# ITC-RWKV: Interactive Tissue–Cell Modeling with Recurrent Key Value Aggregation for Histopathological Subtyping

Yating Huang, Qijun Yating, Lintao Xiang, Hujun Yin

BMVC 2025

MANCHESTER 1824 The University of Manchester


## Abstract

**➢ Challenges**

- Existing pathology models often ignore cell-level morphology and rely solely on patch-level features, resulting in the loss of fine-grained nuclear cues.
- Quadratic attention cost in standard Transformers makes it impractical to aggregate thousands of cells within dense tissue regions.
- Accurate diagnosis equires contextual understanding of cell–tissue interactions, which most models fail to capture.

**➢ Contributions**

- **Dual-stream ITC-RWKV architecture** ➜ Jointly encodes cell- and tissue-level information, mimicking expert pathology reasoning.
- **Aggr-RWKV: Receptance-Weighted Key-Value Aggregation** ➜ Efficiently aggregates large cell sets with linear complexity while retaining morphology.
- **Bidirectional Tissue–Cell Interaction Module** ➜ Enables mutual attention between cellular cues and tissue context for improved accuracy and interpretability.


## Methodology

Digital pathology diagnosis requires analysis across two complementary scales: individual cellular morphology and global tissue architecture. Our framework addresses this challenge through a dual-stream design that mimics the diagnostic process of pathologists. As illustrated in **Figure 1**, it consists of three key components: (i) a cell pathway (**Figure 1a**) that processes individual nuclei and aggregates their features using the proposed Aggr-RWKV module (**Figure 1d**), (ii) a tissue pathway (**Figure 1b**) leveraging the pathological foundation model for contextual understanding, and (iii) a tissue-cell interaction module (**Figure 1c**) enabling cross-scale information exchange.

