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 to ensure the document remains readable. 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 agree that rigorous benchmarking is essential to support any claims of computational advantage. We have addressed this concern through the following revisions:

- **Clarification and Text Revision:** We recognize that the original statement—referring to a "60% reduction in computational overhead"—was imprecisely phrased and lacked direct empirical support within our specific experimental setup. To maintain the highest level of academic rigor, we have **removed this specific claim** from the revised manuscript.
- **Quantitative Benchmarking:** 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**). This metric provides a standard and objective measure of the computational complexity and the performance-efficiency trade-off of our proposed framework.

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 fully concur that multi-staining compatibility and cross-institutional validation are benchmarks for clinical-grade algorithms. We have addressed these concerns as follows:

- **Expanded Cross-Institutional and Cross-Species Validation:** To rigorously test the generalizability of DualSeg, we have incorporated a new external dataset—the **KPMP (Kidney Precision Medicine Project) dataset** (now detailed in **Section IV. A, Dataset III**). This allowed us to perform an extensive validation encompassing: ***Cross-Species Generalization:*** From mouse models to human clinical specimens. ***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.
- **Addressing Multi-Stain Limitations:** We acknowledge that public datasets featuring synchronized multi-stain (e.g., H&E, Masson) annotations for this specific task remain extremely scarce. To maintain scientific integrity, we have added a dedicated discussion on this in **Section VI. D (Limitations and Future Work)**. We explicitly state that while DualSeg shows exceptional performance on PAS-stained images, its direct clinical deployment would require further validation on other staining protocols as such data becomes available.

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.

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 acknowledge that InceptionNeXt and Mamba-based models (VM-UNet and U-Mamba) represent the current state-of-the-art in efficient deep

learning and provide critical benchmarks for evaluating our DualSeg framework. We have addressed the perceived "methodological gap" by conducting a comprehensive re-evaluation **(Section IV. B, Baselines)**. The results, detailed in **Tables 1, 2, and 3** and visualized in **Figures 6, 7, 8, and 10**, consistently demonstrate that DualSeg achieves superior segmentation accuracy and maintains a more favorable performance-to-complexity ratio.

Page 8, 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].

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 agree that for an AI tool to be truly impactful, its performance gains must be both statistically robust and practically feasible. We have addressed these concerns as follows:

- **Deployment Feasibility and Hardware Constraints:** We acknowledge the reviewer's point regarding the resource-constrained nature of clinical environments (e.g., 8GB GPU limits). While this study primarily focuses on architectural innovation and accuracy benchmarks, we have added a comprehensive discussion in **Section VI. D (Limitations and Future Work)**. This section now outlines the roadmap for future clinical deployment, including planned optimizations for inference on consumer-grade GPUs and CPU-only workstations through model quantization and pruning.
- **Clinical Relevance of Performance Gains:** To address the clinical significance of the 2.73% mDSC improvement, we have enriched **Section VI. C (Clinical Relevance)** with the following evidence: ***Statistical Significance:*** As illustrated in the updated **Figure 10**, we performed rigorous t-tests across three diverse datasets (**KPIs, HuBMAP, and KPMP**). The results indicate that DualSeg's performance gains are statistically significant ($p < 0.001$ in most metrics), suggesting a robust improvement rather than marginal fluctuation. ***Reduction in Diagnostic Omissions:*** Beyond the mDSC metric, we have included an analysis of "omission cases" in **Tables 6-8**. Our results show that DualSeg significantly reduces the frequency of missed glomerular structures compared to baseline models. In a clinical setting, this directly translates to higher diagnostic sensitivity and reduced workloads for pathologists by minimizing manual corrections.

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 (0.004) 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.

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 pointing that out. We have addressed these concerns as follows:

- **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 **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*** and ***dataset-specific anatomical statistics***. We have significantly expanded **Section III. A. 2 (Dynamic Swin Mechanism)** to detail this rationale.

Page 5, 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.

Minor Concerns

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

- **Response:** To better balance technical density with clinical utility, we have enriched **Section VI. C (Clinical Relevance)** with a rigorous 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.

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

Response: Thank you for pointing that out. We have addressed these concerns as follows:

- **Response:** We have addressed the concern regarding computational overhead by incorporating a system-level efficiency analysis in **Fig. 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.

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

- **Response:** To address the gap in performance interpretation, we have incorporated **Error Maps** into the comparative visualizations in **Figures 6, 7, 8, and 9**. 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.

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.

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: Thanks to your suggestions, we have incorporated a rigorous cross-center and cross-species validation by performing direct inference on the human **KPMP** dataset using models trained exclusively on mouse data. This ***zero-shot*** evaluation demonstrates DualSeg's exceptional stability across significant biological and technical variations without the need for domain adaptation. 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)**. Furthermore, we have integrated and discussed the suggested literature in Section II. B, which better positions DualSeg within the current landscape of robust renal histopathology analysis.

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.

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: Thanks for pointing that out. We have addressed this suggestion by expanding the discussion on the diagnostic implications of our results in **Section VI. C (Clinical Relevance)**. Specifically, we have incorporated a rigorous t-test analysis in Figure 10, which demonstrates that DualSeg's performance gains are statistically significant ($p<0.05$-0.001), supporting its reliability for standardized quantification of biomarkers such as glomerulosclerosis and fibrosis. To enhance interpretability from a pathologist's perspective, we have introduced **Error Maps** in **Figures 6, 7, 8, and 9**. These maps provide a spatial audit of model performance by color-coding under-segmentation (yellow) and over-segmentation (green), allowing clinicians to assess diagnostic confidence and effectively mitigate inter-observer variability in complex histopathological scenarios.

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 (0.004) 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.

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 thoroughly 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 (6)**—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.

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 addressed these concerns as follows:

- We have significantly strengthened the statistical and methodological support for our experimental analysis. As illustrated in the newly added **Figure 10**, we performed rigorous 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).
- Furthermore, we incorporated a **performance-vs-FLOPs comparison in Figure 1** to justify computational efficiency and expanded our qualitative evaluation to include

challenging subtypes such as fragmented and sclerotic glomeruli.

- Finally, we have provided a balanced assessment in **Section VI. B and Figure 9** by analyzing specific failure cases, offering a transparent view of the model's current limitations and future improvement directions.

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

Fig. 9 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.

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 sincerely apologize for these presentation oversights and 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.
- **Table Reformatting: Table I** has been thoroughly 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 and References:** 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 meticulously verified for completeness and adherence to the journal’s formatting standards.