RESPONES-TO-REVIEWER-PSZR

Anonymous authors
Paper under double-blind review

Dear Reviewer PSzr, we sincerely appreciate your valuable feedback on our submission. Below, we provide our responses to the concerns you raised. We have incorporated these revisions into the updated version of the manuscript, and we believe they will help improve the overall quality of the submission.

WEAKNESSES:

Comment 1

While the theoretical framework is novel, the core idea of using frequency-domain information for computer vision is well-established. (i.e, ICCV'23 works like SPANet [1]) and recent methods FAD [2]. The paper doesn't sufficiently differentiate its approach from these existing frequency-based methods.

We appreciate you for highlighting this point. We have rewritten the **Related Work** section to more clearly articulate the differences between our method and existing frequency-domain approaches, and have supplemented it with corresponding references. Although our method also leverages frequency-domain information, it differs fundamentally from existing methods in terms of theoretical motivation, fusion strategy, and its role within the overall architecture.

Theory-Driven Design vs. Heuristic Application: Most existing methods (e.g., SPANet[1], FAD[2], FreqMamba[3], etc.) utilize frequency information as a heuristic tool for feature enhancement or domain adaptation. In contrast, our pre-modulation mechanism is rigorously derived from our theoretical framework. Our designed GSSM does not merely "add" frequency features. We employ a DFT-based pre-modulation approach to inject global properties into the input signal u_t before it enters the SSM system. This ensures that the model satisfies the rigorous mathematical definition of global perception and non-causal constraints.

Based on this theoretical foundation, our integration strategy and structural role differ:

Existing Methods (Post-Fusion): Methods such as FAD and SPANet typically treat frequency features as an auxiliary branch, fusing them with spatial features in a parallel or post-processing manner as a complement to spatial-domain information.

Our Method (Pre-Modulation): GSSM uses frequency coefficients to modulate the state transition process itself. By embedding global information into the input at the front end, we explicitly condition the SSM's selective scanning and state updates through the frequency domain. This fundamentally changes how the model processes information, rather than merely enriching the final features.

To experimentally validate this theoretical advantage, we have added new experiments in **Appendix K.7** (Table 1). We compare our "pre-modulation" strategy with a "dual-branch fusion" strategy (ex-

isting frequency-domain-based methods). The results demonstrate that our pre-modulation approach achieves 86.00% mIoU, significantly outperforming the conventional approach (84.10% mIoU).

[1]Guhnoo Yun, Juhan Yoo, Kijung Kim, Jeongho Lee, and Dong Hwan Kim. Spanet: Frequency-balancing token mixer using spectral pooling aggregation modulation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6113–6124, 2023.

[2] Zhen Zou, Hu Yu, Jie Huang, and Feng Zhao. *Freqmamba: Viewing mamba from a frequency perspective for image deraining*. In Proceedings of the 32nd ACM International Conference on Multimedia, pp. 1905–1914, 2024.

[3]Ruixiao Shi, Fu Feng, Yucheng Xie, Jing Wang, and Xin Geng. Fad: Frequency adaptation and diversion for cross-domain few-shot learning. arXiv preprint arXiv:2505.08349, 2025a.

Table 1: Comparison of Frequency Enhancement Strategies in GMamba

Method	Frequency Strategy	Position	Mechanism	mIoU (%)
Baseline (GMamba)	None	None	None	83.11
GMamba w/ Dual-Branch Fusion	Dual-Branch Fusion	After SSM	Post-hoc feature fusion	84.10
GMamba w/ Pre-Modulation	Frequency Pre-Modulation	Before SSM	Frequency-guided state update	**86.00**

Comment 2

Existing work, such as Vim and VMamb, has demonstrated that SSMs can achieve global receptive fields through bidirectional processing and multi-directional scanning. The paper's claim that current SSMs lack global perception seems overstated.

We fully agree with you that, in principle, Vim and VMamba can achieve a global receptive field through bidirectional or multidirectional scanning. However, our core argument concerns how such globality is realized in the context of image data, and how its implementation mechanism, quality, and adaptability manifest.

Structural mismatch: As analyzed in our paper, the core of SSMs (including Mamba/Vim) is the stepwise autoregressive state update mechanism. This design originates from causal time-series modeling [1][2]. However, images are spatially non-causal; pixel relationships are parallel and lack inherent sequential order. Vim and VMamba mitigate this issue via different scanning strategies, but they do not alter the underlying "sequential recursive modeling" paradigm. In essence, a causal model is being imposed on non-causal data.

