrinsic 2D spatial structure and the anatomical adjacency of the tissue microenvironment. This is the main advantage of recent permutation-variant methods (e.g., MmabaMIL [24]) compared to early permutation-invariant methods (e.g., ABMIL [13]). To compensate for this loss of spatial context and enable PathRWKV to model geometry-aware dependencies, we introduce a 2D PE. Formally, let $\mathbf{z}_i \in \mathbb{R}^D$ denote the feature embedding of the i-th tile, and (x_i, y_i) represent its normalized spatial coordinates. We

employ sinusoidal functions to map these coordinates into a continuous embedding space:

$$\begin{aligned} \mathrm{PE}(p, 2k) &= \sin\left(p/\Omega^{4k/D}\right) \\ \mathrm{PE}(p, 2k+1) &= \cos\left(p/\Omega^{4k/D}\right) \end{aligned} \tag{10}$$

where Ω is a scaling factor and $p \in \{x_i, y_i\}$. The final spatial embedding $\mathbf{P}_i$ is constructed by concatenating the encodings of horizontal and vertical coordinates and injected into the tile features via addition: $\hat{\mathbf{z}}_i = \mathbf{z}_i + \mathbf{P}_i$. This ensures that geometric relationships are restored regardless of the sampling order.

E. Multi-task Learning

Finally, to maximize the utility of limited annotated data and enhance training efficiency, we incorporate a multi-task learning (MTL) module. It comprises multiple prediction heads, allowing the model to learn from diverse clinical objectives simultaneously. By leveraging task-wise correlations, the model extracts more discriminative features, further reducing the risk of overfitting on the feature distribution of a single task. We employ Cross-Entropy, Cox proportional hazards, and L1 losses for classification, survival analysis, and regression tasks, respectively. The total loss is aggregated as $\mathcal{L}_{total} = \sum_{i=1}^{T} \mathcal{L}_i$, where T represents the number of task heads. Crucially, gradients are computed only for tasks with available labels, enabling flexible training on partially annotated datasets.

IV. Experiment

A. Datasets and Downstream Tasks

We evaluated PathRWKV on 11 datasets across 9 downstream tasks covering diverse diagnostic scenarios, as shown in Fig. 3, to demonstrate its performance and generalizability. The PANDA [35] dataset with the ISUP Grade task assesses prostate cancer aggressiveness; CAMELYON16 [4] with the Breast Metastasis task classifies lymph nodes as normal or tumorous; IMP-CRS-2024 [36] with the CRC-Tumor task identifies tumor tissues in colorectal images. The TCGA [37] datasets cover multiple cancer types and tasks: TCGA-BRCA with the IHC-HER2 task predicts HER2 receptor status from H&E-stained slides; TCGA-GBM with the Histological Diagnosis task classifies glioblastoma subtypes based on morphology; TCGA-LGG with the Tumor Stage task predicts the WHO grade of lower-grade gliomas; TCGA-CESC with the Lymphovascular Involvement task detects the presence of lymphovascular invasion; TCGA-ESCA with the Cancer Grade task assesses the histological differentiation grade of esophageal carcinoma; and TCGA-LIHC, TCGA-BLCA, and TCGA-LUNG, the combination of TCGA-LUAD and TCGA-LUSC, with the Overall Survival task predict patient survival time in months from liver, bladder, and lung tissue morphology, respectively.

B. Implementation Details

We use the UnPuzzle framework [32] for preprocessing, where each WSI is tessellated into 224×224 patches at

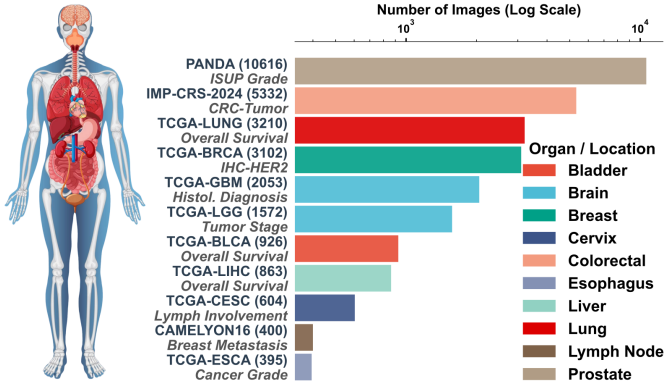

Fig. 3. Summary of implemented datasets and clinical tasks. The horizontal bars display the number of WSIs for each cohort on a logarithmic scale, alongside their corresponding prediction targets. This collection covers a wide spectrum of sample sizes and clinical objectives, ensuring a comprehensive evaluation of model generalizability.

0.5 mpp and embedded using the tile-level encoder of Prov-GigaPath [14]. All models are initialized from scratch and trained for 100 epochs using the AdamW optimizer and a cosine decay scheduler with a final learning rate factor of 0.1. We conduct a grid search over ten learning rates (1×10^{-6} to 1×10^{-3}) and employ early stopping with a patience of 10 epochs based on validation loss. During training, we use a batch size of 4 and randomly sample a maximum of 2,000 tiles per WSI. For evaluation, we select the checkpoint with the lowest validation loss, use a batch size of 1, and process all tiles per WSI. We report the average of the top-3 results for each metric and calculate P-values using Welch's t-test [38] to compare each method against PathRWKV. All experiments were conducted on 4 NVIDIA RTX4090 GPUs using Python 3.12.12, PyTorch 2.9.1, and CUDA 12.8.

C. Comparison with SOTA Methods

To demonstrate the effectiveness of PathRWKV for slide-level WSI modeling, we compared it against 11 state-of-the-art (SOTA) methods, including SlideAve and SlideMax from MINNs [28], ABMIL [13], CLAM [6], DSMIL [29], DTFD-MIL [16], TransMIL [15], Prov-GigaPath [14], S4MIL [30], MambaMIL [24], and MamMIL [31].

As presented in Tab. I, PathRWKV demonstrates superior performance and robust generalizability, achieving SOTA results on 10 out of 11 datasets across 9 distinct downstream tasks. Specifically, on standard classification benchmarks such as CAMELYON16 and IMP-CRS-2024, most deep learning-based methods achieve high metrics ($> 90\%$ Accuracy, AUC, and F1), with the exception of simple pooling strategies, SlideAve and SlideMax, that lack the representational capacity for comprehensive slide-level modeling. Among high-performing models, PathRWKV consistently secures the highest scores across all three metrics.