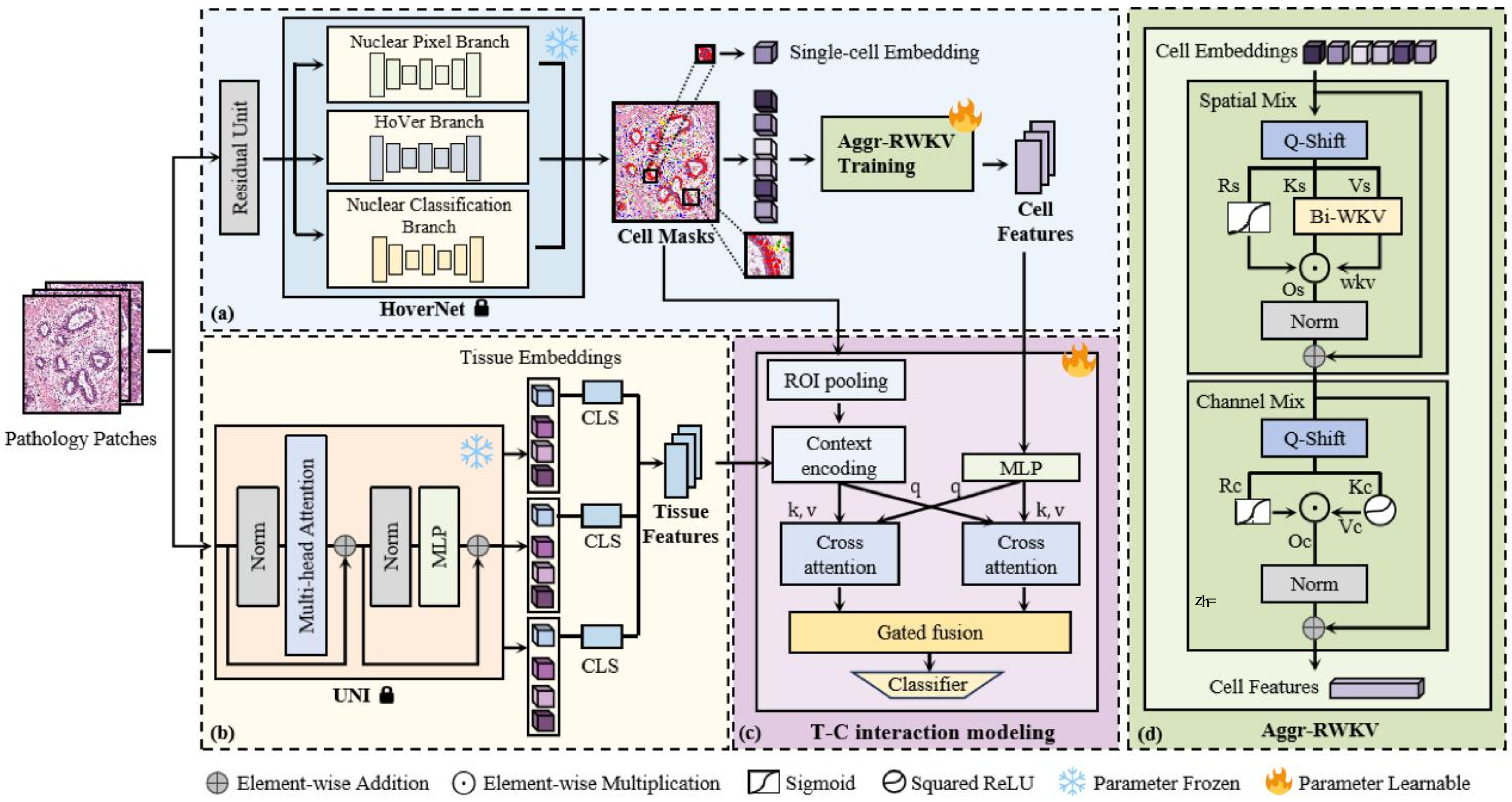


Figure 1: Overview of the proposed ITC-RWKV model.

### ➢ Cell Pathway with Aggr-RWKV

**Spatial Mix:**

The input features (cell embeddings) H are first processed by Q-Shift to obtain shifted features $H_{shifted}$. As shown in Figure 1d, three independent linear projections are then applied to generate the receptance vector $R_s$, key vector $K_s$ and value vector $V_s$. With linear complexity, the Bi-WKVmechanism recurrently aggregates Ks and Vs in both forward and backward directions, functioning like a recurrent model to capture bidirectional context and produce the output wkv and the final output $H_s$

$$\mathbf{wkv = \text{Bi-WKV}(K_s, V_s)} \qquad \mathbf{O_s = \sigma(R_s)\odot wkv,\ H_s = H + \text{LayerNorm}(O_s)}$$

**Channel Mix:**

This module models intra-feature interactions from the previous output $H_s$. The input is first processed by Q-Shift to obtain $H_{s\_shifted}$, which is linearly projected into a receptance vector $R_c = H_{s\_shifted}W_r$ and Key vector $K_c = H_{s\_shifted}W_k$. The output $O_c$ is then computed by applying a SquaredReLU activation to $K_c$ and gating it with $\sigma(R_c)$ through element-wise multiplication:

$$\mathbf{O_c = \sigma(Rc)\odot \text{SquaredReLU}(K_c)}$$

The final output of the entire Aggr-RWKV block is obtained from the output of the channel mixing stage after layer normalization and a residual connection:

$$\mathbf{H_{cell} = H_s + \text{LayerNorm}(O_c)}$$

This updated feature matrix contains enhanced cell representations that incorporate inter-cell relationship information and feature-level interactions. By stacking multiple Aggr-RWKV blocks, cell representations are progressively refined and ultimately aggregated into a global cell feature vector for downstream tasks.

### ➢ Tissue-Cell Interaction Modeling

We begin by aligning spatially corresponding features from the two branches. For each nucleus with mask $R_k$, we extract its contextual tissue representation $r_k$ by applying ROI pooling over the token features from the tissue branch. This pooling aggregates token embeddings from the UNI model that overlap with the cell's position and its local microenvironment. As a result, we obtain paired sequences: $\{h_k\}$ from the cell pathway, and corresponding tissue contexts $\{r_j\}$ . To enable information exchange between the two modalities, we perform dual cross-attention: $\mathbf{\tilde{h}_k = Attn(h_k, \{r_j\}, \{r_j\})}$ $\mathbf{\tilde{r}_k = Attn(r_k, \{h_j\}, \{h_j\})}$. After cross-attention, we aggregate the context-enriched cell features $\{\tilde{h}_k\}$ into a global cellular representation c. Similarly, the cell-aware tissue features $\{\tilde{r}_k\}$ are combined with the [CLS] token embedding $t_{[CLS]}$ to form a comprehensive tissue representation. The attended outputs are then combined through our tissue-cell interaction function φ, which implements a gated fusion mechanism:

$$z = \sigma(W_g[c; t_{[CLS]}])\odot c + (1 - \sigma(W_g[c; t_{[CLS]}]))\odot t_{[CLS]}$$

## Experiments

We evaluated our method on four public histopathology benchmarks covering two cancer types and multiple domain shifts.

