Dear Editor and Reviewers,

Thank you for your hard work on our manuscript “DualSeg: Unified Multi-Scale Framework With Dual-Stage Encoder For Glomerular Segmentation” (ID: [JBHI-05124-2025]). We have carefully considered all your comments and made changes to the manuscript's content. In this revised version, we have significantly expanded our experimental validation by including new state-of-the-art baselines (e.g., InceptionNeXt, U-Mamba), adding a new external validation dataset (KPMP), and visualizing Effective Receptive Fields (ERF) to clarify our architectural novelty. Please note that as the manuscript was extensively condensed (from 17 to 14 pages) to meet JBHI page limits, we have restricted highlighting to substantive changes. Our responses to the comments are provided below.

**Response to reviewer #1:**

COMMENTS TO THE AUTHOR(S)

1. **Experimental Validation and Efficiency Claims**

   The paper's central efficiency claim—60% computational reduction via VRWKV—remains entirely unsubstantiated. A core contribution built on efficiency must provide comprehensive benchmarks including FLOPs, memory consumption, and inference latency across varying input sizes. The absence of these fundamental metrics suggests either inadequate experimental rigor or awareness that actual gains may not support the claimed advantages. This gap is particularly damaging given that computational efficiency differentiates DualSeg from existing methods.

**Response:** Thank you for your advice. We have addressed the concerns regarding computational benchmarking through the following revisions, as detailed below:

- **Text Revision and Clarification**: We have revised the text in **Section I. Introduction** to remove the claim regarding a "60% reduction in computational overhead." This figure was originally cited from the VRWKV framework literature (Reference [23] *Y. Duan, W. Wang, Z. Chen, X. Zhu, L. Lu, T. Lu, Y. Qiao, H. Li, J. Dai, and W. Wang, “Vision-rwkv: Efficient and scalable visual perception with rwkv-like architectures,” arXiv preprint arXiv:2403.02308, 2024.*) to illustrate general efficiency, and this statement has been deleted.
- **Expanded FLOPs Comparison**: To provide a concrete evaluation of computational efficiency, we have introduced a new comparative analysis in the revised manuscript. Specifically, we have added a **FLOPs (Floating Point Operations) vs. Dice Score comparison** across different models (now illustrated in the bottom panel of **Figure 1 and Figure 3**). This metric provides a standard and objective measure of the computational complexity and the performance-efficiency trade-off of our proposed framework.

The revisions can be found in Section I. Introduction, Figure 1 and 3:

> **Page 2, Section I. Introduction:**
>
> Conversely, Vision Transformers (ViTs) tackle spatial heterogeneity by leveraging self-attention for global context modeling [20], [21] (Fig. 1(II)). While ViTs typically outperform CNNs in maintaining structural continuity, the quadratic computational complexity of self-attention incurs high FLOPs, limiting their scalability in high-resolution histology [10]. To mitigate this, computationally efficient alternatives like Wave-MLP [22] and VRWKV [23] (employing linear-time recurrent kernels) have emerged. Similarly, VM-UNet [24] introduced State Space Models (SSM) to balance computational cost and segmentation performance.
>
> …
>
> To address this, we propose **DualSeg**, a novel hybrid framework that synergizes Wave Vision and VRWKV within a pyramid structure (Fig. 2(f)). Our architecture features a dual-stage encoder where early-stage Wave-Swin blocks execute hierarchical local feature extraction to enhance texture discriminability, followed by

VRWKV blocks that model long-range dependencies via linear attention to resolve spatial heterogeneity. This design effectively amalgamates the texture sensitivity of CNNs, the generalization of MLPs, and the scalability of VRWKV. By bridging local and global processing, DualSeg achieves robust multi-scale mapping of morphological priors. As shown in Fig. 1(IV) and Bottom, our method attains a global, clean ERF and SOTA segmentation performance. Notably, it achieves this with significantly reduced computational overhead (9.51 G FLOPs), offering an optimal accuracy-complexity trade-off superior to resource-intensive baselines.

**Page 1, Figure 1:**

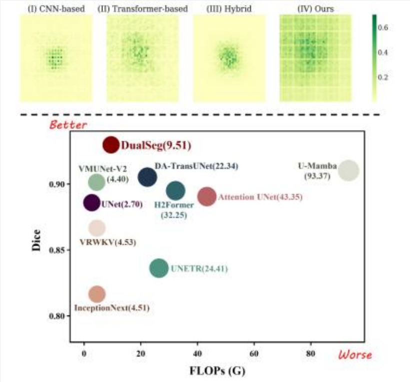

Fig. 1: **Top:** Comparison of *Effective Receptive Fields* (ERF). CNNs (I) show restricted local focus; Transformers (II) display dispersed global patterns; and current Hybrids (III) demonstrate subtimal coverage, failing to form a full context. Our method (IV) overcomes these limitations to generate a dense, global ERF. **Bottom:** Trade-off analysis. DualSeg occupies the optimal top-left position, delivering SOTA accuracy with significantly lower computational complexity (FLOPs) compared to existing methods.

2. **Dataset Limitations and Clinical Generalizability**

   The evaluation's restriction to PAS-stained specimens fundamentally limits clinical relevance. Real-world pathology predominantly uses H&E staining with supplementary protocols (Masson's trichrome, Jones silver) for specific diagnoses. Each staining method reveals distinct tissue characteristics—algorithms optimized for PAS often fail catastrophically on H&E due to different contrast patterns and feature visibility. The absence of cross-institutional validation or multi-protocol testing renders clinical applicability claims premature.

**Response:** Thank you for pointing that out. We have revised the manuscript to strengthen the evaluation of generalizability and address staining limitations, as detailed below:

- **Expanded Validation Experiments:** We have incorporated a new external dataset—the **KPMP (Kidney Precision Medicine Project) dataset** (now detailed in **Section IV. A. Dataset**). This allowed us to perform an extensive validation encompassing: ***Cross-Institutional Validation:*** Testing on data from completely different centers without any fine-tuning (zero-shot inference). The results demonstrate that even when restricted to PAS staining, DualSeg maintains high robustness against significant biological and technical variations across different cohorts. And the results are illustrated in **Table III and Figure 8**.
- **Text Revision on Staining Limitations:** We have added a discussion in **Section VI. D. Limitations and Future Work**, regarding the scarcity of synchronized multi-stain (e.g., H&E, Masson) annotations in current public datasets. The revised text explicitly states that while DualSeg demonstrates superior performance on PAS-stained images, its clinical deployment across diverse staining protocols remains a subject for future validation as relevant data becomes available.

The revisions can be found in Section IV. A. Dataset, Section VI. D. Limitations and Future Work, Table III and Figure 8:

## Page 7, Section IV. A. Datasets:

Dataset III: Human Glomeruli (KPMP). To assess cross species generalization, we retrieved a second human dataset from the Kidney Precision Medicine Project (KPMP) Atlas Repository [50]. Four PAS-stained SVS format WSIs (avg. resolution 84,000×50,000) were selected with corresponding masks. To rigorously validate generalization, models trained solely on the mouse KPIs dataset were directly applied to this human dataset without retraining. For preprocessing consistency, KPMP WSIs were partitioned into 2,048 × 2,048 patches.

## Page 12, Section VI. D. Limitations and Future Work:

While DualSeg demonstrates superior performance in glomerular segmentation, three primary limitations remain to be addressed in future iterations. First, the model's generalization to rare pathological subtypes, such as global glomerulosclerosis, is currently constrained by data scarcity. We plan to mitigate this by employing domain adaptation and transfer learning techniques to enhance feature robustness for these underrepresented classes. Second, our current validation is restricted to PAS-stained images. To ensure broad clinical applicability, we will extend our evaluation to include Hematoxylin-Eosin (HE) stained datasets, verifying the model's adaptability to varying histological protocols. Finally, despite the linear complexity of the VRWKV block, the computational overhead for gigapixel WSI processing remains significant. Future work will focus on model quantization and lightweight optimization to facilitate deployment on resource-constrained platforms and edge devices.

## Page 6, Table III:

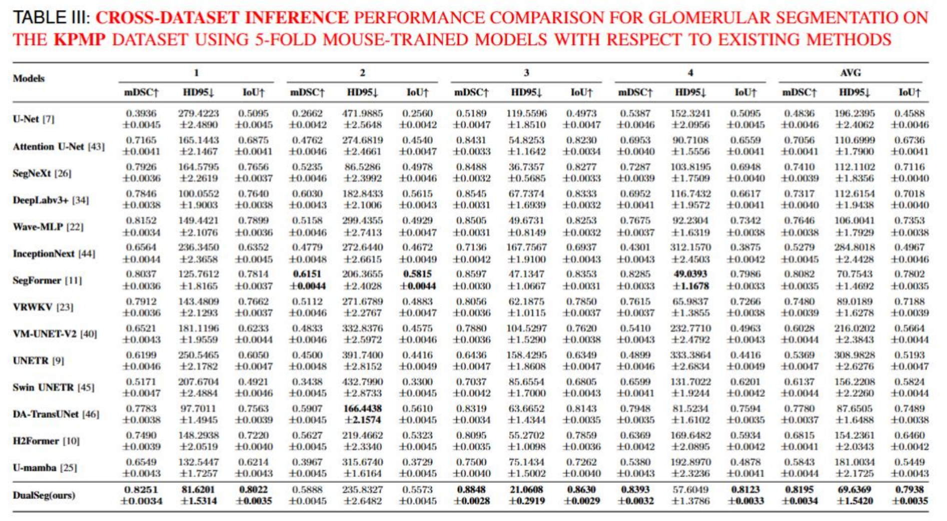

TABLE III: **CROSS-DATASET INFERENCE** PERFORMANCE COMPARISON FOR GLOMERULAR SEGMENTATIO ON THE **KPMP** DATASET USING 5-FOLD MOUSE-TRAINED MODELS WITH RESPECT TO EXISTING METHODS