In contrast, the TCGA datasets present a significantly more challenging scenario, where the average performance of most methods drops to approximately 70% due to the intrinsic complexity and heterogeneity of the samples. Despite these challenges, PathRWKV establishes its efficacy on the majority of TCGA datasets, validating the capability of the proposed TimeMix and ChannelMix modules to capture complex pathological dependencies. However, we observe a performance gap

TABLE I

THE PERFORMANCE COMPARISON WITH SOTA METHODS ON ELEVEN DOWNSTREAM DATASETS.

Dataset	Metric	SlideAve [28]		SlideMax [28]		ABMIL [13]		CLAM [6]		DSMIL [29]		DTFD-MIL [16]		TransMIL [15]		GigaPath [14]		S4MIL [30]		MamMIL [31]		MambaMIL [24]		PathRWKV
		Value	P-Value	Value	P-Value	Value	P-Value	Value	P-Value	Value	P-Value	Value	P-Value	Value	P-Value	Value	P-Value	Value	P-Value	Value	P-Value	Value	P-Value	Value
PANDA	Acc.[%]	63.92	< 0.001	62.74	< 0.001	76.06	0.008	75.97	0.029	76.39	0.042	75.53	0.092	75.79	0.014	73.58	0.002	75.93	0.004	75.70	0.006	75.99	0.009	**76.45**
	AUC[%]	89.53	< 0.001	88.54	< 0.001	94.71	0.005	94.74	0.018	94.20	0.012	93.93	0.035	93.87	0.044	93.67	0.032	94.07	0.014	94.81	0.040	94.30	0.006	**94.89**
	F1[%]	60.76	< 0.001	55.94	< 0.001	70.30	0.007	69.90	0.023	70.12	0.022	70.03	0.017	69.02	0.009	68.06	0.006	70.70	0.003	69.94	0.042	70.56	0.008	**70.81**
CAMELYON16	Acc.[%]	68.99	< 0.001	72.87	< 0.001	**98.45**	0.018	**98.45**	0.010	92.25	0.049	**98.45**	0.012	**98.45**	0.025	**98.45**	0.020	**98.45**	0.028	97.67	0.037	97.67	0.014	**98.45**
	AUC[%]	55.19	< 0.001	74.44	< 0.001	97.90	0.026	98.77	0.015	96.51	0.004	98.65	0.035	98.80	0.025	97.93	0.004	98.92	0.027	98.67	0.015	99.03	0.008	**99.11**
	F1[%]	62.70	< 0.001	70.29	0.004	**98.34**	0.018	**98.34**	0.014	91.47	0.008	**98.34**	0.010	**98.34**	0.022	**98.34**	0.020	**98.34**	0.023	97.52	0.037	97.50	0.039	**98.34**
IMP-CRS-2024	Acc.[%]	92.33	< 0.001	91.67	0.003	94.67	0.022	94.22	0.004	94.44	0.035	94.33	0.014	94.44	0.022	94.56	0.018	**94.78**	0.020	94.22	0.007	94.33	0.013	**94.78**
	AUC[%]	98.59	< 0.001	98.56	0.001	99.36	0.003	99.42	0.011	99.40	0.006	99.41	0.008	99.44	0.005	99.43	0.006	99.44	0.028	99.41	0.009	99.43	0.019	**99.45**
	F1[%]	92.67	< 0.001	92.08	0.006	94.90	0.021	94.40	0.004	94.69	0.027	94.67	0.011	94.54	0.004	94.83	0.012	94.85	0.019	94.41	0.015	94.52	0.017	**95.09**
TCGA-BRCA	Acc.[%]	59.20	0.015	55.19	0.015	59.30	0.008	59.39	0.005	59.39	0.009	59.30	0.007	59.84	0.006	59.30	0.007	59.39	0.002	59.30	0.008	59.30	0.011	**60.11**
	AUC[%]	61.14	0.009	56.97	0.022	63.42	0.008	63.10	0.024	64.32	0.034	64.18	0.037	61.47	0.009	64.12	0.026	63.71	0.026	59.84	0.023	59.67	0.004	**64.35**
	F1[%]	25.82	0.018	26.09	0.020	27.54	0.004	30.00	0.019	30.04	0.040	28.64	0.018	26.57	0.021	29.41	0.024	29.59	0.004	28.73	0.014	25.88	0.018	**30.40**
TCGA-GBM	Acc.[%]	98.60	0.034	97.90	0.004	98.60	0.025	98.60	0.022	99.30	0.007	98.60	0.025	97.90	0.004	99.30	0.028	99.30	0.027	98.60	0.022	99.30	0.003	**100.00**
	AUC[%]	66.27	0.007	65.71	0.016	**66.67**	0.023	**66.67**	0.010	66.51	0.035	66.19	0.016	65.87	0.016	66.59	0.004	**66.67**	0.023	**66.67**	0.014	**66.67**	0.037	**66.67**
	F1[%]	74.65	0.023	49.47	0.034	74.65	0.008	74.65	0.028	74.65	0.002	74.65	0.015	49.47	0.034	89.82	0.003	89.82	0.019	74.65	0.025	89.82	0.026	**100.00**
TCGA-LGG	Acc.[%]	68.81	0.004	68.81	0.003	68.14	0.006	61.07	0.028	68.47	0.019	68.15	0.006	68.44	0.007	68.46	0.001	66.44	0.024	68.50	0.008	68.14	0.037	**69.13**
	AUC[%]	71.80	0.004	71.67	0.009	71.89	0.017	67.59	0.010	71.95	0.009	71.79	0.008	67.88	0.004	71.68	0.019	72.10	0.024	71.68	0.034	71.89	0.028	**72.61**
	F1[%]	68.54	0.002	68.00	0.007	68.23	0.001	60.28	0.027	68.46	0.013	68.38	0.014	66.44	0.030	62.35	0.056	63.89	0.023	68.18	0.017	66.69	0.003	**68.96**
TCGA-CESC	Acc.[%]	62.96	0.024	62.96	0.007	59.26	0.005	66.67	0.023	**70.37**	0.009	59.26	0.020	66.67	0.034	67.78	0.003	55.56	0.001	66.67	0.001	59.26	0.013	**70.37**
	AUC[%]	68.13	0.005	56.59	0.002	62.09	0.004	65.93	0.008	70.88	0.029	68.13	0.007	70.33	0.022	71.43	0.012	56.04	< 0.001	67.03	0.017	58.24	< 0.001	**75.27**
	F1[%]	62.50	0.019	62.91	0.024	59.03	0.003	66.67	0.031	69.32	0.009	58.35	0.017	66.48	0.021	67.75	0.001	55.49	0.002	65.92	0.037	58.35	0.010	**70.33**
TCGA-ESCA	Acc.[%]	47.22	0.009	38.89	0.006	47.22	0.009	**55.56**	0.007	47.22	0.008	47.22	0.0007	47.22	0.014	**55.56**	0.026	50.00	0.009	50.00	0.012	**55.56**	0.021	**55.56**
	AUC[%]	65.00	0.017	48.57	0.002	63.35	0.012	65.09	0.005	64.39	0.004	65.33	0.013	64.68	0.002	65.35	0.007	65.10	0.004	65.09	0.022	65.47	0.006	**65.96**
	F1[%]	38.67	0.017	28.94	0.006	39.09	0.005	37.20	0.009	26.67	0.023	32.46	0.007	41.44	0.008	44.14	0.003	43.52	0.004	35.08	0.006	32.22	0.005	**48.31**
TCGA-LIHC	C-Index	0.522	0.002	0.533	0.010	0.549	0.009	0.546	0.030	0.553	0.027	0.581	0.002	0.575	0.033	0.573	0.011	0.550	0.004	0.529	0.004	0.574	0.016	**0.584**
TCGA-BLCA	C-Index	0.472	0.013	0.561	0.016	0.573	0.019	0.572	0.006	0.571	0.002	0.567	0.007	0.559	0.009	0.572	0.024	0.576	0.002	0.538	0.007	0.575	0.022	**0.579**
TCGA-LUNG	C-Index	0.490	0.007	0.534	0.017	0.537	0.001	0.550	0.009	0.518	0.025	0.535	0.011	0.543	0.013	0.505	0.005	0.525	0.003	0.529	0.009	**0.547**	0.014	0.518