| Model | Normal | Benign | UDH | ADH | FEA | DCIS | IC | Total |
|---|---|---|---|---|---|---|---|---|
| CGC-Net[44] | 30.8 ± 5.3 | 31.6 ± 4.7 | 17.3 ± 3.4 | 24.5 ± 5.2 | 59.0 ± 3.6 | 49.4 ± 3.4 | 75.3 ± 3.2 | 43.6 ± 0.5 |
| Patch-GNN[4] | 52.5 ± 3.3 | 47.6 ± 2.2 | 23.7 ± 4.6 | 30.7 ± 1.8 | 60.7 ± 5.3 | 58.8 ± 1.1 | 81.6 ± 2.2 | 52.1 ± 0.6 |
| CG-GNN[31] | 63.6 ± 4.9 | 47.7 ± 3.1 | 34.7 ± 4.9 | 28.5 ± 4.3 | 72.1 ± 3.6 | 54.6 ± 3.2 | 82.2 ± 4.0 | 56.6 ± 1.3 |
| HACT-Net[31] | 61.6 ± 2.1 | 47.5 ± 2.9 | 43.6 ± 1.9 | 40.4 ± 2.5 | 74.2 ± 4.6 | 66.4 ± 3.6 | 88.4 ± 0.2 | 61.5 ± 0.9 |
| CLAM[28] | 59.4 ± 2.0 | 47.7 ± 1.2 | 31.7 ± 0.7 | 20.1 ± 3.4 | 68.3 ± 4.0 | 59.9 ± 1.7 | 86.8 ± 3.6 | 54.8 ± 1.0 |
| TransMIL[34] | 47.6 ± 9.8 | 42.9 ± 3.6 | 31.5 ± 5.3 | 38.4 ± 5.9 | 72.7 ± 3.6 | 62.7 ± 2.9 | 87.1 ± 3.9 | 57.5 ± 0.7 |
| ScoreNet[35] | 64.3 ± 1.5 | **54.0 ± 2.2** | 45.3 ± 3.4 | **46.7 ± 1.0** | 78.1 ± 2.8 | 62.9 ± 2.0 | 91.0 ± 1.4 | 64.4 ± 0.9 |
| **Ours** | **76.3 ± 3.7** | 51.6 ± 1.5 | **47.5 ± 2.8** | 45.0 ± 2.6 | **80.3 ± 3.5** | **67.1 ± 2.3** | **94.6 ± 1.2** | **66.5 ± 0.8** |

Table 1: Comparison with the prior art for breast cancer subtyping on the BRACS dataset.

| Aggregator | Complexity | GPU Mem (GB) | Latency (ms) | Throughput (patch/s) | GFLOPs (G) | Speed-up (↑) | Weighted F1 (%) |
|---|---|---|---|---|---|---|---|
| Self-Attention[37] | $\mathcal{O}(n^2)$ | 5.2 | 42.8 | 23 | 10.52 | 1.0× | 64.2 |
| Set Transformer[26] | $\mathcal{O}(n^2)$ | 4.8 | 38.4 | 26 | 9.47 | 1.1× | 65.7 |
| DeepSets[42] (mean) | $\mathcal{O}(n)$ | 2.1 | 15.6 | 64 | 3.82 | 2.8× | 62.3 |
| **Aggr-RWKV (ours)** | $\mathcal{O}(n)$ | 2.4 | 16.8 | 71 | 4.09 | 3.1× | 66.5 |

Table 2: Comparison of cell aggregation methods.

| Experiments | In-domain | | Cross-domain | |
|---|---|---|---|---|
| | Acc (%) | F1 | Acc (%) | F1 |
| ResNet-50 [17] | 78.3 | 0.656 | 70.9 | 0.651 |
| ViT-B32 [11] | 77.4 | 0.665 | **72.4** | 0.637 |
| CTransPath [39] | 64.5 | 0.643 | 61.1 | 0.614 |
| DCAH-Net [43] | 54.1 | 0.420 | 61.9 | 0.526 |
| HiFuse [20] | 62.7 | 0.457 | 61.2 | 0.501 |
| **ITC-RWKV (ours)** | **79.5** | **0.673** | 72.1 | **0.662** |

Table 3: Generalization performance on two prostate cancer test sets.

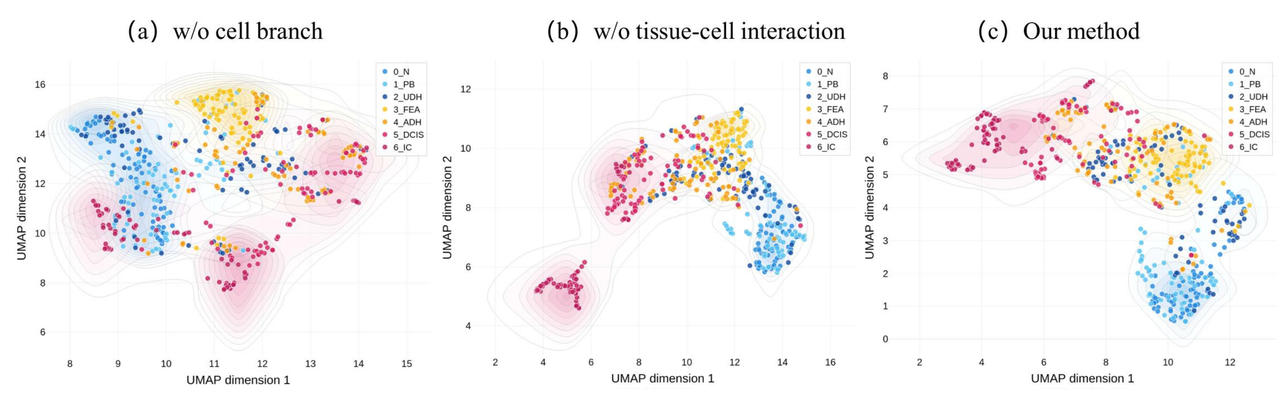


Figure 2: UMAP visualization of feature embeddings from the ablation study: (a) without cell branch, (b) without tissue–cell interaction (simple concatenation), and (c) full model.