| Models | 1 | | | 2 | | | 3 | | | 4 | | | AVG | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | mDSC↑ | HD95↓ | IoU↑ | mDSC↑ | HD95↓ | IoU↑ | mDSC↑ | HD95↓ | IoU↑ | mDSC↑ | HD95↓ | IoU↑ | mDSC↑ | HD95↓ | IoU↑ |
| U-Net [7] | 0.3936<br>±0.0045 | 279.4223<br>±2.4890 | 0.5095<br>±0.0045 | 0.2662<br>±0.0042 | 471.9885<br>±2.5648 | 0.2560<br>±0.0042 | 0.5189<br>±0.0047 | 119.5596<br>±1.8510 | 0.4973<br>±0.0047 | 0.5387<br>±0.0046 | 152.3241<br>±2.0956 | 0.5095<br>±0.0045 | 0.4836<br>±0.0046 | 196.2395<br>±2.4062 | 0.4588<br>±0.0046 |
| Attention U-Net [43] | 0.7165<br>±0.0041 | 165.1443<br>±2.1467 | 0.6875<br>±0.0041 | 0.4762<br>±0.0046 | 274.6819<br>±2.4661 | 0.4540<br>±0.0047 | 0.8431<br>±0.0033 | 54.8253<br>±1.1642 | 0.8230<br>±0.0034 | 0.6953<br>±0.0040 | 90.7108<br>±1.5556 | 0.6559<br>±0.0041 | 0.7056<br>±0.0041 | 110.6999<br>±1.7900 | 0.6736<br>±0.0041 |
| SegNeXt [26] | 0.7926<br>±0.0036 | 164.5795<br>±2.2619 | 0.7656<br>±0.0037 | 0.5235<br>±0.0046 | 86.5286<br>±2.3992 | 0.4978<br>±0.0046 | 0.8488<br>±0.0032 | 36.7357<br>±0.5685 | 0.8277<br>±0.0033 | 0.7287<br>±0.0039 | 103.8195<br>±1.7509 | 0.6948<br>±0.0040 | 0.7410<br>±0.0039 | 112.1102<br>±1.8356 | 0.7116<br>±0.0040 |
| DeepLabv3+ [34] | 0.7846<br>±0.0038 | 100.0552<br>±1.9003 | 0.7640<br>±0.0038 | 0.6030<br>±0.0043 | 182.8433<br>±2.1006 | 0.5615<br>±0.0043 | 0.8545<br>±0.0031 | 67.7374<br>±1.6939 | 0.8333<br>±0.0032 | 0.6952<br>±0.0041 | 116.7432<br>±1.9572 | 0.6617<br>±0.0041 | 0.7317<br>±0.0040 | 112.6154<br>±1.9438 | 0.7018<br>±0.0040 |
| Wave-MLP [22] | 0.8152<br>±0.0034 | 149.4421<br>±2.1076 | 0.7899<br>±0.0036 | 0.5158<br>±0.0046 | 299.4355<br>±2.7413 | 0.4929<br>±0.0047 | 0.8505<br>±0.0031 | 49.6731<br>±0.8149 | 0.8253<br>±0.0032 | 0.7675<br>±0.0037 | 92.2304<br>±1.6319 | 0.7342<br>±0.0038 | 0.7646<br>±0.0038 | 106.0041<br>±1.7929 | 0.7353<br>±0.0038 |
| InceptionNext [44] | 0.6564<br>±0.0044 | 236.3450<br>±2.3658 | 0.6352<br>±0.0045 | 0.4779<br>±0.0048 | 272.6440<br>±2.6615 | 0.4672<br>±0.0049 | 0.7136<br>±0.0042 | 167.7567<br>±1.9100 | 0.6937<br>±0.0043 | 0.4301<br>±0.0043 | 312.1570<br>±2.4503 | 0.3875<br>±0.0042 | 0.5279<br>±0.0045 | 284.8018<br>±2.4428 | 0.4967<br>±0.0046 |
| SegFormer [11] | 0.8037<br>±0.0036 | 125.7612<br>±1.8165 | 0.7814<br>±0.0037 | **0.6151**<br>**±0.0044** | 206.3655<br>±2.4028 | **0.5815**<br>**±0.0044** | 0.8597<br>±0.0030 | 47.1347<br>±1.0667 | 0.8353<br>±0.0031 | 0.8285<br>±0.0033 | **49.0393**<br>**±1.1678** | 0.7986<br>±0.0033 | 0.8082<br>±0.0035 | 70.7543<br>±1.4692 | 0.7802<br>±0.0035 |
| VRWKV [23] | 0.7912<br>±0.0036 | 143.4809<br>±2.1293 | 0.7662<br>±0.0037 | 0.5112<br>±0.0046 | 271.6789<br>±2.2767 | 0.4883<br>±0.0047 | 0.8056<br>±0.0036 | 62.1875<br>±1.0115 | 0.7850<br>±0.0037 | 0.7615<br>±0.0037 | 65.9837<br>±1.3855 | 0.7266<br>±0.0038 | 0.7480<br>±0.0039 | 89.0189<br>±1.6278 | 0.7188<br>±0.0039 |
| VM-UNET-V2 [40] | 0.6521<br>±0.0043 | 181.1196<br>±1.9559 | 0.6233<br>±0.0044 | 0.4833<br>±0.0046 | 332.8376<br>±2.5972 | 0.4575<br>±0.0046 | 0.7880<br>±0.0036 | 104.5297<br>±1.5290 | 0.7620<br>±0.0038 | 0.5410<br>±0.0043 | 232.7710<br>±2.4792 | 0.4963<br>±0.0043 | 0.6028<br>±0.0044 | 216.0202<br>±2.3843 | 0.5664<br>±0.0044 |
| UNETR [9] | 0.6199<br>±0.0046 | 250.5465<br>±2.1782 | 0.6050<br>±0.0047 | 0.4500<br>±0.0048 | 391.7400<br>±2.8152 | 0.4416<br>±0.0049 | 0.6436<br>±0.0047 | 158.4295<br>±1.8608 | 0.6349<br>±0.0047 | 0.4899<br>±0.0046 | 333.3864<br>±2.6834 | 0.4416<br>±0.0049 | 0.5369<br>±0.0047 | 308.9828<br>±2.6276 | 0.5193<br>±0.0047 |
| Swin UNETR [45] | 0.5171<br>±0.0047 | 207.6704<br>±2.4884 | 0.4921<br>±0.0046 | 0.3438<br>±0.0045 | 432.7990<br>±2.8733 | 0.3300<br>±0.0045 | 0.7037<br>±0.0042 | 85.6554<br>±1.7000 | 0.6805<br>±0.0043 | 0.6599<br>±0.0041 | 131.7022<br>±1.9244 | 0.6201<br>±0.0042 | 0.6137<br>±0.0044 | 156.2208<br>±2.2260 | 0.5824<br>±0.0044 |
| DA-TransUNet [46] | 0.7783<br>±0.0038 | 97.7011<br>±1.4945 | 0.7563<br>±0.0039 | 0.5907<br>±0.0045 | **166.4438**<br>**±2.1574** | 0.5610<br>±0.0045 | 0.8319<br>±0.0034 | 63.6652<br>±1.4344 | 0.8143<br>±0.0035 | 0.7948<br>±0.0035 | 81.5234<br>±1.6102 | 0.7594<br>±0.0035 | 0.7780<br>±0.0037 | 87.6505<br>±1.6488 | 0.7489<br>±0.0038 |
| H2Former [10] | 0.7490<br>±0.0039 | 148.2938<br>±2.0519 | 0.7220<br>±0.0040 | 0.5627<br>±0.0045 | 219.4662<br>±2.3340 | 0.5323<br>±0.0045 | 0.8095<br>±0.0035 | 55.2702<br>±1.0098 | 0.7859<br>±0.0036 | 0.6369<br>±0.0042 | 169.6482<br>±2.0895 | 0.5934<br>±0.0042 | 0.6815<br>±0.0041 | 154.2361<br>±2.0343 | 0.6460<br>±0.0042 |
| U-mamba [25] | 0.6549<br>±0.0043 | 132.5447<br>±1.7257 | 0.6214<br>±0.0043 | 0.3967<br>±0.0045 | 315.6740<br>±1.6164 | 0.3729<br>±0.0045 | 0.7500<br>±0.0040 | 75.1434<br>±1.5002 | 0.7262<br>±0.0040 | 0.5380<br>±0.0043 | 192.8970<br>±2.3236 | 0.4878<br>±0.0041 | 0.5843<br>±0.0044 | 181.0034<br>±2.1725 | 0.5449<br>±0.0043 |
| DualSeg(ours) | **0.8251**<br>**±0.0034** | **81.6201**<br>**±1.5314** | **0.8022**<br>**±0.0035** | 0.5888<br>±0.0045 | 235.8327<br>±2.6482 | 0.5573<br>±0.0045 | **0.8848**<br>**±0.0028** | **21.0608**<br>**±0.2919** | **0.8630**<br>**±0.0029** | **0.8393**<br>**±0.0032** | 57.6049<br>±1.3786 | **0.8123**<br>**±0.0033** | **0.8195**<br>**±0.0034** | **69.6369**<br>**±1.5420** | **0.7938**<br>**±0.0035** |

## Page 9, Figure 8:

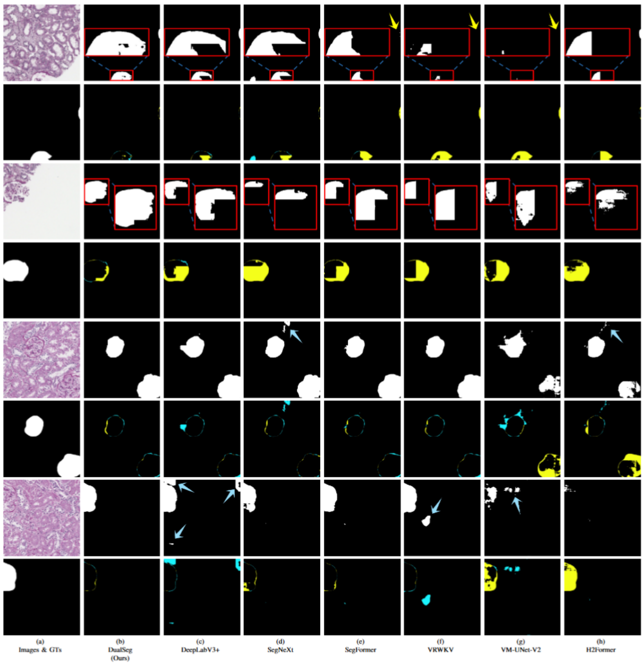

Fig. 8: Visual comparison on the held-out KPMP test set. **Odd rows** display original images and inference masks; **even rows** show GT and error maps, where **yellow** and **green** denote under-segmentation and over-segmentation, respectively. Red boxes and matching arrows highlight magnified details and segmentation errors. **DualSeg** exhibits unprecedented stability when handling cross-center and cross-species data.

3. **Incomplete Architectural Comparisons**

The baseline selection reveals a critical methodological flaw through systematic omission of contemporary efficient architectures. InceptionNeXt, which achieves transformer-comparable performance with CNN efficiency, directly challenges DualSeg's value proposition yet remains unexamined. Similarly, Mamba-based segmentation models (VM-UNet, U-Mamba) already address the linear complexity challenge DualSeg claims to solve. These omissions appear deliberate rather than oversight, suggesting the authors recognize these comparisons might undermine their architectural superiority claims. Without these essential benchmarks, the true contribution remains indeterminate.

**Response:** Thank you for pointing that out. We have addressed the perceived "methodological gap" by conducting a comprehensive re-evaluation. In the revised manuscript, we have integrated contemporary efficient architectures, including InceptionNeXt, VM-UNet, and U-Mamba, as primary baselines. These models are now formally described in **Section IV. B. Baselines**. The quantitative and qualitative results of these comparisons are presented in **Section V. A. Glomeruli Segmentation Results and Section V. C. Visualization Results**, detailed in **Tables I, II, and III**, and visualized in **Figure 8**. A detailed analysis and comparative discussion regarding our model's superiority over these state-of-the-art baselines are provided in **Section VI. A. Glomeruli Segmentation Results**. Our findings consistently demonstrate that DualSeg achieves superior segmentation accuracy while maintaining a more favorable performance-to-complexity ratio.

The revisions can be found in Section IV. B. Baselines, Section VI. A. Discussion, Section V. A. Results, Tables I-III and Figure 9:

**Page 7, Section IV. B. Baselines:**

We benchmark DualSeg against 14 representative methods spanning three architectural paradigms. CNN-based models include canonical baselines like U-Net [7] and Attention U-Net [41], the receptive-field-enhanced DeepLabV3+ [34], and efficient modern architectures such as SegNext [26], InceptionNext [42], and Wave-MLP [22]. Transformer-based models encompass SegFormer [11] for hierarchical encoding, VRWKV [23] utilizing linear recurrent operators, and the SSM-integrated VM-UNet-V2 [40]. Finally, Hybrid models feature U-shaped Transformer variants like UNETR [9] and Swin UNETR [43], alongside advanced fusion frameworks including H2Former [10], DA-TransUNet [44], and the Mamba-based U-Mamba [25].

**Page 9, Section VI. A. Glomeruli Segmentation Results:**

3) KPMP Dataset (Cross-Species Inference): To assess generalizability, we evaluated models pre-trained solely on the murine KPIs dataset directly on the human KPMP dataset without additional fine-tuning (Table III). DualSeg achieved a highly competitive average mDSC of 81.95%, surpassing SegFormer (80.82%) and significantly outperforming InceptionNext (52.79%). In terms of boundary delineation, DualSeg achieved an HD95 of 69.64, which is approximately one third of the error recorded by VM-UNet-V2 (216.02units). Furthermore, its IoU of 79.38% surpassed other hybrid models by over 4%, validating the model's capacity for robust cross-species and cross-center transfer learning.

**Page 11, Section VI. A. Comparative Analysis with SOTA Methods:**

DualSeg effectively addresses the limitations of existing architectural paradigms through a unified framework. Unlike conventional CNNs (e.g., DeepLabV3+, InceptionNext) which are constrained by fixed receptive fields, DualSeg utilizes the Wave-Swin Block's dynamic propagation window (Table V). This innovation enhances local texture discriminability, reducing HD95 by 26–37 on the KPIs dataset compared to CNN baselines and outperforming InceptionNext by 29.16% mDSC in cross-species tasks.

Furthermore, while Transformers typically excel at global context but compromise local detail due to patch flattening, DualSeg integrates a Z-Shift operator within the VRWKV block to preserve edge integrity. This design mitigates the information loss inherent in standard Q-Shift operations (Table VI) and models spatial heterogeneity more effectively, allowing DualSeg to exceed VM-UNet-V2 by 21.67% on the external KPMP dataset.

Finally, in contrast to existing hybrids (e.g., H2Former, DA TransUNet) that suffer from a "semantic gap" or the high computational costs of U-Mamba (Fig. 1), DualSeg employs a sequential *local-to-global* refinement strategy. This structured integration ensures precise morphological prior mapping, enabling the model to outperform H2Former by 13.68% on HuBMAP and surpass U-Mamba by 23.52% on KPMP. By dynamically adapting to morphological variability and seamlessly integrating features, DualSeg establishes a robust and efficient backbone for renal histology analysis.