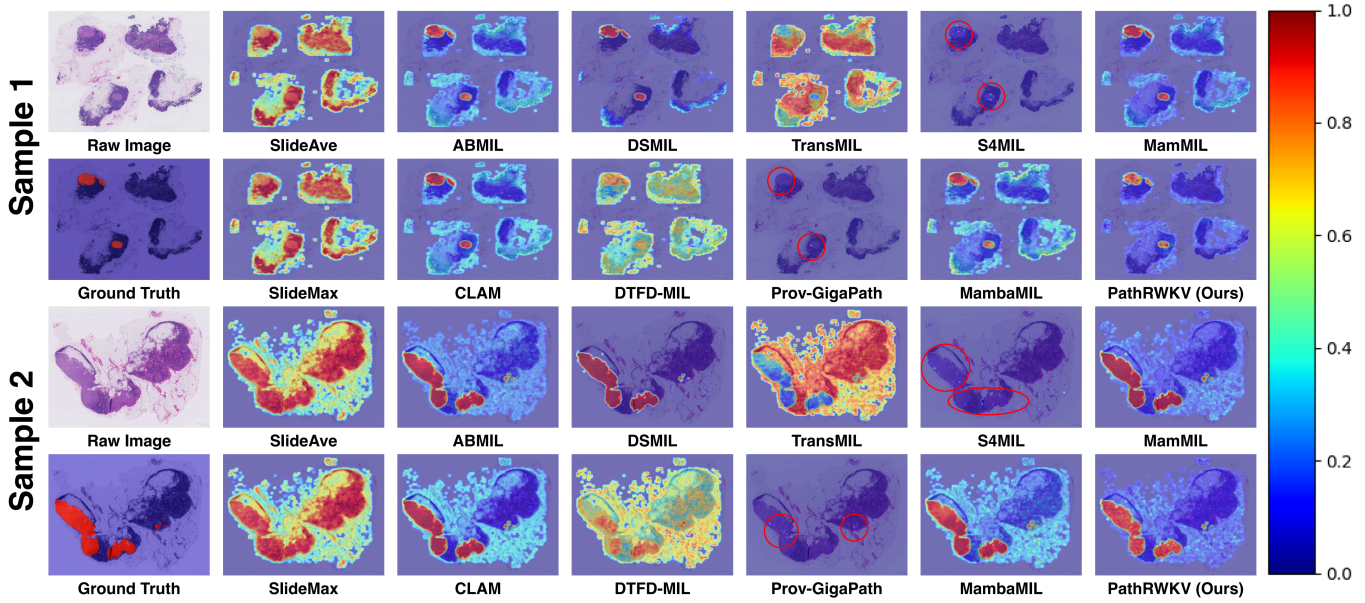

Fig. 4. Visualization of a high-grade lesion WSI sample from the CAMELYON16 dataset. From left to right: the raw image, the ground truth label annotated by the pathologist, feature embeddings extracted by the Prov-GigaPath tile-level encoder, and CAMs from each model.

on the TCGA-LUNG overall survival task, where MambaMIL achieves the leading performance. We attribute this to the inherent trade-off of our max pooling strategy. While it ensures $\mathcal{O}(1)$ inference memory, it focuses on the most salient features and may inadvertently discard global contextual cues (e.g., total tumor burden) that are beneficial for specific prognostic predictions, which Attention mechanism in MambaMIL preserve better. Nevertheless, the strong performance of both PathRWKV and MambaMIL underscores the structural advantage of SSMs over conventional pooling models and Transformers in modeling multi-scale relationships within WSIs.

A core innovation of PathRWKV is its asymmetric design, which fundamentally optimizes GPU memory utilization during inference. Fig. 5 compares the memory consumption profiles of various methods as the number of input tiles increases. Conventional attention-based methods (e.g., ABMIL, CLAM), Transformers (e.g., TransMIL, Prov-GigaPath), and even recent linear-complexity models (e.g., MambaMIL) exhibit linear memory growth ($\mathcal{O}(N)$). Although these SSM-based approaches utilize linear attention mechanisms similar to PathRWKV, they typically rely on the Gated Attention mechanism from ABMIL [13] for final aggregation. This strategy requires matrix multiplication between the complete input and output tensors to compute attention scores, necessitating the retention of the entire input tensor in GPU memory until the output is generated. This dependency causes memory usage to scale linearly with sequence length, effectively negating the inherent efficiency advantages of the SSM backbone and severely constraining applicability to large-scale WSIs in resource-limited environments. In contrast, by equal contribution of linear attention and max aggregation strategy, PathRWKV achieves constant memory consumption ($\mathcal{O}(1)$), as evidenced by the flat trajectory in Fig. 5. This efficiency confirms that our recurrent inference formulation enables sequential iteration over WSI tiles without caching historical states, successfully resolving the trade-off between training efficiency and inference scalability. Notably, while PathRWKV exhibits linear time complexity ($\mathcal{O}(N)$) same with other methods, it maintains a competitive inference speed, effectively balancing its relatively large parameter size for complex feature modeling with computational efficiency.