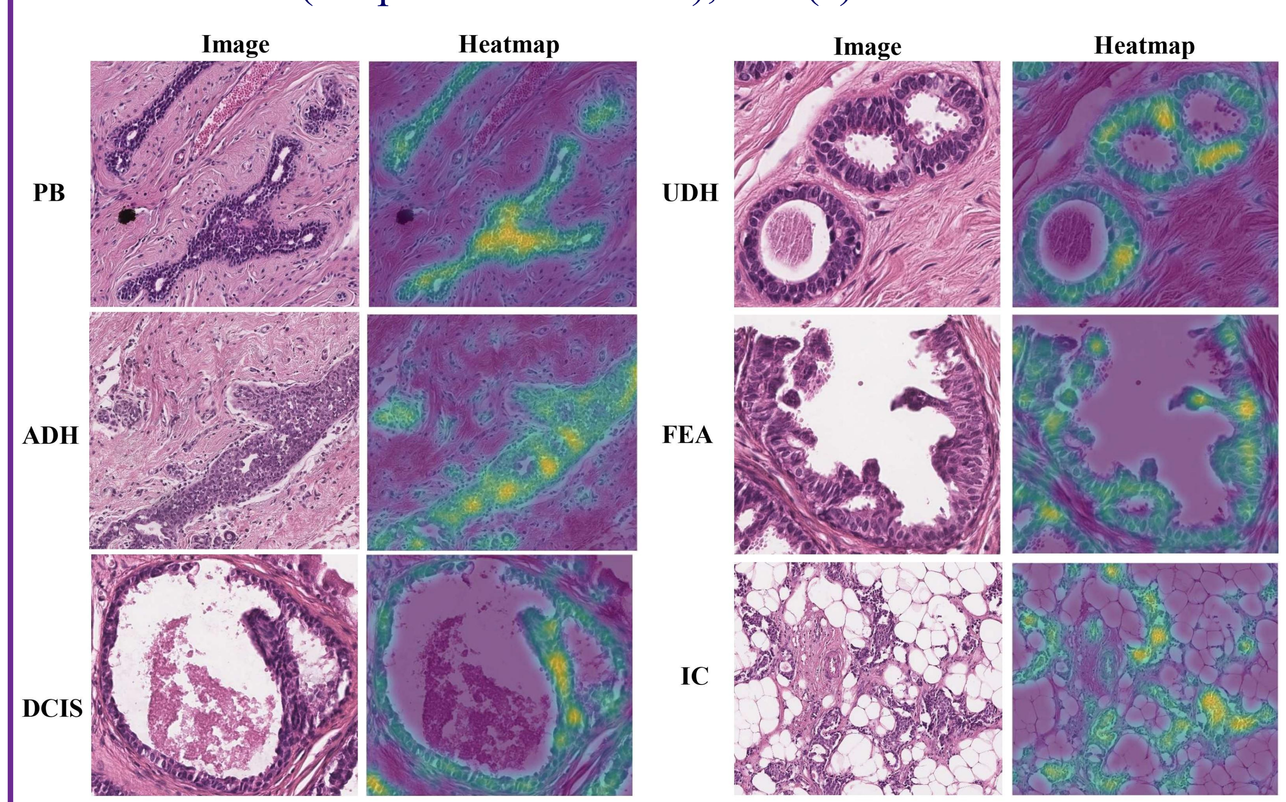


Figure 3: Tissue region importance heatmaps across different breast pathologies. For each pair, the left shows original H&E images and the right shows corresponding importance maps where yellow-green highlights indicate regions most influential for classification.

## Conclusion

**➢ Bridging micro- and macro-level reasoning:**

The proposed dual-stream framework, combining the Aggr-RWKV module and the tissue–cell interaction mechanism, effectively integrates cell-level and tissue-level information for histopathological analysis.

**➢ Scalability and flexibility:**

The linear-complexity Aggr-RWKV ensures scalability to dense cellular regions, while the interaction module adapts well to diverse diagnostic scenarios.

**➢ Future directions:**

Future work will extend the framework to whole-slide image inference and explore integration with spatial priors or multimodal clinical data.