## Page 4, Table I:

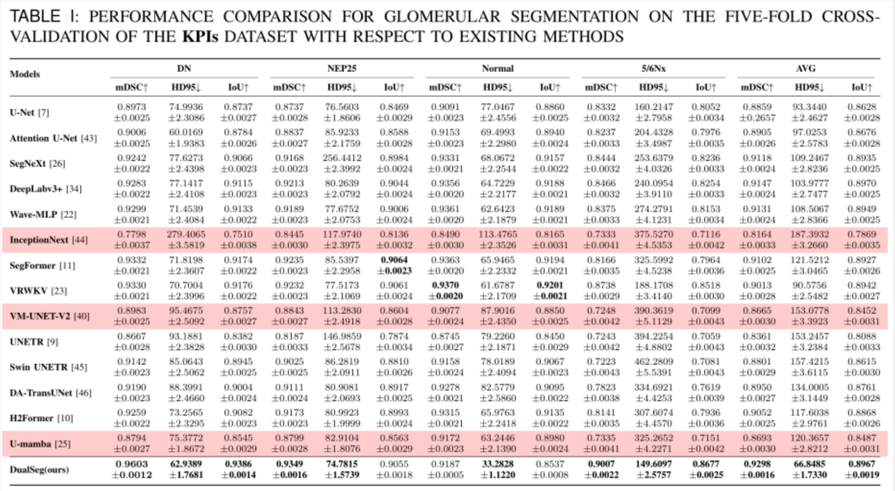

TABLE I: PERFORMANCE COMPARISON FOR GLOMERULAR SEGMENTATION ON THE FIVE-FOLD CROSS-VALIDATION OF THE **KPIs** DATASET WITH RESPECT TO EXISTING METHODS

| **Models** | **DN** | | | **NEP25** | | | **Normal** | | | **5/6Nx** | | | **AVG** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **mDSC↑** | **HD95↓** | **IoU↑** | **mDSC↑** | **HD95↓** | **IoU↑** | **mDSC↑** | **HD95↓** | **IoU↑** | **mDSC↑** | **HD95↓** | **IoU↑** | **mDSC↑** | **HD95↓** | **IoU↑** |
| **U-Net** [7] | 0.8973<br>±0.0025 | 74.9936<br>±2.3086 | 0.8737<br>±0.0027 | 0.8737<br>±0.0028 | 76.5603<br>±1.8606 | 0.8469<br>±0.0029 | 0.9091<br>±0.0023 | 77.0467<br>±2.4556 | 0.8860<br>±0.0025 | 0.8332<br>±0.0032 | 160.2147<br>±2.7958 | 0.8052<br>±0.0034 | 0.8859<br>±0.2657 | 93.3440<br>±2.4627 | 0.8628<br>±0.0028 |
| **Attention U-Net** [43] | 0.9006<br>±0.0025 | 60.0169<br>±1.9383 | 0.8784<br>±0.0026 | 0.8837<br>±0.0027 | 85.9233<br>±2.1759 | 0.8588<br>±0.0028 | 0.9153<br>±0.0023 | 69.4993<br>±2.2980 | 0.8940<br>±0.0024 | 0.8237<br>±0.0033 | 204.4328<br>±3.4987 | 0.7976<br>±0.0035 | 0.8905<br>±0.0026 | 97.0253<br>±2.5783 | 0.8676<br>±0.0028 |
| **SegNeXt** [26] | 0.9242<br>±0.0022 | 77.6273<br>±2.4398 | 0.9066<br>±0.0023 | 0.9168<br>±0.0023 | 256.4412<br>±2.3992 | 0.8984<br>±0.0024 | 0.9331<br>±0.0021 | 68.0672<br>±2.2544 | 0.9157<br>±0.0022 | 0.8444<br>±0.0032 | 253.6379<br>±4.0326 | 0.8236<br>±0.0033 | 0.9118<br>±0.0024 | 109.2467<br>±2.8236 | 0.8935<br>±0.0025 |
| **DeepLabv3+** [34] | 0.9283<br>±0.0022 | 77.1417<br>±2.4108 | 0.9115<br>±0.0023 | 0.9213<br>±0.0023 | 80.2639<br>±2.0792 | 0.9044<br>±0.0024 | 0.9356<br>±0.0020 | 64.7229<br>±2.2177 | 0.9188<br>±0.0021 | 0.8466<br>±0.0032 | 240.0954<br>±3.9110 | 0.8254<br>±0.0033 | 0.9147<br>±0.0024 | 103.9777<br>±2.7477 | 0.8970<br>±0.0025 |
| **Wave-MLP** [22] | 0.9299<br>±0.0021 | 71.4539<br>±2.4084 | 0.9133<br>±0.0022 | 0.9189<br>±0.0023 | 77.6752<br>±2.0753 | 0.9006<br>±0.0024 | 0.9361<br>±0.0020 | 62.6423<br>±2.1879 | 0.9189<br>±0.0021 | 0.8375<br>±0.0033 | 274.2791<br>±4.1231 | 0.8153<br>±0.0034 | 0.9131<br>±0.0024 | 108.5067<br>±2.8366 | 0.8949<br>±0.0025 |
| **InceptionNext** [44] | 0.7798<br>±0.0037 | 279.4065<br>±3.5819 | 0.7510<br>±0.0038 | 0.8445<br>±0.0030 | 117.9740<br>±2.3975 | 0.8136<br>±0.0032 | 0.8490<br>±0.0030 | 113.4765<br>±2.3526 | 0.8165<br>±0.0031 | 0.7333<br>±0.0041 | 375.5270<br>±4.5353 | 0.7116<br>±0.0042 | 0.8164<br>±0.0033 | 187.3932<br>±3.2660 | 0.7869<br>±0.0035 |
| **SegFormer** [11] | 0.9332<br>±0.0021 | 71.8198<br>±2.3607 | 0.9174<br>±0.0022 | 0.9235<br>±0.0023 | 85.5397<br>±2.2958 | **0.9064<br>±0.0023** | 0.9363<br>±0.0020 | 65.9465<br>±2.2332 | 0.9194<br>±0.0021 | 0.8166<br>±0.0035 | 325.5992<br>±4.5238 | 0.7964<br>±0.0036 | 0.9102<br>±0.0025 | 121.5212<br>±3.0465 | 0.8927<br>±0.0026 |
| **VRWKV** [23] | 0.9330<br>±0.0021 | 70.7004<br>±2.3996 | 0.9176<br>±0.0022 | 0.9232<br>±0.0023 | 77.5173<br>±2.1069 | 0.9061<br>±0.0024 | **0.9370<br>±0.0020** | 61.6787<br>±2.1709 | **0.9201<br>±0.0021** | 0.8738<br>±0.0029 | 188.1708<br>±3.4140 | 0.8518<br>±0.0030 | 0.9013<br>±0.0028 | 90.5756<br>±2.5482 | 0.8942<br>±0.0027 |
| **VM-UNET-V2** [40] | 0.8983<br>±0.0025 | 95.4675<br>±2.5092 | 0.8757<br>±0.0027 | 0.8843<br>±0.0027 | 113.2830<br>±2.4918 | 0.8604<br>±0.0028 | 0.9077<br>±0.0024 | 87.9016<br>±2.4350 | 0.8850<br>±0.0025 | 0.7248<br>±0.0042 | 390.3619<br>±5.1129 | 0.7099<br>±0.0043 | 0.8665<br>±0.0030 | 153.0778<br>±3.3923 | 0.8452<br>±0.0031 |
| **UNETR** [9] | 0.8667<br>±0.0028 | 93.1881<br>±2.3828 | 0.8382<br>±0.0030 | 0.8187<br>±0.0033 | 146.9859<br>±2.5678 | 0.7874<br>±0.0034 | 0.8745<br>±0.0027 | 79.2260<br>±2.1871 | 0.8450<br>±0.0029 | 0.7243<br>±0.0042 | 394.2254<br>±4.8802 | 0.7059<br>±0.0043 | 0.8361<br>±0.0032 | 153.2457<br>±3.2384 | 0.8088<br>±0.0033 |
| **Swin UNETR** [45] | 0.9142<br>±0.0023 | 85.0643<br>±2.5062 | 0.8945<br>±0.0025 | 0.9025<br>±0.0025 | 86.2819<br>±2.0911 | 0.8810<br>±0.0026 | 0.9158<br>±0.0024 | 78.0189<br>±2.4094 | 0.9067<br>±0.0023 | 0.7223<br>±0.0043 | 462.2809<br>±5.5391 | 0.7081<br>±0.0043 | 0.8801<br>±0.0029 | 157.4215<br>±3.6115 | 0.8615<br>±0.0030 |
| **DA-TransUNet** [46] | 0.9190<br>±0.0023 | 88.3991<br>±2.4660 | 0.9004<br>±0.0024 | 0.9111<br>±0.0024 | 80.9081<br>±2.0693 | 0.8917<br>±0.0025 | 0.9278<br>±0.0021 | 82.5779<br>±2.5860 | 0.9095<br>±0.0022 | 0.7823<br>±0.0038 | 334.6921<br>±4.4253 | 0.7619<br>±0.0039 | 0.8950<br>±0.0027 | 134.0005<br>±3.1449 | 0.8761<br>±0.0028 |
| **H2Former** [10] | 0.9259<br>±0.0022 | 73.2565<br>±2.3295 | 0.9082<br>±0.0023 | 0.9173<br>±0.0023 | 80.9923<br>±1.9999 | 0.8993<br>±0.0024 | 0.9315<br>±0.0021 | 65.9763<br>±2.2418 | 0.9135<br>±0.0022 | 0.8141<br>±0.0035 | 307.6074<br>±4.4570 | 0.7936<br>±0.0036 | 0.9052<br>±0.0025 | 117.6038<br>±2.9761 | 0.8868<br>±0.0026 |
| **U-mamba** [25] | 0.8794<br>±0.0027 | 75.3772<br>±1.8672 | 0.8545<br>±0.0029 | 0.8799<br>±0.0028 | 82.9104<br>±1.8076 | 0.8563<br>±0.0029 | 0.9172<br>±0.0023 | 63.2446<br>±2.1390 | 0.8980<br>±0.0024 | 0.7335<br>±0.0041 | 325.2652<br>±4.2271 | 0.7151<br>±0.0042 | 0.8693<br>±0.0030 | 120.3657<br>±2.8212 | 0.8487<br>±0.0031 |
| **DualSeg(ours)** | **0.9603<br>±0.0012** | **62.9389<br>±1.7681** | **0.9386<br>±0.0014** | **0.9349<br>±0.0016** | **74.7815<br>±1.5739** | 0.9055<br>±0.0018 | 0.9187<br>±0.0005 | **33.2828<br>±1.1220** | 0.8537<br>±0.0008 | **0.9007<br>±0.0022** | **149.6097<br>±2.5757** | **0.8677<br>±0.0025** | **0.9298<br>±0.0016** | **66.8485<br>±1.7330** | **0.8967<br>±0.0019** |

## Page 5, Table II:

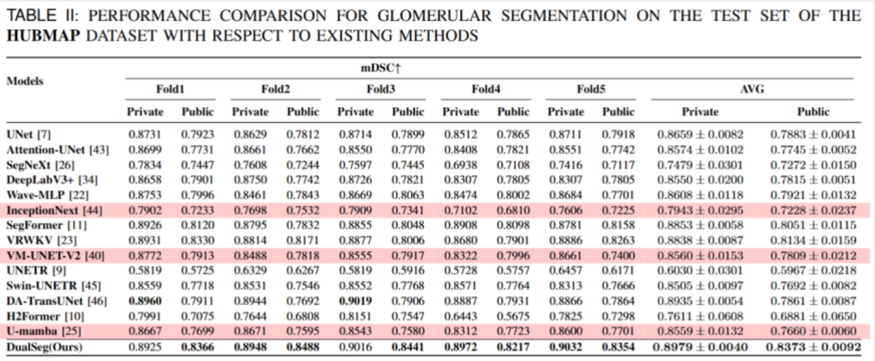

TABLE II: PERFORMANCE COMPARISON FOR GLOMERULAR SEGMENTATION ON THE TEST SET OF THE **HUBMAP** DATASET WITH RESPECT TO EXISTING METHODS

| **Models** | **mDSC↑** | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Fold1** | | **Fold2** | | **Fold3** | | **Fold4** | | **Fold5** | | **AVG** | |
| | **Private** | **Public** | **Private** | **Public** | **Private** | **Public** | **Private** | **Public** | **Private** | **Public** | **Private** | **Public** |
| **UNet** [7] | 0.8731 | 0.7923 | 0.8629 | 0.7812 | 0.8714 | 0.7899 | 0.8512 | 0.7865 | 0.8711 | 0.7918 | 0.8659 ± 0.0082 | 0.7883 ± 0.0041 |
| **Attention-UNet** [43] | 0.8699 | 0.7731 | 0.8661 | 0.7662 | 0.8550 | 0.7770 | 0.8408 | 0.7821 | 0.8551 | 0.7742 | 0.8574 ± 0.0102 | 0.7745 ± 0.0052 |
| **SegNeXt** [26] | 0.7834 | 0.7447 | 0.7608 | 0.7244 | 0.7597 | 0.7445 | 0.6938 | 0.7108 | 0.7416 | 0.7117 | 0.7479 ± 0.0301 | 0.7272 ± 0.0150 |
| **DeepLabV3+** [34] | 0.8658 | 0.7901 | 0.8750 | 0.7742 | 0.8726 | 0.7821 | 0.8307 | 0.7805 | 0.8307 | 0.7805 | 0.8550 ± 0.0200 | 0.7815 ± 0.0051 |
| **Wave-MLP** [22] | 0.8753 | 0.7996 | 0.8461 | 0.7843 | 0.8669 | 0.8063 | 0.8474 | 0.8002 | 0.8684 | 0.7701 | 0.8608 ± 0.0118 | 0.7921 ± 0.0132 |
| **InceptionNext** [44] | 0.7902 | 0.7233 | 0.7698 | 0.7532 | 0.7909 | 0.7341 | 0.7102 | 0.6810 | 0.7606 | 0.7225 | 0.7943 ± 0.0295 | 0.7228 ± 0.0237 |
| **SegFormer** [11] | 0.8926 | 0.8120 | 0.8795 | 0.7832 | 0.8855 | 0.8048 | 0.8908 | 0.8098 | 0.8781 | 0.8158 | 0.8853 ± 0.0058 | 0.8051 ± 0.0115 |
| **VRWKV** [23] | 0.8931 | 0.8330 | 0.8814 | 0.8171 | 0.8877 | 0.8006 | 0.8680 | 0.7901 | 0.8886 | 0.8263 | 0.8838 ± 0.0087 | 0.8134 ± 0.0159 |
| **VM-UNET-V2** [40] | 0.8772 | 0.7913 | 0.8488 | 0.7818 | 0.8555 | 0.7917 | 0.8322 | 0.7996 | 0.8661 | 0.7400 | 0.8560 ± 0.0153 | 0.7809 ± 0.0212 |
| **UNETR** [9] | 0.5819 | 0.5725 | 0.6329 | 0.6267 | 0.5819 | 0.5916 | 0.5728 | 0.5757 | 0.6457 | 0.6171 | 0.6030 ± 0.0301 | 0.5967 ± 0.0218 |
| **Swin-UNETR** [45] | 0.8559 | 0.7718 | 0.8531 | 0.7546 | 0.8552 | 0.7768 | 0.8571 | 0.7764 | 0.8313 | 0.7666 | 0.8505 ± 0.0097 | 0.7692 ± 0.0082 |
| **DA-TransUNet** [46] | **0.8960** | 0.7911 | 0.8944 | 0.7692 | **0.9019** | 0.7906 | 0.8887 | 0.7931 | 0.8866 | 0.7864 | 0.8935 ± 0.0054 | 0.7861 ± 0.0087 |
| **H2Former** [10] | 0.7991 | 0.7075 | 0.7644 | 0.6808 | 0.8151 | 0.7547 | 0.6443 | 0.5675 | 0.7825 | 0.7298 | 0.7611 ± 0.0608 | 0.6881 ± 0.0650 |
| **U-mamba** [25] | 0.8667 | 0.7699 | 0.8671 | 0.7595 | 0.8543 | 0.7580 | 0.8312 | 0.7723 | 0.8600 | 0.7701 | 0.8559 ± 0.0132 | 0.7660 ± 0.0060 |
| **DualSeg(Ours)** | 0.8925 | **0.8366** | **0.8948** | **0.8488** | 0.9016 | **0.8441** | **0.8972** | **0.8217** | **0.9032** | **0.8354** | **0.8979 ± 0.0040** | **0.8373 ± 0.0092** |