To further assess the interpretability of our framework, Fig. 4 visualizes the Class Activation Maps (CAMs) for representative samples from the CAMELYON16 dataset. Note

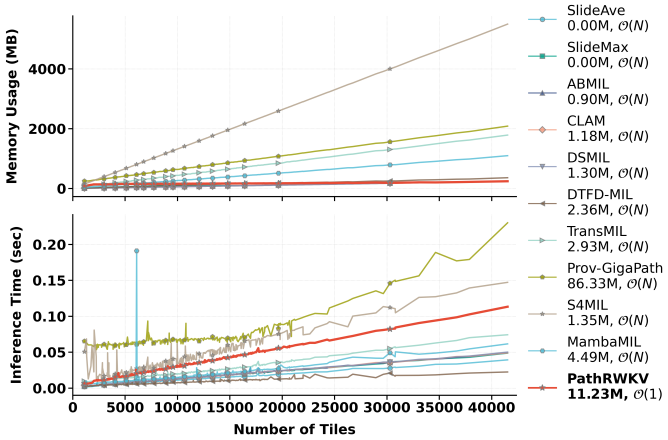

Fig. 5. GPU memory consumption and inference time versus the number of input tiles, where "M" denotes the number of parameters in millions. PathRWKV demonstrates superior efficiency with constant $\mathcal{O}(1)$ memory usage, significantly outperforming all baseline methods that exhibit linear $\mathcal{O}(N)$ memory growth, including SSMs with attention aggregation. Regarding inference speed, despite the inherent linear $\mathcal{O}(N)$ time complexity, PathRWKV achieves competitive and stable performance relative to its parameter size.

that, as standard attention maps are not directly obtainable from Prov-GigaPath and S4MIL, we visualize their saliency maps instead. Consistent with quantitative findings, SlideAve and SlideMax fail to generate meaningful activation patterns due to their simplistic aggregation logic. Compared with the ground truth, attention-based methods (e.g., ABMIL, CLAM) tend to assign uniform weights across the entire region of interest, demonstrating a limited capacity to distinguish fine-grained intratumoral heterogeneity. Furthermore, they exhibit relatively higher attention scores in normal regions. Among transformers, TransMIL misdirects attention to incorrect areas, while Prov-GigaPath focuses on normal regions in the second sample; this is likely attributable to overfitting on the small-scale dataset. In contrast, all SSMs successfully detect the accurate regions. Notably, PathRWKV not only accurately highlights global tumor regions but also delineates local feature intensity variances through its heatmap distribution. This visualization validates the effectiveness of the multi-scale modeling facilitated by the Time Mix and Channel Mix modules, demonstrating PathRWKV's ability to extract hierarchically significant pathological features.

V. Discussion

To rigorously evaluate the contribution of each component within PathRWKV and validate our design choices, we conducted a comprehensive series of ablation studies. The results are summarized in Fig. 6.

A. Validation of Asymmetric Design and Aggregation

A core innovation of PathRWKV is the asymmetric structure, designed to resolve the conflict between training throughput and inference memory efficiency. This design is predicated on two critical hypotheses: first, that a model trained on short sequences can effectively generalize to full-length WSIs during inference; and second, that max pooling serves as a sufficient and efficient aggregator for slide-level features. To validate these premises, we conducted ablation studies on each component.

Inference Scalability. Our asymmetric protocol involves training on a fixed subsample (2,000 tiles) while inferring on the entire WSI (up to 40,000+ tiles). Fig. 6a analyzes the impact of inference sequence length on performance. Despite the potential distribution shift caused by the length discrepancy, PathRWKV exhibits a continuous performance improvement as the number of inference tiles increases. The steepest performance gains coincide with the peak of the WSI tile count distribution (approx. 8,000–12,000 tiles), indicating that the model effectively integrates information from the entire slide. This confirms that our recurrent backbone successfully captures long-range dependencies and generalizes well to sequence lengths far exceeding those seen during training.

Aggregation Strategy. The choice of aggregation function dictates both the representation quality and memory complexity. While Gated Attention (Attn) is the standard in MIL, it necessitates storing all tile features in memory to compute global softmax weights, leading to $\mathcal{O}(N)$ memory usage. In contrast, our proposed streaming max pooling strategy maintains $\mathcal{O}(1)$ complexity. As shown in Fig. 6b, max pooling demonstrates remarkable competitiveness. Compared to Attention, it achieves comparable performance on CAMELYON16 and substantially higher Accuracy and F1 scores on TCGA-GBM (Acc: 0.993 vs. 0.986; F1: 1.000 vs. 0.747). Crucially, max pooling aligns with the worst-pattern diagnostic principle in pathology, where the presence of a specific high-grade lesion often dictates the diagnosis, rendering the global average less relevant. However, we acknowledge the limitation that strictly selecting the maximum feature may discard information regarding tumor burden and the global microenvironment, which are valuable for survival analysis. This is reflected in the TCGA-LIHC task, where Attention slightly outperforms max pooling (0.600 vs. 0.584) by capturing global context. Nevertheless, given the massive efficiency gain ($\mathcal{O}(1)$ vs. $\mathcal{O}(N)$ memory), max pooling represents an optimal trade-off for efficient WSI modeling.

B. Spatial-Temporal Robustness

To address the overfitting risks on small-scale datasets and the loss of spatial structure due to sampling, we introduced random sampling paired with 2D PE.

Sampling Strategy. Fig. 6c compares four sampling strategies. Sequential sampling feeds image features and coordinates according to their original extraction order, maintaining a deterministic sequence. Random sampling introduces a global random permutation, mitigating potential biases associated with the raster scanning order. Z-order sampling [39] utilizes a space-filling curve to retain 2D spatial locality within the flattened 1D sequence. Local-box sampling prioritizes dense local contexts by randomly selecting centroids and querying spatial neighbors. Intuitively, structure-preserving strategies like Sequential or Z-Order might seem superior for an RNN-based model like PathRWKV. However, the empirical results counter-intuitively favor random sampling, which achieves the highest metrics on CAMELYON16 and TCGA-GBM. We

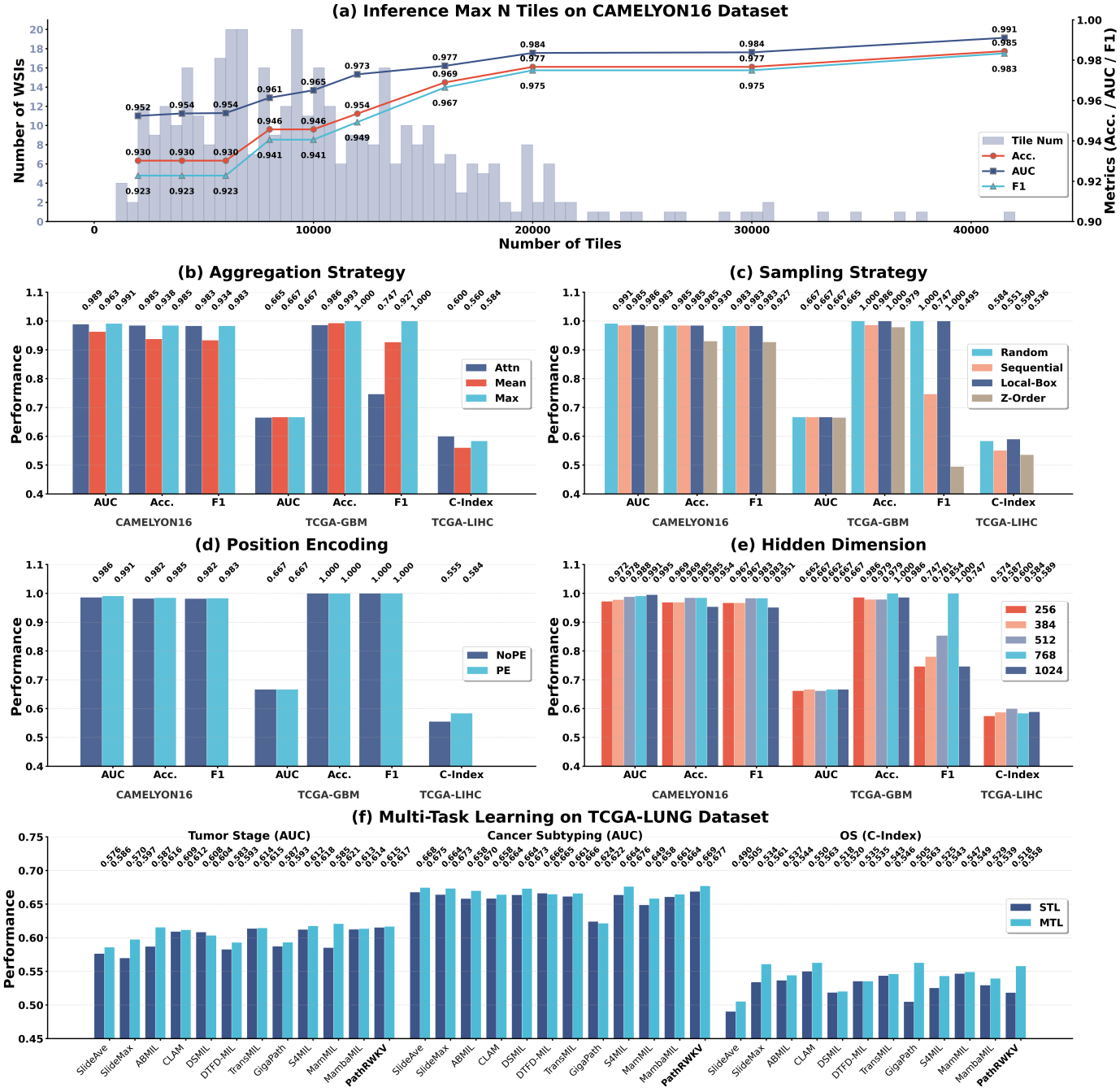

Fig. 6. Ablation studies and performance analysis. (a) Sensitivity analysis of inference performance with respect to the maximum number of input tiles on the CAMELYON16 dataset. (b)–(e) Impact of key components on model performance, including sampling strategies, aggregation methods, position encoding, and hidden dimension sizes across CAMELYON16, TCGA-GBM, and TCGA-LIHC datasets. The baseline strategies are consistently shown in cyan. (f) Comparison between Single-Task Learning (STL) and multi-task learning (MTL) on the TCGA-LUNG dataset for tumor staging, cancer subtyping, and overall survival prediction.

hypothesize that random sampling acts as a potent data augmentation technique, breaking the model's reliance on specific raster-scanning orders and preventing overfitting to incidental sequence patterns. While it disrupts local spatial continuity, it forces the model to learn more robust, permutation-invariant representations.

2D Position Encoding. The efficacy of random sampling is intrinsically linked to the inclusion of 2D PE. As random sampling discards the implicit spatial order, PE is essential for explicitly injecting coordinate information back into the features. Fig. 6d corroborates this, showing that adding PE consistently maintains or enhances performance across tasks (e.g., boosting TCGA-LIHC C-Index from 0.555 to 0.584). This confirms that PathRWKV utilizes these encodings to reconstruct the spatial context of the tissue microenvironment, thereby mitigating the structural information loss caused by random sampling.

C. Generalization and Optimization

Finally, we analyzed the components designed to enhance model generalization under data-constrained conditions.

Hidden Dimension. We investigated the impact of model capacity by varying the hidden dimension D (Fig. 6e). Increasing D does not linearly translate to better performance. While

$D = 1024$ yields the highest AUC on CAMELYON16, it degrades performance on the smaller TCGA-LIHC dataset, likely due to overfitting. Conversely, $D = 768$ offers the optimal balance, achieving the highest performance on TCGA-GBM and competitive results elsewhere. This finding underscores the importance of matching model complexity to the scale of available pathological data, validating our choice of 768 as the default configuration.