## Page 6, Table III:

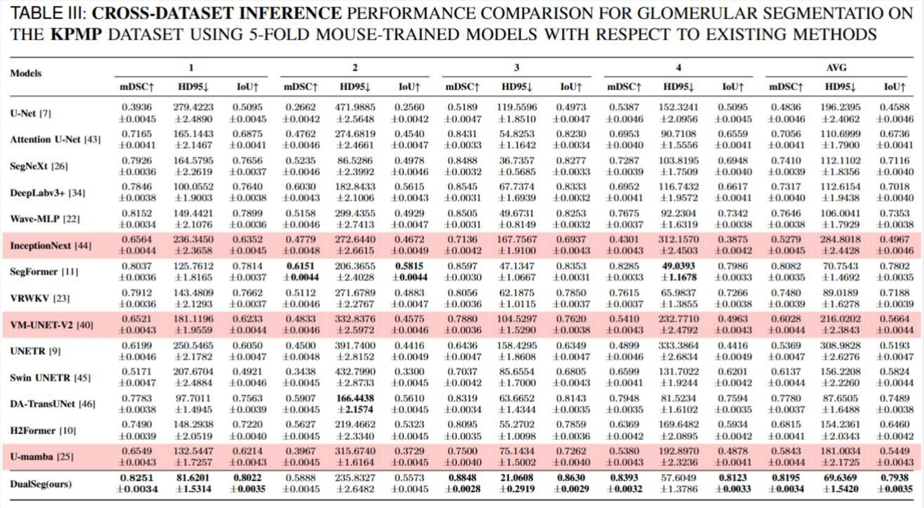

TABLE III: **CROSS-DATASET INFERENCE** PERFORMANCE COMPARISON FOR GLOMERULAR SEGMENTATIO ON THE **KPMP** DATASET USING 5-FOLD MOUSE-TRAINED MODELS WITH RESPECT TO EXISTING METHODS

| **Models** | **1** | | | **2** | | | **3** | | | **4** | | | **AVG** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **mDSC↑** | **HD95↓** | **IoU↑** | **mDSC↑** | **HD95↓** | **IoU↑** | **mDSC↑** | **HD95↓** | **IoU↑** | **mDSC↑** | **HD95↓** | **IoU↑** | **mDSC↑** | **HD95↓** | **IoU↑** |
| **U-Net** [7] | 0.3936<br>±0.0045 | 279.4223<br>±2.4890 | 0.5095<br>±0.0045 | 0.2662<br>±0.0042 | 471.9885<br>±2.5648 | 0.2560<br>±0.0042 | 0.5189<br>±0.0047 | 119.5596<br>±1.8510 | 0.4973<br>±0.0047 | 0.5387<br>±0.0046 | 152.3241<br>±2.0956 | 0.5095<br>±0.0045 | 0.4836<br>±0.0046 | 196.2395<br>±2.4062 | 0.4588<br>±0.0046 |
| **Attention U-Net** [43] | 0.7165<br>±0.0041 | 165.1443<br>±2.1467 | 0.6875<br>±0.0041 | 0.4762<br>±0.0046 | 274.6819<br>±2.4661 | 0.4540<br>±0.0047 | 0.8431<br>±0.0033 | 54.8253<br>±1.1642 | 0.8230<br>±0.0034 | 0.6953<br>±0.0040 | 90.7108<br>±1.5556 | 0.6559<br>±0.0041 | 0.7056<br>±0.0041 | 110.6999<br>±1.7900 | 0.6736<br>±0.0041 |
| **SegNeXt** [26] | 0.7926<br>±0.0036 | 164.5795<br>±2.2619 | 0.7656<br>±0.0037 | 0.5235<br>±0.0046 | 86.5286<br>±2.3992 | 0.4978<br>±0.0046 | 0.8488<br>±0.0032 | 36.7357<br>±0.5685 | 0.8277<br>±0.0033 | 0.7287<br>±0.0039 | 103.8195<br>±1.7509 | 0.6948<br>±0.0040 | 0.7410<br>±0.0039 | 112.1102<br>±1.8356 | 0.7116<br>±0.0040 |
| **DeepLabv3+** [34] | 0.7846<br>±0.0038 | 100.0552<br>±1.9003 | 0.7640<br>±0.0038 | 0.6030<br>±0.0043 | 182.8433<br>±2.1006 | 0.5615<br>±0.0043 | 0.8545<br>±0.0031 | 67.7374<br>±1.6939 | 0.8333<br>±0.0032 | 0.6952<br>±0.0041 | 116.7432<br>±1.9572 | 0.6617<br>±0.0041 | 0.7317<br>±0.0040 | 112.6154<br>±1.9438 | 0.7018<br>±0.0040 |
| **Wave-MLP** [22] | 0.8152<br>±0.0034 | 149.4421<br>±2.1076 | 0.7899<br>±0.0036 | 0.5158<br>±0.0046 | 299.4355<br>±2.7413 | 0.4929<br>±0.0047 | 0.8505<br>±0.0031 | 49.6731<br>±0.8149 | 0.8253<br>±0.0032 | 0.7675<br>±0.0037 | 92.2304<br>±1.6319 | 0.7342<br>±0.0038 | 0.7646<br>±0.0038 | 106.0041<br>±1.7929 | 0.7353<br>±0.0038 |
| **InceptionNext** [44] | 0.6564<br>±0.0044 | 236.3450<br>±2.3658 | 0.6352<br>±0.0045 | 0.4779<br>±0.0048 | 272.6440<br>±2.6615 | 0.4672<br>±0.0049 | 0.7136<br>±0.0042 | 167.7567<br>±1.9100 | 0.6937<br>±0.0043 | 0.4301<br>±0.0043 | 312.1570<br>±2.4503 | 0.3875<br>±0.0042 | 0.5279<br>±0.0045 | 284.8018<br>±2.4428 | 0.4967<br>±0.0046 |
| **SegFormer** [11] | 0.8037<br>±0.0036 | 125.7612<br>±1.8165 | 0.7814<br>±0.0037 | **0.6151<br>±0.0044** | 206.3655<br>±2.4028 | **0.5815<br>±0.0044** | 0.8597<br>±0.0030 | 47.1347<br>±1.0667 | 0.8353<br>±0.0031 | 0.8285<br>±0.0033 | **49.0393<br>±1.1678** | 0.7986<br>±0.0033 | 0.8082<br>±0.0035 | 70.7543<br>±1.4692 | 0.7802<br>±0.0035 |
| **VRWKV** [23] | 0.7912<br>±0.0036 | 143.4809<br>±2.1293 | 0.7662<br>±0.0037 | 0.5112<br>±0.0046 | 271.6789<br>±2.2767 | 0.4883<br>±0.0047 | 0.8056<br>±0.0036 | 62.1875<br>±1.0115 | 0.7850<br>±0.0037 | 0.7615<br>±0.0037 | 65.9837<br>±1.3855 | 0.7266<br>±0.0038 | 0.7480<br>±0.0039 | 89.0189<br>±1.6278 | 0.7188<br>±0.0039 |
| **VM-UNET-V2** [40] | 0.6521<br>±0.0043 | 181.1196<br>±1.9559 | 0.6233<br>±0.0044 | 0.4833<br>±0.0046 | 332.8376<br>±2.5972 | 0.4575<br>±0.0046 | 0.7880<br>±0.0036 | 104.5297<br>±1.5290 | 0.7620<br>±0.0038 | 0.5410<br>±0.0043 | 232.7710<br>±2.4792 | 0.4963<br>±0.0043 | 0.6028<br>±0.0044 | 216.0202<br>±2.3843 | 0.5664<br>±0.0044 |
| **UNETR** [9] | 0.6199<br>±0.0046 | 250.5465<br>±2.1782 | 0.6050<br>±0.0047 | 0.4500<br>±0.0048 | 391.7400<br>±2.8152 | 0.4416<br>±0.0049 | 0.6436<br>±0.0047 | 158.4295<br>±1.8608 | 0.6349<br>±0.0047 | 0.4899<br>±0.0046 | 333.3864<br>±2.6834 | 0.4416<br>±0.0049 | 0.5369<br>±0.0047 | 308.9828<br>±2.6276 | 0.5193<br>±0.0047 |
| **Swin UNETR** [45] | 0.5171<br>±0.0047 | 207.6704<br>±2.4884 | 0.4921<br>±0.0046 | 0.3438<br>±0.0045 | 432.7990<br>±2.8733 | 0.3300<br>±0.0045 | 0.7037<br>±0.0042 | 85.6554<br>±1.7000 | 0.6805<br>±0.0043 | 0.6599<br>±0.0041 | 131.7022<br>±1.9244 | 0.6201<br>±0.0042 | 0.6137<br>±0.0044 | 156.2208<br>±2.2260 | 0.5824<br>±0.0044 |
| **DA-TransUNet** [46] | 0.7783<br>±0.0038 | 97.7011<br>±1.4945 | 0.7563<br>±0.0039 | 0.5907<br>±0.0045 | **166.4438<br>±2.1574** | 0.5610<br>±0.0045 | 0.8319<br>±0.0034 | 63.6652<br>±1.4344 | 0.8143<br>±0.0035 | 0.7948<br>±0.0035 | 81.5234<br>±1.6102 | 0.7594<br>±0.0035 | 0.7780<br>±0.0037 | 87.6505<br>±1.6488 | 0.7489<br>±0.0038 |
| **H2Former** [10] | 0.7490<br>±0.0039 | 148.2938<br>±2.0519 | 0.7220<br>±0.0040 | 0.5627<br>±0.0045 | 219.4662<br>±2.3340 | 0.5323<br>±0.0045 | 0.8095<br>±0.0035 | 55.2702<br>±1.0098 | 0.7859<br>±0.0036 | 0.6369<br>±0.0042 | 169.6482<br>±2.0895 | 0.5934<br>±0.0042 | 0.6815<br>±0.0041 | 154.2361<br>±2.0343 | 0.6460<br>±0.0042 |
| **U-mamba** [25] | 0.6549<br>±0.0043 | 132.5447<br>±1.7257 | 0.6214<br>±0.0043 | 0.3967<br>±0.0045 | 315.6740<br>±1.6164 | 0.3729<br>±0.0045 | 0.7500<br>±0.0040 | 75.1434<br>±1.5002 | 0.7262<br>±0.0040 | 0.5380<br>±0.0043 | 192.8970<br>±2.3236 | 0.4878<br>±0.0041 | 0.5843<br>±0.0044 | 181.0034<br>±2.1725 | 0.5449<br>±0.0043 |
| **DualSeg(ours)** | **0.8251<br>±0.0034** | **81.6201<br>±1.5314** | **0.8022<br>±0.0035** | 0.5888<br>±0.0045 | 235.8327<br>±2.6482 | 0.5573<br>±0.0045 | **0.8848<br>±0.0028** | **21.0608<br>±0.2919** | **0.8630<br>±0.0029** | **0.8393<br>±0.0032** | 57.6049<br>±1.3786 | **0.8123<br>±0.0033** | **0.8195<br>±0.0034** | **69.6369<br>±1.5420** | **0.7938<br>±0.0035** |

**Page 11, Figure 9:**

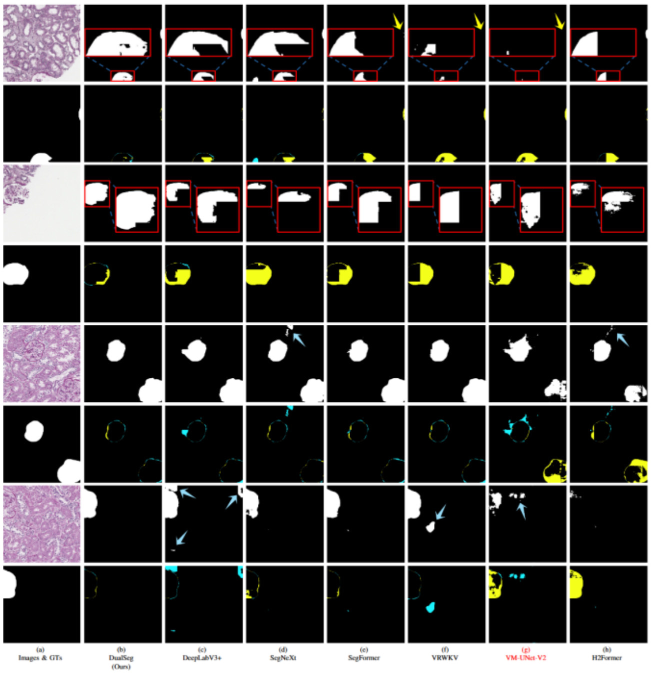

Fig. 9: Visual comparison on the held-out KPMP test set. **Odd rows** display original images and inference masks; **even rows** show GT and error maps, where **yellow** and **green** denote under-segmentation and over-segmentation, respectively. Red boxes and matching arrows highlight magnified details and segmentation errors. **DualSeg** exhibits unprecedented stability when handling cross-center and cross-species data.