Multi-task Learning. The MTL module is designed to regularize feature learning by leveraging auxiliary tasks. Fig. 6f illustrates the performance of various backbones with and without MTL on the TCGA-LUNG dataset. The results demonstrate that our MTL module is a versatile plugin, improving the performance of most baselines (e.g., boosting ABMIL and CLAM). While there are marginal drops in specific cases for single-task specialists (e.g., a 0.4% drop for DSMIL on Tumor Stage), the overall trend signifies that learning shared representations across related clinical tasks effectively reduces overfitting and improves the robustness of the slide-level features.

VI. Conclusion

In conclusion, this work proposed PathRWKV, a novel slide-level modeling framework that introduces an asymmetric training and inference paradigm, provides a principled solution to long-standing challenges in whole slide image analysis. To the best of our knowledge, PathRWKV is the first approach to explicitly decouple slide-level training and inference within a unified structure, enabling robust learning during parallelized training while preserving holistic slide reasoning at inference. Through the integration of asymmetric state space modeling, random tile sampling with multi-task learning regularization, 2D sinusoidal position encoding, and multi-scale feature mixing, PathRWKV effectively addresses weak supervision, data scarcity, disrupted spatial context, and heterogeneous multi-scale feature interactions. Extensive experiments on 29,073 whole slide images across 11 public datasets validating its effectiveness and reliability for clinical-grade pathological inference. We believe PathRWKV establishes a new perspective on slide-level modeling by showing that asymmetry between training and inference is not a limitation, but a powerful inductive principle for scalable and trustworthy computational pathology.

References

[1] N. Ying, Y. Lei, T. Zhang, S. Lyu, C. Li, S. Chen, Z. Liu, Y. Zhao, and G. Zhang, "Cpia dataset: A comprehensive pathological image analysis dataset for self-supervised learning pre-training," *arXiv preprint arXiv:2310.17902*, 2023.

[2] M. G. Hanna, A. Parwani, and S. J. Sirintrapun, "Whole slide imaging: technology and applications," *Advances in Anatomic Pathology*, vol. 27, no. 4, pp. 251–259, 2020.

[3] M. K. K. Niazi, A. V. Parwani, and M. N. Gurcan, "Digital pathology and artificial intelligence," *The lancet oncology*, vol. 20, no. 5, pp. e253–e261, 2019.

[4] B. E. Bejnordi, M. Veta, P. J. Van Diest, B. Van Ginneken, N. Karssemeijer, G. Litjens, J. A. Van Der Laak, M. Hermsen, Q. F. Manson, M. Balkenhol *et al.*, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *Jama*, vol. 318, no. 22, pp. 2199–2210, 2017.

[5] G. Campanella, M. G. Hanna, L. Geneslaw, A. Miraflor, V. Werneck Krauss Silva, K. J. Busam, E. Brogi, V. E. Reuter, D. S. Klimstra, and T. J. Fuchs, "Clinical-grade computational pathology using weakly supervised deep learning on whole slide images," *Nature medicine*, vol. 25, no. 8, pp. 1301–1309, 2019.

[6] M. Y. Lu, D. F. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature Biomedical Engineering*, vol. 5, no. 6, pp. 555–570, 2021.

[7] X. Wang, S. Yang, J. Zhang, M. Wang, J. Zhang, J. Huang, W. Yang, and X. Han, "Transpath: Transformer-based self-supervised learning for histopathological image classification," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, Cham, 2021, pp. 186–195.

[8] T. Nan, S. Zheng, S. Qiao, H. Quan, X. Gao, J. Niu, B. Zheng, C. Guo, Y. Zhang, X. Wang *et al.*, "Deep learning quantifies pathologists' visual patterns for whole slide image diagnosis," *Nature Communications*, vol. 16, no. 1, p. 5493, 2025.

[9] T. Zhang, Y. Feng, Y. Feng, Y. Zhao, Y. Lei, N. Ying, Z. Yan, Y. He, and G. Zhang, "Shuffle instances-based vision transformer for pancreatic cancer rose image classification," *arXiv preprint arXiv:2208.06833*, 2022.

[10] S. Diao, W. Luo, J. Hou, R. Lambo, H. A. Al-Kuhali, H. Zhao, Y. Tian, Y. Xie, N. Zaki, and W. Qin, "Deep multi-magnification similarity learning for histopathological image classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 3, pp. 1535–1545, 2023.

[11] C. Peng, K. Zhao, A. Wiliem, T. Zhang, P. Hobson, A. Jennings, and B. C. Lovell, "To what extent does downsampling, compression, and data scarcity impact renal image analysis?" 2019. [Online]. Available: https://arxiv.org/abs/1909.09945

[12] E. Jenkinson and O. Arandjelović, "Whole slide image understanding in pathology: what is the salient scale of analysis?" *BioMedInformatics*, vol. 4, no. 1, pp. 489–518, 2024.

[13] M. Ilse, J. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," in *International conference on machine learning*. PMLR, 2018, pp. 2127–2136.

[14] H. Xu, N. Usuyama, J. Bagga, S. Zhang, R. Rao, T. Naumann, C. Wong, Z. Gero, J. González, Y. Gu *et al.*, "A whole-slide foundation model for digital pathology from real-world data," *Nature*, pp. 1–8, 2024.

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

[16] H. Zhang, Y. Meng, Y. Zhao, Y. Qiao, X. Yang, S. E. Coupland, and Y. Zheng, "Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18 802–18 812.

[17] J. Wang, Y. Mao, N. Guan, and C. J. Xue, "Advances in multiple instance learning for whole slide image analysis: Techniques, challenges, and future directions," *arXiv preprint arXiv:2408.09476*, 2024.

[18] J. Ding, S. Ma, L. Dong, X. Zhang, S. Huang, W. Wang, N. Zheng, and F. Wei, "Longnet: Scaling transformers to 1,000,000,000 tokens," *arXiv preprint arXiv:2307.02486*, 2023.

[19] B. Peng, D. Goldstein, Q. Anthony, A. Albalak, E. Alcaide, S. Biderman, E. Cheah, T. Ferdinan, H. Hou, P. Kazienko *et al.*, "Eagle and finch: Rwkv with matrix-valued states and dynamic recurrence," *arXiv preprint arXiv:2404.05892*, 2024.

[20] H. Quan, X. Li, D. Hu, T. Nan, and X. Cui, "Dual-channel prototype network for few-shot pathology image classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 7, pp. 4132–4144, 2024.

[21] L. Qu, D. Yang, D. Huang, Q. Guo, R. Luo, S. Zhang, and X. Wang, "Pathology-knowledge enhanced multi-instance prompt learning for few-shot whole slide image classification," in *European conference on computer vision*. Springer, 2024, pp. 196–212.

[22] T. Zhang, Z. Yan, C. Li, N. Ying, Y. Lei, Y. Feng, Y. Zhao, and G. Zhang, "Cellmix: A general instance relationship based method for data augmentation towards pathology image classification," *arXiv preprint arXiv:2301.11513*, 2023.

[23] H. Li, C. Zhu, Y. Zhang, Y. Sun, Z. Shui, W. Kuang, S. Zheng, and L. 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*, 2023, pp. 7454–7463.

[24] S. Yang, Y. Wang, and H. Chen, "Mambamil: Enhancing long sequence modeling with sequence reordering in computational pathology," in

International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2024, pp. 296–306.
[25] M. Kloster, A. M. Burfeid-Castellanos, D. Langenkämper, T. W. Nattkemper, and B. Beszteri, "Improving deep learning-based segmentation of diatoms in gigapixel-sized virtual slides by object-based tile positioning and object integrity constraint," *Plos one*, vol. 18, no. 2, p. e0272103, 2023.
[26] T. Zhang, S. Lyu, Y. Lei, S. Chen, N. Ying, Y. He, Y. Zhao, Y. Feng, H. K. Lee, and G. Zhang, "Puzzletuning: Explicitly bridge pathological and natural image with puzzles," *arXiv preprint arXiv:2311.06712*, 2023.
[27] D. Hu, Z. Dong, K. Liang, H. Yu, S. Wang, and X. Liu, "High-order topology for deep single-cell multiview fuzzy clustering," *IEEE Transactions on Fuzzy Systems*, vol. 32, no. 8, pp. 4448–4459, 2024.
[28] X. Wang, Y. Yan, P. Tang, X. Bai, and W. Liu, "Revisiting multiple instance neural networks," *Pattern recognition*, vol. 74, pp. 15–24, 2018.
[29] B. Li, Y. Li, and K. 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*, 2021, pp. 14 318–14 328.
[30] L. Fillioux, J. Boyd, M. Vakalopoulou, P.-H. Cournède, and S. Christodoulidis, "Structured state space models for multiple instance learning in digital pathology," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2023, pp. 594–604.
[31] Z. Fang, Y. Wang, Y. Zhang, Z. Wang, J. Zhang, X. Ji, and Y. Zhang, "Mammil: Multiple instance learning for whole slide images with state space models," in *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2024, pp. 3200–3205.
[32] D. Liao, S. Chen, N. Xi, Q. Xue, J. Li, L. Hou, Z. Liu, C. H. Low, Y. Wu, Y. Liu, Y. Jiang, D. Li, and S. Lyu, "Unpuzzle: A unified framework for pathology image analysis," 2025. [Online]. Available: https://arxiv.org/abs/2503.03152
[33] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
[34] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[35] W. Bulten, K. Kartasalo, P.-H. C. Chen, P. Ström, H. Pinckaers, K. Nagpal, Y. Cai, D. F. Steiner, H. Van Boven, R. Vink *et al.*, "Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge," *Nature medicine*, vol. 28, no. 1, pp. 154–163, 2022.
[36] S. P. Oliveira, P. C. Neto, J. Fraga, D. Montezuma, A. Monteiro, J. Monteiro, L. Ribeiro, S. Gonçalves, I. M. Pinto, and J. S. Cardoso, "Cad systems for colorectal cancer from wsi are still not ready for clinical acceptance," *Scientific Reports*, vol. 11, no. 1, p. 14358, 2021.
[37] A. Colaprico, T. C. Silva, C. Olsen, L. Garofano, C. Cava, D. Garolini, T. S. Sabedot, T. M. Malta, S. M. Pagnotta, I. Castiglioni *et al.*, "Tcgabiolinks: an r/bioconductor package for integrative analysis of tcga data," *Nucleic acids research*, vol. 44, no. 8, pp. e71–e71, 2016.
[38] B. L. Welch, "The generalization of 'student's' problem when several different population varlances are involved," *Biometrika*, vol. 34, no. 1-2, pp. 28–35, 1947.
[39] J. Peng, J. Jiang, Y. Ying, S. Yun, Q. Long, Y. Zhang, and T. Chen, "One leaf knows autumn: A piece of data-model facilitates efficient cancer prognosis with histological and genomic modalities," in *Workshop on Large Language Models and Generative AI for Health at AAAI 2025*, 2025.

VII. Theoretical Analysis of Asymmetric Structure

In the main text, we introduce an asymmetric structure that utilizes parallel processing during training and recurrent state-passing during inference. Here, we provide the mathematical proof demonstrating that the chunk-based recurrent inference is mathematically equivalent to processing the entire WSI sequence in a single pass.

It is worth noting that despite the difference in execution modes, both approaches rely on the same fundamental linear operations without introducing any complex approximations. As revealed by the non-trivial derivations of the closed-form solution (Eq. (16)) and the state-passing mechanism (Eq. (20)) below, the memory of the model is mathematically preserved through simple linear decays. This ensures that no global context information is lost due to the chunking strategy.

Recall the core state update rule of the PathRWKV block defined in Eq. (4) of the main text. For a sequence of tiles indexed by t, the hidden state matrix S_t and the output y_t are computed as:

$$S_t = w_t \odot S_{t-1} + k_t v_t^\top \tag{11}$$

$$y_t = r_t \odot (S_{t-1} + u \odot k_t v_t^\top) \tag{12}$$

where $\odot$ denotes element-wise multiplication, and w_t represents the data-dependent decay at step t. S_0 is initialized as a zero matrix.

Proposition VII.1 (Associativity and Inference Equivalence)**.** *Let $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$ be the complete sequence of tiles from a WSI. Let $\mathcal{A}$ denote the computation of the final state S_N by processing all tiles continuously:*

$$S_N = \Phi(\mathcal{X}, S_0) \tag{13}$$

Let $\mathcal{B}$ denote the computation where the sequence is split into two contiguous chunks $\mathcal{X}_1 = \{x_1, \ldots, x_M\}$ and $\mathcal{X}_2 = \{x_{M+1}, \ldots, x_N\}$ (where $1 < M < N$). The inference is performed sequentially by passing the intermediate state:

$$S_M = \Phi(\mathcal{X}_1, S_0) \tag{14}$$

$$S'_N = \Phi(\mathcal{X}_2, S_M) \tag{15}$$

Then, the final states are identical: $S_N = S'_N$.