### 4. Clinical Deployment Analysis

Despite positioning DualSeg for practical application, the manuscript provides no deployment feasibility analysis. Clinical environments impose strict constraints: limited GPU memory (often 8GB), CPU-only workstations in many facilities. The reported 2.73% mDSC improvement lacks clinical context—pathologists require understanding of whether such margins affect diagnostic confidence, inter-observer agreement, or treatment decisions. Without this translation from metrics to clinical impact, the practical value remains speculative.

**Response:** Thank you for pointing that out. We have revised the manuscript to demonstrate the statistical robustness and practical feasibility of our results, as detailed below:

- **Text Revision on Deployment Feasibility:** We have added a comprehensive discussion in **Section VI. D. Limitations and Future Work** regarding the practical deployment of DualSeg. This new section outlines the roadmap for future optimizations, including lightweight design to transition our architectural innovations into clinical tools.
- **Expanded Statistical Validation**: To substantiate the 2.73% mDSC improvement, we have enriched **Section VI. C. Clinical Relevance** with quantitative evidence. We performed *t*-tests across three diverse datasets (KPIs, HuBMAP, and KPMP); as illustrated in **Figure 10**, the results confirm that DualSeg's performance gains are statistically significant ($p < 0.001$), ensuring the improvements are robust rather than marginal.
- **Enhanced Qualitative Evaluation:** To supplement the quantitative metrics, we have incorporated a granular spatial audit in **Figures 6 and 8**. By utilizing red bounding boxes for local magnification and colored arrows (yellow for under-segmentation; green for over-segmentation), we explicitly demonstrate DualSeg's superior ability to maintain structural continuity and reduce critical diagnostic omissions in complex anatomical structures._

The revisions can be found in Section VI. D. Limitations and Future Work, Section VI. C. Clinical Relevance, Figures 6, 8 and 10:

## Page 12, Section VI. D. Limitations and Future Work:

While DualSeg demonstrates superior performance in glomerular segmentation, three primary limitations remain to be addressed in future iterations. First, the model's generalization to rare pathological subtypes, such as global glomerulosclerosis, is currently constrained by data scarcity. We plan to mitigate this by employing domain adaptation and transfer learning techniques to enhance feature robustness for these underrepresented classes. Second, our current validation is restricted to PAS-stained images. To ensure broad clinical applicability, we will extend our evaluation to include Hematoxylin-Eosin (HE) stained datasets, verifying the model's adaptability to varying histological protocols. Finally, despite the linear complexity of the VRWKV block, the computational overhead for gigapixel WSI processing remains significant. Future work will focus on model quantization and lightweight optimization to facilitate deployment on resource-constrained platforms and edge devices.

## Page 12, Section VI. C. Clinical Relevance:

DualSeg exhibits statistically significant superiority ($p<0.05$–$0.001$; Fig. 10 and exceptional reproducibility, evidenced by a minimal standard deviation (0.0040) on the HuBMAP dataset. Its ability to accurately resolve diverse morphologies—ranging from mild hypertrophy to severe fragmentation—enables the precise quantification of pathological biomarkers like sclerosis and fibrosis. Furthermore, the model's robust performance on the cross-species KPMP dataset supports standardized CKD monitoring. By mitigating inter-observer variability and reducing manual annotation burdens, DualSeg provides a scalable solution for multi-center clinical trials and routine diagnostic workflows.

## Page 7, Figure 6:

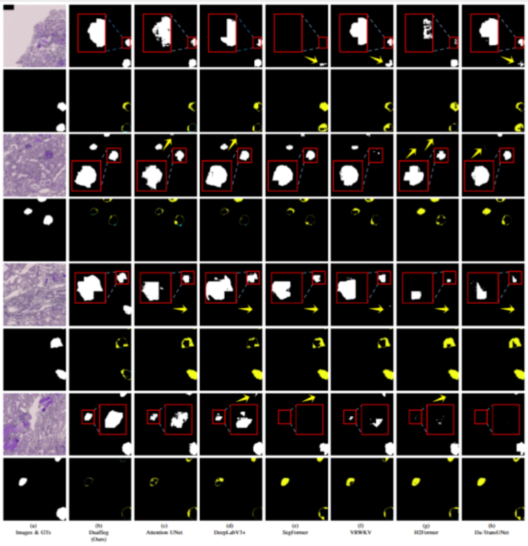

Fig. 6: Visual comparison on the KPIs test set. **Odd rows** display original images and inference masks; **even rows** show GT and error maps, where **yellow** and **green** denote under-segmentation and over-segmentation, respectively. Red boxes and matching arrows highlight magnified details and specific segmentation errors. **DualSeg** demonstrates superior performance in handling heterogeneous glomeruli with intricate boundaries.

## Page 11, Figure 8:

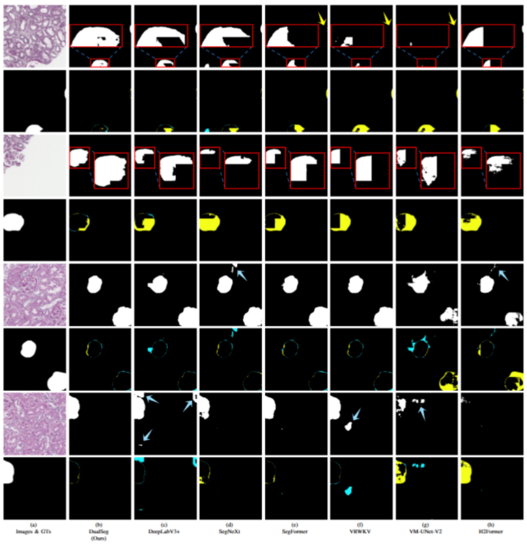

Fig. 8: Visual comparison on the held-out KPMP test set. **Odd rows** display original images and inference masks; **even rows** show GT and error maps, where **yellow** and **green** denote under-segmentation and over-segmentation, respectively. Red boxes and matching arrows highlight magnified details and segmentation errors. **DualSeg** exhibits unprecedented stability when handling cross-center and cross-species data.

**Page 12, Figure 10:**

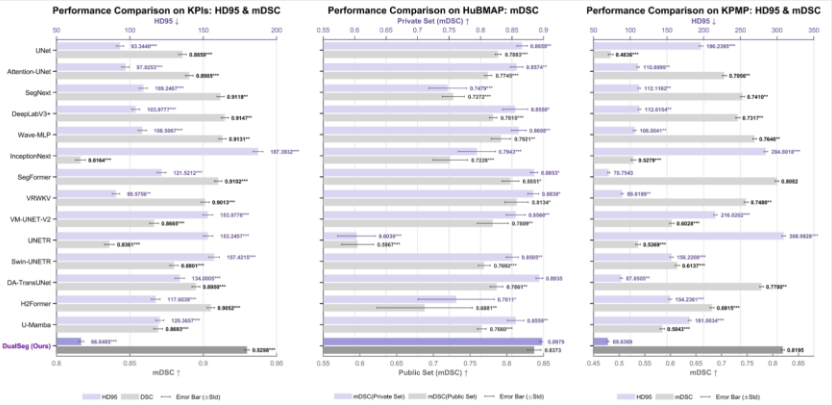

Fig. 10: Performance comparison between our DualSeg model and 14 baseline methods across three datasets. The comparison metrics include mDSC and HD95 for the KPIs and KPMP datasets (left and right panels, respectively), and mDSC for the HuBMAP dataset (middle panel). The error bars represent ± standard deviation. Statistical significance was assessed using paired $t$-tests, with levels indicated by asterisks: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

5. **Architectural Design Justification**

The dual-stage encoder's sequential arrangement (Wave-Swin → VRWKV) lacks theoretical foundation. The manuscript presents this as optimal without exploring alternative configurations (parallel processing, reversed ordering, or hybrid fusion strategies). The dynamic window sizing, acknowledged as "manually defined based on empirical observations," reveals methodological weakness—critical design parameters derived through trial-and-error rather than principled analysis. This empirical approach, while sometimes necessary, requires thorough justification absent here.

**Response:** Thank you for your valuable comments. We have addressed the concerns regarding the model's practical utility by adding a limitation analysis on deployability and a dedicated section on clinical significance, as detailed below:

- **Ablation of Architectural Ordering:** To substantiate the theoretical superiority of the proposed sequential arrangement (Wave-Swin→VRWKV), we have conducted additional ablation experiments evaluating alternative configurations, including reversed ordering (VRWKV→Wave-Swin) and Attention-Wave hybrids. The comparative results (detailed in the **Section V. B. Ablation Studies** and **Table IV**) demonstrate that our current design optimizes the transition from local spatial feature extraction to global linear complexity modeling, yielding the highest segmentation accuracy.
- **Principled Justification for Dynamic Window Sizing:** We clarify that the selection of the window candidate set S = {7, 11, 15, …} is not the result of unprincipled trial-and-error, but is rooted in domain-validated baselines. We have significantly expanded **Section III. A. 2. Dynamic Swin Mechanism** to detail this rationale.

The revisions can be found in Section III. A. Wave-Swin Block, Section V. B. Ablation Studies and Table IV:

**Page 4, Section III. A. Wave-Swin Block:**

First, previous studies on Wave-MLP have empirically demonstrated that windows smaller than 7 lack the generality necessary to capture spatial dependencies in medical images [22]; meanwhile, anchor sizes 7 and 11 align with kernel sizes employed in SOTA encoders like SegNeXt [26]. Second, the average glomerular bounding box in our murine dataset measures approximately 154px [33]. After the 4× and 8× downsampling stages, this dimension reduces to roughly 38px and 19px, respectively. Accordingly, selecting a maximum window of 15 (instead of 21) prevents the network from integrating extraneous background noise while ensuring full coverage of the target glomerular structure.

**Page 10, Section V. B. Ablation Studies:**

1) Effect of Dual-Stage Encoder: Table IV contrasts single stage architectures with the proposed dual-stage design. Single-stage variants employing only Wave-Swin Blocks, SegFormer-style self-attention, or VRWKV Blocks achieved average mDSCs of 90.36%, 90.99%, and 91.08%, respectively. The integrated Dual-Stage (Wave-VRWKV) architecture outperformed all single-stage counterparts with an average mDSC of 92.98%. We further investigated the impact of module sequencing. Reversing the feature extraction order (placing attention mechanisms before wave blocks) resulted in a significant performance decrease, lowering the average mDSC to 85.78%(Attention-Wave) and 91.63%(VRWKV-Wave). These findings corroborate the critical role of the proposed local-to-global refinement strategy.

**Page 10, Table IV:**

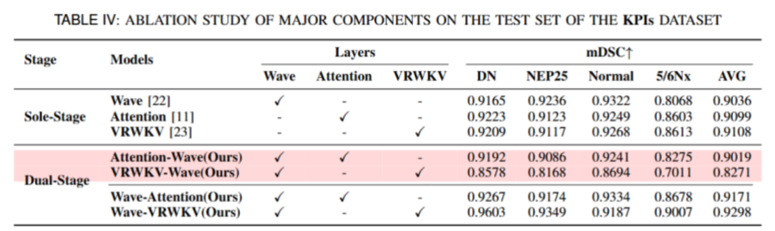

TABLE IV: ABLATION STUDY OF MAJOR COMPONENTS ON THE TEST SET OF THE **KPIs** DATASET

| Stage | Models | Layers | | | mDSC↑ | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Wave | Attention | VRWKV | DN | NEP25 | Normal | 5/6Nx | AVG |
| Sole-Stage | Wave [22] | ✓ | - | - | 0.9165 | 0.9236 | 0.9322 | 0.8068 | 0.9036 |
| | Attention [11] | - | ✓ | - | 0.9223 | 0.9123 | 0.9249 | 0.8603 | 0.9099 |
| | VRWKV [23] | - | - | ✓ | 0.9209 | 0.9117 | 0.9268 | 0.8613 | 0.9108 |
| Dual-Stage | Attention-Wave(Ours) | ✓ | ✓ | - | 0.9192 | 0.9086 | 0.9241 | 0.8275 | 0.9019 |
| | VRWKV-Wave(Ours) | ✓ | - | ✓ | 0.8578 | 0.8168 | 0.8694 | 0.7011 | 0.8271 |
| | Wave-Attention(Ours) | ✓ | ✓ | - | 0.9267 | 0.9174 | 0.9334 | 0.8678 | 0.9171 |
| | Wave-VRWKV(Ours) | ✓ | - | ✓ | 0.9603 | 0.9349 | 0.9187 | 0.9007 | 0.9298 |

**Minor Concerns**

**Ablation scope:** Component analysis limited to mDSC ignores computational overhead—does each module justify its complexity?