Proof. The recursive update rule in Eq. (11) implies that the current state depends on the immediate past state, which in turn depends on its predecessor. By recursively unrolling this dependency back to the initial step $t = 1$, we can observe a pattern: the contribution of an input $k_i v_i^\top$ at step i to the current state S_t is scaled by the cumulative product of all subsequent decay factors. Mathematically, this accumulation allows us to express S_t in a non-trivial closed form:

$$S_t = \sum_{i=1}^{t} \left(\prod_{j=i+1}^{t} w_j \right) \odot (k_i v_i^\top) + \left(\prod_{j=1}^{t} w_j \right) \odot S_0 \tag{16}$$

Assuming $S_0 = \mathbf{0}$, the term involving S_0 vanishes.

Case 1: Global Continuous Inference ($\mathcal{A}$) Applying Eq. (16) to the full sequence $t = N$:

$$S_N = \sum_{i=1}^{N} \left(\prod_{j=i+1}^{N} w_j \right) \odot (k_i v_i^\top) \tag{17}$$

We can split this summation into two parts at index M.

$$\begin{aligned} S_N = & \underbrace{\sum_{i=1}^{M} \left(\prod_{j=i+1}^{N} w_j \right) \odot k_i v_i^\top}_{\text{Part 1}} \\ & + \underbrace{\sum_{i=M+1}^{N} \left(\prod_{j=i+1}^{N} w_j \right) \odot k_i v_i^\top}_{\text{Part 2}} \end{aligned} \tag{18}$$

Notice that for Part 1, the decay product can be factored: $\prod_{j=i+1}^{N} w_j = (\prod_{j=M+1}^{N} w_j) \odot (\prod_{j=i+1}^{M} w_j)$.

Case 2: Chunked Sequential Inference ($\mathcal{B}$) First, compute the state S_M after the first chunk $\mathcal{X}_1$:

$$S_M = \sum_{i=1}^{M} \left(\prod_{j=i+1}^{M} w_j \right) \odot k_i v_i^\top \tag{19}$$

Next, use S_M as the initial state for the second chunk $\mathcal{X}_2$. We apply the recursive definition starting from step $M+1$ to N. Here, we treat S_M similarly to S_0 in Eq. (16), but with a crucial difference: the historical information carried by S_M must continue to decay as it propagates through the new sequence from $M+1$ to N. By induction, the final state S'_N comprises two components: the decayed history from the previous chunk and the accumulated information from the current chunk:

$$S'_N = \left(\prod_{j=M+1}^{N} w_j \right) \odot S_M + \sum_{i=M+1}^{N} \left(\prod_{j=i+1}^{N} w_j \right) \odot k_i v_i^\top \tag{20}$$

Substitute S_M into the equation above:

$$\begin{aligned} S'_N = &\left(\prod_{j=M+1}^{N} w_j \right) \odot \left[\sum_{i=1}^{M} \left(\prod_{j=i+1}^{M} w_j \right) \odot k_i v_i^\top \right] \\ &+ \sum_{i=M+1}^{N} \left(\prod_{j=i+1}^{N} w_j \right) \odot k_i v_i^\top \end{aligned} \tag{21}$$

Distributing the decay term $\prod_{j=M+1}^{N} w_j$ into the summation bracket exactly reconstructs the Part 1 term from Eq. (18), and the second term is identical to Part 2.

$$S'_N \equiv S_N \tag{22}$$

Thus, chunked inference with state passing is mathematically exact to global inference. □

VIII. Implementation Details and Hardware Acceleration

In this section, we provide a detailed description of the implementation of the PathRWKV backbone, specifically focusing on the custom CUDA kernels designed to enable the asymmetric training and inference structure described in Section III.

A. Custom CUDA Kernels

To efficiently implement the mathematical duality of the Linear Attention mechanism, we implemented two distinct sets of CUDA kernels, corresponding to the parallel (training) and recurrent (inference) modes.

Parallel Kernel for Training. During training, we utilize the **wkv6 parallel** kernel. This kernel is optimized for maximizing throughput when the entire sequence is available in memory. It implements the time-decayed aggregation described in Eq. 3. Crucially, it fuses the computation of receptance (r), key (k), value (v), and time-decay (w) processing into a single GPU kernel to minimize memory access overhead (HBM reads/writes). The backward pass kernel analytically computes gradients for all parameters, including the time-dependent decay rates, ensuring stable backpropagation through long sequences without the vanishing gradient problem typical of RNNs. The kernel leverages shared memory tiling and loop unrolling to accelerate the accumulation of attention scores along the sequence dimension T.

State-based Kernel for Inference. For inference on gigapixel WSIs, we utilize the **wkv6 state** kernel. This kernel explicitly manages the recurrent state to support the chunked processing strategy. Unlike standard attention kernels, this kernel accepts an additional input tensor S_{in} (the hidden state from the previous chunk) and outputs S_{out} (the updated state). This directly implements the update rule derived in Proposition VII.1. To further accelerate inference, the state update loop is vectorized using `float4` data types, allowing the GPU to process 4 floating-point numbers simultaneously per thread. The kernel performs internal accumulation in `float32` to maintain numerical precision during the recursive updates, preventing error accumulation over long WSI sequences.

B. Chunked Inference Implementation

The code implementation orchestrates the interaction between the CUDA kernel and the WSI data to minimize memory footprint. Let N be the total number of tiles and B_{chunk} be the chunk size. The inference process proceeds as follows:

Initialization. A state tensor S of shape $(B, H, N_{head}, N_{head})$ is initialized to zeros, where H is the number of heads and N_{head} is the head dimension.

Sequential Processing. The WSI is split into $\lceil N/B_{chunk} \rceil$ chunks. For each chunk k:

$$Y_k, S_k = \text{Block}(X_k, S_{k-1}) \tag{23}$$

Here, Block calls the **wkv6 state** CUDA kernel, and S_k represents the hidden state after processing the k-th chunk. The state S is passed strictly from CPU to GPU memory only once per chunk, minimizing PCI-e bandwidth usage.

Streaming Aggregation. Concurrently with feature extraction, the slide-level representation is updated using the streaming max pooling aggregation: $h_{global} = \max(h_{global}, \max(Y_k))$. This ensures the GPU memory usage remains constant $\mathcal{O}(1)$ regardless of slide size.