➢ **Response:** We have addressed the concern regarding computational overhead by incorporating a system-level efficiency analysis in **Figure 1 (Bottom)**, which demonstrates that DualSeg achieves an optimal trade-off between segmentation accuracy and FLOPs compared to 14 baseline methods. While our ablation studies (**Tables IV-VI**) prioritize the synergy of the dual-stage architecture, the selection of the lightweight HamDecoder using Non-negative Matrix Factorization (NMF) further justifies our design by enhancing multi-scale fusion without significant complexity. This strategic balance ensures that the proposed local-to-global refinement remains computationally viable for high-resolution histopathology.

The supporting data for these revisions can be found in Figure 1 (pertaining to Comment 1), Table IV (pertaining to Comment 5), and Tables V and VI, which are provided on the Page 10 of the revised manuscript (Page 8 of the original manuscript).

**Presentation imbalance:** Technical density overshadows clinical motivation, limiting accessibility.

➢ **Response:** Thank you for the constructive suggestions. To better balance technical density with clinical utility, we have enriched **Section VI. C. Clinical Relevance** with a statistical analysis of diagnostic impact. As illustrated in the newly added **Fig. 10**, DualSeg demonstrates statistically significant superiority ($p<0.05$ to $p<0.001$) across three diverse datasets, confirming that the mDSC improvements are robust and clinically meaningful. These performance gains directly translate to a reduction in diagnostic omissions and more precise quantification of biomarkers like glomerulosclerosis, thereby effectively reducing the manual annotation burden for pathologists.

The supporting data for these revisions can be found in Section VI. C. Clinical Relevance and Figure 10 (pertaining to Comment 4).

**Visualization gaps:** Comparison figures lack error maps or uncertainty quantification essential for understanding performance differences.

➢ **Response:** Thank you for the advice. To address the gap in performance interpretation, we have incorporated **Error Maps** into the comparative visualizations in **Figures 6 and 8**. These maps provide a spatial quantification of segmentation uncertainty and errors, allowing for a more nuanced understanding of where the model excels and where its limitations lie compared to baseline methods.

The supporting data for these revisions can be found in Figures 6 and 8 (pertaining to Comment 4).

**Response to reviewer #2:**

COMMENTS TO THE AUTHOR(S)

This manuscript, entitled "DualSeg: Unified multi-scale framework with dual-stage encoder for glomerular segmentation," presents a dual-stage hybrid segmentation model that combines convolutional and recurrent attention mechanisms (Wave-Swin and VRWKV) to improve glomerular segmentation performance in kidney histopathology images. The topic is timely and relevant to renal pathology and computational histology. The model demonstrates good quantitative results on both mouse and human datasets.

However, despite these merits, the work still contains several limitations in both methodology and presentation, and a major revision is required before it can be considered for publication.

1) **the claimed novelty is not entirely convincing.** The proposed framework, while integrating CNN and VRWKV modules, appears conceptually similar to previously published hybrid architectures such as TransUNet, H2Former, and DA-TransUNet. The paper does not provide sufficient theoretical or empirical evidence to show that DualSeg fundamentally differs from these approaches. The authors should clearly articulate the unique contribution of their dual-stage design, beyond incremental improvements or architectural recombination. It would be helpful to include visual or quantitative analysis (e.g., feature map visualization, attention distribution comparison) to demonstrate how the proposed design contributes to performance beyond existing hybrid models.

**Response:** Thank you for your advice. We have addressed the concern regarding architectural novelty by clarifying the fundamental differences between DualSeg and existing hybrid designs like TransUNet and H2Former. Unlike traditional U-shaped cascades, DualSeg employs a hierarchical pyramid structure that sequentially integrates local wave-based texture refinement and global linear-complexity modeling (VRWKV). To provide empirical evidence of this advantage, we have included a visualization of the **Effective Receptive Field (ERF)** in **Fig. 1 (Top)**. The comparison demonstrates that while prior hybrid architectures often exhibit restricted or noisy receptive fields **(Fig. 1-III)**, our dual-stage design achieves a distinctively "clean and global" ERF **(Fig. 1-IV)**. This visualization confirms that DualSeg uniquely bridges the gap between local precision and global structural continuity, facilitating more robust feature extraction than previous architectural recombinations.

The revisions can be found in Section I. Introduction and Figure 1:

> **Page 2, Section I. Introduction:**
>
> Beyond standalone architectures, synergistic hybrid frame works have been explored. As illustrated in Fig. 2(a-d), most prior hybrids (e.g., TransUNet [8], H2Former [10] and U mamba [25]) adopt U-shaped paradigms. However, these often suffer from limitations: TransUNet compromises multi-scale capture; H2Former's shallow integration underutilizes ViTs; and their ERFs often remain suboptimal (Fig. 1(III)). Distinct from U-shaped models, pyramid architectures like SegFormer [11] (Fig. 2(e)) and SegNext [26] address the third chal lenge via feature fusion but rely on unidirectional extraction, weakening robustness against morphological variations. This prompts a critical inquiry: *Is it possible to integrate local and global features within a unified multi-scale framework to tackle all three segmentation challenges simultaneously?*
>
> To address this, we propose **DualSeg**, a novel hybrid framework that synergizes Wave Vision and VRWKV within a pyramid structure (Fig. 2(f)). Our architecture features a dual-stage encoder where early-stage Wave-

Swin blocks execute hierarchical local feature extraction to enhance texture discriminability, followed by VRWKV blocks that model long-range dependencies via linear attention to resolve spatial heterogeneity. This design effectively amalgamates the texture sensitivity of CNNs, the generalization of MLPs, and the scalability of VRWKV. By bridging local and global processing, DualSeg achieves robust multi-scale mapping of morphological priors. As shown in Fig. 1(IV) and Bottom, our method attains a global, clean ERF and SOTA segmentation performance. Notably, it achieves this with significantly reduced computational overhead (9.51 G FLOPs), offering an optimal accuracy-complexity trade-off superior to resource-intensive baselines.

**Page 1, Figure 1:**

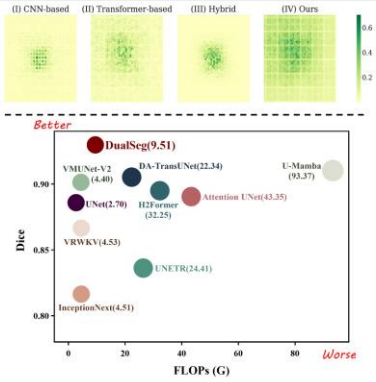

Fig. 1: **Top:** Comparison of *Effective Receptive Fields* (ERF). CNNs (I) show restricted local focus; Transformers (II) display dispersed global patterns; and current Hybrids (III) demonstrate subtimal coverage, failing to form a full context. Our method (IV) overcomes these limitations to generate a dense, global ERF. **Bottom:** Trade-off analysis. DualSeg occupies the optimal top-left position, delivering SOTA accuracy with significantly lower computational complexity (FLOPs) compared to existing methods.

2) **the generalization capability of the model remains insufficiently evaluated.** The experiments are limited to PAS-stained mouse and human datasets that share similar imaging conditions. Without cross-stain or cross-center validation, the robustness of the model under real-world variations in staining or scanner parameters cannot be confirmed. The authors should include additional experiments or at least discuss the expected behavior of the model under stain variability. It is also recommended to cite and discuss two closely related works—"Unsupervised stain augmentation enhanced glomerular instance segmentation on pathology images" and "Identifying and matching 12-level multistained glomeruli via deep learning for diagnosis of glomerular diseases"—to better position this study within the current research landscape.

**Response:** Thank you for your constructive suggestions. We have **incorporated the KPMP dataset** for a more comprehensive evaluation of generalizability and **updated our reference**, as detailed below:

- **Expanded Validation Experiments:** We have incorporated a cross-center validation by performing direct inference on the **KPMP** dataset using models trained exclusively on KPIs data, which is detailed in **Section IV. A. Datasets** and results are illustrated in **Table III and Figure 9**. This evaluation demonstrates DualSeg's exceptional stability across significant biological and technical variations without the need for domain adaptation.
- **Expanded Discussion on Staining Variability:** While public datasets for non-PAS stains remain scarce, we have added a dedicated discussion on stain variability and its impact on clinical deployment in **Section VI. D. Limitations and Future Work**.
- **Updated Literature and Citations:** we have integrated and discussed the suggested literature in **Section II. B. Technological Evolution of Glomerular Segmentation Architectures**,

which better positions DualSeg within the current landscape of robust renal histopathology analysis.

The revisions can be found in Section II. B. Related Work, Section IV. A. Dataset, Section VI. D. Limitations and Future Work, Table III and Figure 8:

**Page 3, Section II. B. Related Work:**

To bridge the gap between local precision and global context, Transformer-based methods like SegFormer [11] introduced self-attention mechanisms, the effectiveness of which has been corroborated in glomerular segmentation studies [37]–[39]. However, standard self-attention faces scalability constraints when processing the extensive spatial dimensions characteristic of Whole Slide Images (WSIs). This limitation has catalyzed the emergence of alternative global modeling paradigms—notably VRWKV [23] and U-Mamba [25], [40]—which utilize recurrent formulations or SSM to achieve effective global receptivity on high-resolution inputs. Although existing hybrid frameworks attempt to synergize the strengths of convolution and attention mechanism [13], achieving seam less multi-scale integration that fully preserves structural con tinuity remains an ongoing challenge.

**Page 7, Section IV. A. Datasets:**

Dataset III: Human Glomeruli (KPMP). To assess cross species generalization, we retrieved a second human dataset from the Kidney Precision Medicine Project (KPMP) Atlas Repository [50]. Four PAS-stained SVS format WSIs (avg. resolution 84,000×50,000) were selected with corresponding masks. To rigorously validate generalization, models trained solely on the mouse KPIs dataset were directly applied to this human dataset without retraining. For preprocessing consistency, KPMP WSIs were partitioned into 2,048 × 2,048 patches.

**Page 12, Section VI. D. Limitations and Future Work:**

While DualSeg demonstrates superior performance in glomerular segmentation, three primary limitations remain to be addressed in future iterations. First, the model's generalization to rare pathological subtypes, such as global glomerulosclerosis, is currently constrained by data scarcity. We plan to mitigate this by employing domain adaptation and transfer learning techniques to enhance feature robustness for these underrepresented classes. Second, our current validation is restricted to PAS-stained images. To ensure broad clinical applicability, we will extend our evaluation to include Hematoxylin-Eosin (HE) stained datasets, verifying the model's adaptability to varying histological protocols. Finally, despite the linear complexity of the VRWKV block, the computational overhead for gigapixel WSI processing remains significant. Future work will focus on model quantization and lightweight optimization to facilitate deployment on resource-constrained platforms and edge devices.

**Page 6, Table III:**

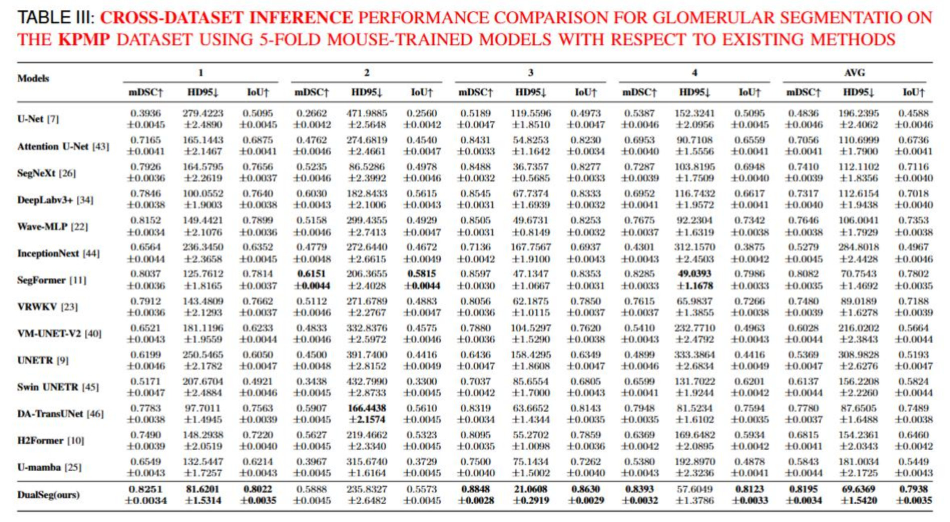

TABLE III: **CROSS-DATASET INFERENCE** PERFORMANCE COMPARISON FOR GLOMERULAR SEGMENTATIO ON THE **KPMP** DATASET USING 5-FOLD MOUSE-TRAINED MODELS WITH RESPECT TO EXISTING METHODS

| Models | 1 | | | 2 | | | 3 | | | 4 | | | AVG | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | mDSC↑ | HD95↓ | IoU↑ | mDSC↑ | HD95↓ | IoU↑ | mDSC↑ | HD95↓ | IoU↑ | mDSC↑ | HD95↓ | IoU↑ | mDSC↑ | HD95↓ | IoU↑ |
| U-Net [7] | 0.3936<br>±0.0045 | 279.4223<br>±2.4890 | 0.5095<br>±0.0045 | 0.2662<br>±0.0042 | 471.9885<br>±2.5648 | 0.2560<br>±0.0042 | 0.5189<br>±0.0047 | 119.5596<br>±1.8510 | 0.4973<br>±0.0047 | 0.5387<br>±0.0046 | 152.3241<br>±2.0956 | 0.5095<br>±0.0045 | 0.4836<br>±0.0046 | 196.2395<br>±2.4062 | 0.4588<br>±0.0046 |
| Attention U-Net [43] | 0.7165<br>±0.0041 | 165.1443<br>±2.1467 | 0.6875<br>±0.0041 | 0.4762<br>±0.0046 | 274.6819<br>±2.4661 | 0.4540<br>±0.0047 | 0.8431<br>±0.0033 | 54.8253<br>±1.1642 | 0.8230<br>±0.0034 | 0.6953<br>±0.0040 | 90.7108<br>±1.5556 | 0.6559<br>±0.0041 | 0.7056<br>±0.0041 | 110.6999<br>±1.7900 | 0.6736<br>±0.0041 |
| SegNeXt [26] | 0.7926<br>±0.0036 | 164.5795<br>±2.2619 | 0.7656<br>±0.0037 | 0.5235<br>±0.0046 | 86.5286<br>±2.3992 | 0.4978<br>±0.0046 | 0.8488<br>±0.0032 | 36.7357<br>±0.5685 | 0.8277<br>±0.0033 | 0.7287<br>±0.0039 | 103.8195<br>±1.7509 | 0.6948<br>±0.0040 | 0.7410<br>±0.0039 | 112.1102<br>±1.8356 | 0.7116<br>±0.0040 |
| DeepLabv3+ [34] | 0.7846<br>±0.0038 | 100.0552<br>±1.9003 | 0.7640<br>±0.0038 | 0.6030<br>±0.0043 | 182.8433<br>±2.1006 | 0.5615<br>±0.0043 | 0.8545<br>±0.0031 | 67.7374<br>±1.6939 | 0.8333<br>±0.0032 | 0.6952<br>±0.0041 | 116.7432<br>±1.9572 | 0.6617<br>±0.0041 | 0.7317<br>±0.0040 | 112.6154<br>±1.9438 | 0.7018<br>±0.0040 |
| Wave-MLP [22] | 0.8152<br>±0.0034 | 149.4421<br>±2.1076 | 0.7899<br>±0.0036 | 0.5158<br>±0.0046 | 299.4355<br>±2.7413 | 0.4929<br>±0.0047 | 0.8505<br>±0.0031 | 49.6731<br>±0.8149 | 0.8253<br>±0.0032 | 0.7675<br>±0.0037 | 92.2304<br>±1.6319 | 0.7342<br>±0.0038 | 0.7646<br>±0.0038 | 106.0041<br>±1.7929 | 0.7353<br>±0.0038 |
| InceptionNext [44] | 0.6564<br>±0.0044 | 236.3450<br>±2.3658 | 0.6352<br>±0.0045 | 0.4779<br>±0.0048 | 272.6440<br>±2.6615 | 0.4672<br>±0.0049 | 0.7136<br>±0.0042 | 167.7567<br>±1.9100 | 0.6937<br>±0.0043 | 0.4301<br>±0.0043 | 312.1570<br>±2.4503 | 0.3875<br>±0.0042 | 0.5279<br>±0.0045 | 284.8018<br>±2.4428 | 0.4967<br>±0.0046 |
| SegFormer [11] | 0.8037<br>±0.0036 | 125.7612<br>±1.8165 | 0.7814<br>±0.0037 | **0.6151**<br>**±0.0044** | 206.3655<br>±2.4028 | **0.5815**<br>**±0.0044** | 0.8597<br>±0.0030 | 47.1347<br>±1.0667 | 0.8353<br>±0.0031 | 0.8285<br>±0.0033 | **49.0393**<br>**±1.1678** | 0.7986<br>±0.0033 | 0.8082<br>±0.0035 | 70.7543<br>±1.4692 | 0.7802<br>±0.0035 |
| VRWKV [23] | 0.7912<br>±0.0036 | 143.4809<br>±2.1293 | 0.7662<br>±0.0037 | 0.5112<br>±0.0046 | 271.6789<br>±2.2767 | 0.4883<br>±0.0047 | 0.8056<br>±0.0036 | 62.1875<br>±1.0115 | 0.7850<br>±0.0037 | 0.7615<br>±0.0037 | 65.9837<br>±1.3855 | 0.7266<br>±0.0038 | 0.7480<br>±0.0039 | 89.0189<br>±1.6278 | 0.7188<br>±0.0039 |
| VM-UNET-V2 [40] | 0.6521<br>±0.0043 | 181.1196<br>±1.9559 | 0.6233<br>±0.0044 | 0.4833<br>±0.0046 | 332.8376<br>±2.5972 | 0.4575<br>±0.0046 | 0.7880<br>±0.0036 | 104.5297<br>±1.5290 | 0.7620<br>±0.0038 | 0.5410<br>±0.0043 | 232.7710<br>±2.4792 | 0.4963<br>±0.0043 | 0.6028<br>±0.0044 | 216.0202<br>±2.3843 | 0.5664<br>±0.0044 |
| UNETR [9] | 0.6199<br>±0.0046 | 250.5465<br>±2.1782 | 0.6050<br>±0.0047 | 0.4500<br>±0.0048 | 391.7400<br>±2.8152 | 0.4416<br>±0.0049 | 0.6436<br>±0.0047 | 158.4295<br>±1.8608 | 0.6349<br>±0.0047 | 0.4899<br>±0.0046 | 333.3864<br>±2.6834 | 0.4416<br>±0.0049 | 0.5369<br>±0.0047 | 308.9828<br>±2.6276 | 0.5193<br>±0.0047 |
| Swin UNETR [45] | 0.5171<br>±0.0047 | 207.6704<br>±2.4884 | 0.4921<br>±0.0046 | 0.3438<br>±0.0045 | 432.7990<br>±2.8733 | 0.3300<br>±0.0045 | 0.7037<br>±0.0042 | 85.6554<br>±1.7000 | 0.6805<br>±0.0043 | 0.6599<br>±0.0041 | 131.7022<br>±1.9244 | 0.6201<br>±0.0042 | 0.6137<br>±0.0044 | 156.2208<br>±2.2260 | 0.5824<br>±0.0044 |
| DA-TransUNet [46] | 0.7783<br>±0.0038 | 97.7011<br>±1.4945 | 0.7563<br>±0.0039 | 0.5907<br>±0.0045 | **166.4438**<br>**±2.1574** | 0.5610<br>±0.0045 | 0.8319<br>±0.0034 | 63.6652<br>±1.4344 | 0.8143<br>±0.0035 | 0.7948<br>±0.0035 | 81.5234<br>±1.6102 | 0.7594<br>±0.0035 | 0.7780<br>±0.0037 | 87.6505<br>±1.6488 | 0.7489<br>±0.0038 |
| H2Former [10] | 0.7490<br>±0.0039 | 148.2938<br>±2.0519 | 0.7220<br>±0.0040 | 0.5627<br>±0.0045 | 219.4662<br>±2.3340 | 0.5323<br>±0.0045 | 0.8095<br>±0.0035 | 55.2702<br>±1.0098 | 0.7859<br>±0.0036 | 0.6369<br>±0.0042 | 169.6482<br>±2.0895 | 0.5934<br>±0.0042 | 0.6815<br>±0.0041 | 154.2361<br>±2.0343 | 0.6460<br>±0.0042 |
| U-mamba [25] | 0.6549<br>±0.0043 | 132.5447<br>±1.7257 | 0.6214<br>±0.0043 | 0.3967<br>±0.0045 | 315.6740<br>±1.6164 | 0.3729<br>±0.0045 | 0.7500<br>±0.0040 | 75.1434<br>±1.5002 | 0.7262<br>±0.0040 | 0.5380<br>±0.0043 | 192.8970<br>±2.3236 | 0.4878<br>±0.0041 | 0.5843<br>±0.0044 | 181.0034<br>±2.1725 | 0.5449<br>±0.0043 |
| DualSeg(ours) | **0.8251**<br>**±0.0034** | **81.6201**<br>**±1.5314** | **0.8022**<br>**±0.0035** | 0.5888<br>±0.0045 | 235.8327<br>±2.6482 | 0.5573<br>±0.0045 | **0.8848**<br>**±0.0028** | **21.0608**<br>**±0.2919** | **0.8630**<br>**±0.0029** | **0.8393**<br>**±0.0032** | 57.6049<br>±1.3786 | **0.8123**<br>**±0.0033** | **0.8195**<br>**±0.0034** | **69.6369**<br>**±1.5420** | **0.7938**<br>**±0.0035** |

**Page 11, Figure 8:**

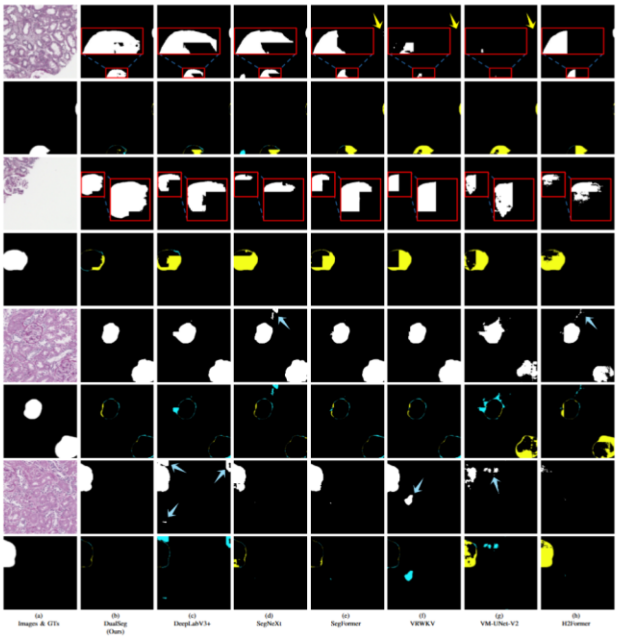

Fig. 8: Visual comparison on the held-out KPMP test set. **Odd rows** display original images and inference masks; **even rows** show GT and error maps, where **yellow** and **green** denote under-segmentation and over-segmentation, respectively. Red boxes and matching arrows highlight magnified details and segmentation errors. **DualSeg** exhibits unprecedented stability when handling cross-center and cross-species data.

3) **the clinical relevance and interpretability of the segmentation results are not adequately discussed. Although the paper emphasizes segmentation accuracy, it does not explain how this improvement translates into clinical or diagnostic benefits, such as better lesion quantification, assessment of glomerulosclerosis, or prediction of disease progression. Including examples or quantitative analyses connecting segmentation quality to potential diagnostic metrics would significantly strengthen the manuscript. A brief discussion of the model's interpretability from a pathologist's perspective would also be valuable.**

**Response:** Thank you for the feedback. We have revised the manuscript to bridge the gap between technical accuracy and diagnostic utility, as detailed below:

- **Statistical Validation of Clinical Utility:** We have enriched **Section VI. C. Clinical Relevance** with statistical analysis to demonstrate the diagnostic value of our mode. Specifically, we incorporated a paired *t*-test analysis in **Figure 10**, demonstrating that DualSeg's performance gains are statistically significant (p<0.05 to p<0.001) across diverse cohorts. This statistical robustness supports the model's reliability for the standardized quantification of critical biomarkers, such as glomerulosclerosis and fibrosis, which are essential for predicting disease progression.
- **Adding Error Map:** To enhance interpretability from a pathologist's perspective, we have introduced Error Maps in **Figures 6 and 8**. These maps utilize red bounding boxes to magnify local anatomical details and provide a spatial audit of model performance by color-coding under-segmentation (yellow) and over-segmentation (green). This visualization allows clinicians to assess diagnostic confidence more effectively and helps mitigate inter-observer variability in complex histopathological scenarios .

The revisions can be found in Section VI. C. Clinical Relevance, Figures 6, 8 and 10:

**Page 12, Section VI. C. Clinical Relevance:**

DualSeg exhibits statistically significant superiority (p<0.05–0.001; Fig. 10) and exceptional reproducibility, evidenced by a minimal standard deviation (0.0040) on the HuBMAP dataset. Its ability to accurately resolve diverse morphologies—ranging from mild hypertrophy to severe fragmentation—enables the precise quantification of pathological biomarkers like sclerosis and fibrosis. Furthermore, the model's robust performance on the cross-species KPMP dataset supports standardized CKD monitoring. By mitigating inter-observer variability and reducing manual annotation burdens, DualSeg provides a scalable solution for multi-center clinical trials and routine diagnostic workflows.

**Page 7, Figure 6:**

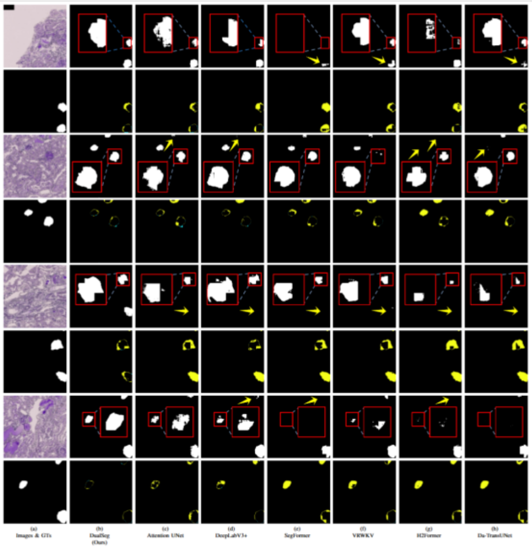

Fig. 6: Visual comparison on the KPIs test set. **Odd rows** display original images and inference masks; **even rows** show GT and error maps, where **yellow** and **green** denote under-segmentation and over-segmentation, respectively. Red boxes and matching arrows highlight magnified details and specific segmentation errors. **DualSeg** demonstrates superior performance in handling heterogeneous glomeruli with intricate boundaries.

**Page 9, Figure 8:**

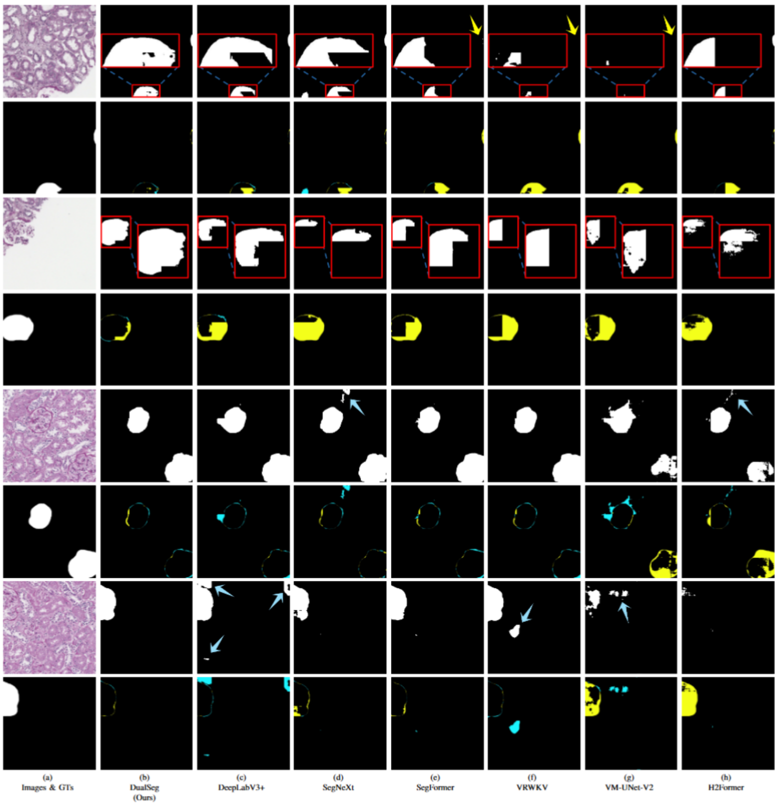

Fig. 8: Visual comparison on the held-out KPMP test set. **Odd rows** display original images and inference masks; **even rows** show GT and error maps, where **yellow** and **green** denote under-segmentation and over-segmentation, respectively. Red boxes and matching arrows highlight magnified details and segmentation errors. **DualSeg** exhibits unprecedented stability when handling cross-center and cross-species data.

**Page 12, Figure 10:**

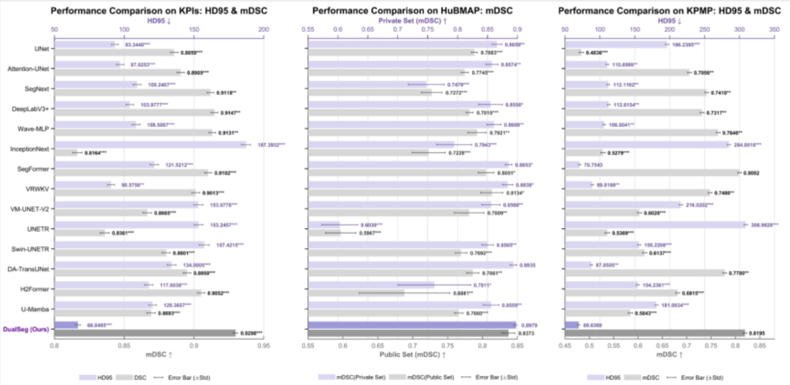

Fig. 10: Performance comparison between our DualSeg model and 14 baseline methods across three datasets. The comparison metrics include mDSC and HD95 for the KPIs and KPMP datasets (left and right panels, respectively), and mDSC for the HuBMAP dataset (middle panel). The error bars represent ± standard deviation. Statistical significance was assessed using paired $t$-tests, with levels indicated by asterisks: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

4) **the methodological presentation is unnecessarily complicated.** The mathematical description of the Wave-Swin and VRWKV blocks is long and dense, making it difficult for readers from the biomedical community to grasp the main idea. Some mathematical symbols, such as Θ and W^T_jk, are not clearly defined, and equation formatting is occasionally inconsistent. The authors should simplify the explanation of equations and focus on the conceptual understanding of each component's role in the overall framework, leaving detailed derivations to supplementary materials if necessary.

**Response:** Thank you for your advice. We have streamlined the methodological presentation in **Section III** to improve accessibility for the biomedical community while maintaining technical precision. The mathematical descriptions of the Wave-Swin and VRWKV blocks have been simplified to focus on their conceptual roles in resolving texture discriminability and spatial heterogeneity. We have standardized the formatting across all mathematical expressions—specifically **Equations (1) through (5)**—and provided explicit definitions for all symbols to ensure clarity and consistency. By prioritizing the structural intuition of the dual-stage framework, we have ensured the methodology remains both rigorous and accessible to a broader audience.

The revisions can be found in Section III Methodology:

**Page 3, Section III Methodology:**

**A. Wave-Swin Block**

…

1) *Wave Formulation*: Let the input feature map be denoted as X=[x1,x2,...,xn],where each xj represents a token. Instead of standard linear projections, we map these tokens into a complex wave representation to capture both semantic intensity and spatial structural priors. We define the complex wave form zj for the j-th token as:

$$z_j = \mathcal{A}(x_j) \odot \exp\left(i \cdot \mathcal{P}(x_j)\right), \qquad (1)$$

where A(·) and P(·) denote learnable linear transformations that project the input to amplitude and phase terms, respectively. The operator ⊙ represents element-wise multiplication. ⋯ These components are then aggregated via adynamic token mixing operation:

$$y_j = \sum_{k \in \Omega_j} \left(\mathrm{Re}(z_k) W_{Re} + \mathrm{Im}(z_k) W_{Im}\right), \qquad (2)$$

Where yj is the output token, Ωj denotes the dynamic propagation window centered at j, and WRe, Wim are learnable weights governing the fusion of spatial components.

…

**B. VRWKV Block**

…

1) *Z-Shift and Spatial Mixing*: …

After the spatial shift, the feature maps are flattened into token sequences and projected. The generation of the receptor, key, and value matrices is formulated as:

$$N_s = \text{Linear}_N\left(Flatten(\text{Z-Shift}(X))\right), \quad N \in \{R, K, V\}, \qquad (3)$$

where X denotes the input 2D feature map, and LinearN represents the learnable projection weights. The flattened tokens then undergo the linear-complexity bidirectional attention aggregation:

$$S = \sigma(R_s) \odot \text{Bi-WKV}(K_s, V_s), \qquad (4)$$

where $\sigma(\cdot)$ is the sigmoid activation, $\odot$ denotes element-wise multiplication, and Bi-WKV is the time-mixing operator that aggregates global context with linear complexity O(L), efficiently capturing pairwise affinities between distant glomerular candidates.

2) Channel Mixing: Following spatial aggregation, the features undergo Channel Mixing to enable inter-channel communication. This module mirrors the gating mechanism of the spatial stage but focuses on feature refinement within the channel dimension. The transition is expressed as:

$$O_c = \sigma(R_c) \odot (\text{SqReLU}(K_c) \cdot W_v), \qquad (5)$$

where Rc and Kc are derived from the spatially mixed features via linear projections, and SqReLU denotes the squared ReLU activation.

5) **the experimental analysis requires stronger statistical and methodological support.** Although the authors conduct ablation studies, the reported improvements are small, and the absence of variance analysis or statistical testing makes it unclear whether the gains are significant. The paper would benefit from a more comprehensive evaluation, including inference time, parameter count, and performance on challenging subtypes such as sclerotic or crescentic glomeruli. Additional qualitative examples demonstrating both the strengths and failure cases of DualSeg would help provide a balanced assessment of the model's robustness.

**Response:** Thank you for your advice. We have included a summary of t-tests, FLOPs comparison, and failure case analysis to strengthen the empirical foundation of our study. We have addressed the specific concerns as follows:

➢ **Enhanced Statistical Analysis:** We have significantly strengthened the statistical and methodological support for our experimental analysis. As detailed in newly added **Section VI. C. Clinical Relevance** and **Figure 10**, we performed ***t*-tests across all three datasets**, confirming that DualSeg's performance gains are statistically significant ($p < 0.05$ to $p<0.001$). To address reproducibility, we included variance analysis (error bars) in all primary results, with DualSeg exhibiting exceptional stability (e.g., a minimal standard deviation of 0.0040 on the HuBMAP dataset).

➢ **Expanded FLOPs Comparison:** We incorporated a **Performance-vs-FLOPs comparison in Figure 1 (Bottom)** to justify computational efficiency and expanded our qualitative evaluation to include challenging subtypes such as fragmented and sclerotic glomeruli.

➢ **Addition of Failure Case Assessment:** Finally, we have provided a balanced assessment in **Section VI. B. Failure Case Analysis and Figure 10** by analyzing specific failure cases, offering a transparent view of the model's current limitations and future improvement directions.

The supporting data for these revisions can be found in Figure 1, 3(pertaining to Comment 1), 11, Section VI. C. Clinical Relevance (pertaining to Comment 3), Figure 10 and Section VI. B. Failure Case Analysis, which are provided below:

**Page 12, Section VI. B. Failure Case Analysis:**

Fig. 10 reveals that the model occasionally fails to detect globally sclerotic glomeruli in the KPMP dataset. This limitation stems primarily from two factors: the partial truncation of peripheral glomeruli during WSI tiling, which compromises morphological context, and the significant divergence of unseen, extreme pathological variants. To mitigate this, future work could increase patch sizes to preserve boundary information or, more efficiently, integrate uncertainty-guided semi-supervised learning. This strategy aims to enhance robustness against rare phenotypes without incurring excessive computational overhead.

**Page 12, Figure 10:**

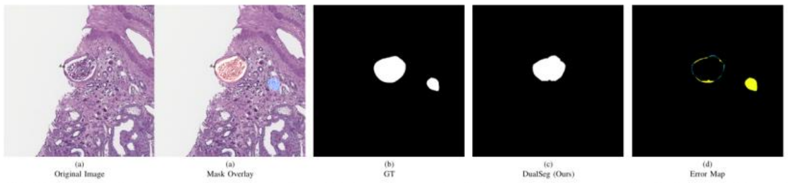

Fig. 10: Visual analysis of segmentation limitations. In the error map (far right), the yellow region highlights a significant false negative (under-segmentation). This failure is primarily attributed to the absence of globally sclerotic samples in the training set, preventing the model from generalizing to the *high heterogeneity* and *distinct morphological* features of this unseen pathology.

**Finally, several minor issues should be corrected:**

1) Figure fonts are too small to read, and color schemes in Figures 3–6 are inconsistent. Figure 5 panels are misaligned, and the arrows indicating regions of interest are too faint.

2) Table I has several misaligned columns and overlapping text.

3) There are typographical errors such as "uqualitative" (should be "qualitative") and repeated line-break hyphenations such as "glomeru- lar," which should be corrected.

4) Ensure consistent capitalization of dataset names (e.g., HuBMAP, KPIs) and verb tense consistency throughout the text. References should be carefully checked for completeness and formatting.

**Response:** We have performed a comprehensive editorial overhaul of the manuscript:

- **Figure Enhancements:** We have revised all figures to ensure maximum legibility and stylistic consistency. Specifically, font sizes were increased across all plots, and the color schemes in **Figures 3–6** were unified to maintain a cohesive visual identity. Panel alignment in **Figure 5** has been corrected, and the indicating arrows have been thickened and brightened to clearly highlight regions of interest.

The revisions can be found in Figure 4 and 5:

**Page 4, Figure 4:**

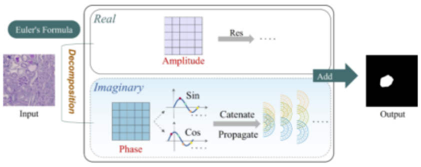

Fig. 4: Illustration of the Token-Mixing mechanism within the **Wave-Swin** Block. The input is initially decomposed into a *real* component and an *imaginary* component via Euler's formula. The modulated phase undergoes two distinct cosine transformations and window-based propagation, and is subsequently fused with the amplitude information to generate the output.

**Page 5, Figure 5:**

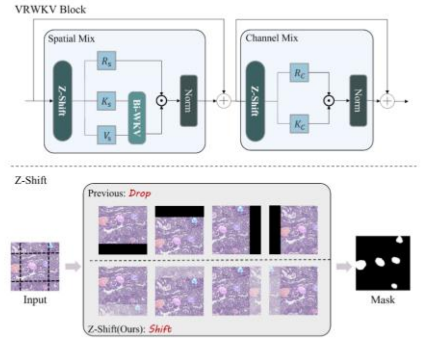

Fig. 5: Structure of the improved VRWKV block. The black margins in the lower panel illustrate the prior approach where pixels shifted beyond the boundary are simply dropped. In contrast, our **Z-Shift** wraps these pixels to the opposite edge, filling the empty regions and eliminating information loss.

- **Table Reformatting:** **Table I** has been reformatted to resolve column misalignments and overlapping text. The presentation of metrics, including mDSC, HD95, and IoU, is now clear and professionally aligned .
- **Textual Corrections:** All typographical errors, such as "uqualitative," have been corrected. We have also removed improper line-break hyphenations (e.g., "glomeru-lar") throughout the text to ensure linguistic fluidity.
- **Consistency Unification and Reference Validation:** We have conducted a full audit of the manuscript to ensure consistent capitalization of dataset names, such as **HuBMAP, KPIs, and KPMP**. Verb tenses have been unified, and the reference list has been verified for completeness and adherence to the journal's formatting standards.