Indirect vs. Explicit mechanisms: Vim/VMamba rely on recursive state transitions, where the global context is accumulated indirectly through multiple steps. Consequently, the acquisition of global information is still derived progressively and indirectly from local information. In contrast, GSSM employs DFT-based pre-modulation to explicitly inject global information into each pixel representation before the SSM scan begins. Due to the position-invariance and non-causality of the DFT, it aligns naturally with image properties, realizing an explicit, parallel, and non-causal global receptive field.

We do not deny that Vim/VMamba possess a global receptive field. Our emphasis is that their mechanism is indirect, non-uniform, and built upon a "causal" modeling paradigm that conflicts with the inherent characteristics of images. We sincerely thank the reviewer for pointing out potential ambiguities in our phrasing, and we have refined the wording in the revised manuscript to more clearly and accurately reflect our contributions.

[1] Weihao Yu and Xinchao Wang. *MambaOut: Do we really need Mamba for vision?* In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 4484–4496, 2025.

[2] Qinfeng Zhu, Yuan Fang, Yuanzhi Cai, Cheng Chen, and Lei Fan. *Rethinking scanning strategies with Vision Mamba in semantic segmentation of remote sensing imagery: An experimental study*. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 17:18223–18234, 2024. doi: 10.1109/JSTARS.2024.3472296.

Comment 3

The related work and the comparison are missing recent state-space models (i.e, GroupMamba[3] (CVPR'25), MambaVision[4] (CVPR'25). Also, the provided implementation of GMamba_Block.py in the supplementary material appears to be largely derived from the MambaVisionMixer module introduced in MambaVision [4]. The authors have added their frequency components on top of that code. However, they neither cite nor compare their results with MambaVision paper [4].

We sincerely apologize for this oversight and deeply appreciate the reviewer's careful attention to this matter. We have now supplemented the "Related Work" section with references to GroupMamba [1] and MambaVision [2], and conducted a thorough review of existing methods, including the most recent works such as Spatial Mamba [3] and FreqMamba [4]. Additionally, we have included comparative results with these methods in the main manuscript tables.

We acknowledge that our Mixer implementation draws inspiration from the MambaVision architecture, following standard open-source practices to ensure system stability and compatibility. We recognize that this has resulted in structural similarities, and we appreciate the opportunity to clarify this aspect. However, we would like to respectfully emphasize that our core contribution lies in the GSSM module and its theoretical foundation, particularly the design and implementation of the Frequency Encoding Module and the Frequency-Guided Modulation Module. These components are developed based on our proposed frequency modulation theoretical framework and represent the unique innovations of our approach.

We have added detailed comments in the supplementary code to explicitly distinguish the novel modules implemented under our theoretical framework from the standard Mixer backbone. We hope this clarification better conveys the originality of our work while acknowledging the foundations upon which we built.

Once again, we sincerely apologize for any confusion this may have caused and are grateful for your constructive feedback, which has helped us improve the clarity and transparency of our manuscript.

[1] Abdelrahman Shaker, Syed Talal Wasim, Salman Khan, Juergen Gall, and Fahad Shahbaz Khan. *Groupmamba: Efficient group-based visual state space model*. In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 14912–14922, 2025.

[2] Ali Hatamizadeh and Jan Kautz. *Mambavision: A hybrid mamba-transformer vision backbone*. In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 25261–25270, 2025.

[3] Chaodong Xiao, Minghan Li, Zhengqiang Zhang, Deyu Meng, and Lei Zhang. *Spatial-mamba: Effective visual state space models via structure-aware state fusion*. In The Thirteenth International Conference on Learning Representations, 2025.

[4] Zhen Zou, Hu Yu, Jie Huang, and Feng Zhao. *Freqmamba: Viewing mamba from a frequency perspective for image deraining*. In Proceedings of the 32nd ACM International Conference on Multimedia, pp. 1905–1914, 2024.

Comment 4

The reported improvements are relatively modest (around 2–3% mIoU) while introducing a substantial parameter increase of 20–35%. It remains unclear whether these gains stem from the additional parameters or from the model's inherent effectiveness. For example, in Table 2, the baseline UNet-ConvNeXt(S) has 58.42M parameters, whereas incorporating GMamba increases this to 71.06M, yielding a 2.89% improvement in mIoU. A fair comparison would require scaling the baseline (e.g., by increasing the number of channels or blocks) to match 71.06M parameters, in order to determine whether the observed gains are truly architectural rather than parameter-driven.

We sincerely thank you for raising this important concern. To address it, we conducted a fair comparison experiment in **Appendix K.5** (Table 2) of the revised manuscript, using parameter-matched baselines.

Setup: We scaled the baseline networks (ResNet34, Swin-T, ConvNeXt-S) by increasing their channel widths to match the number of parameters of the corresponding "baseline + GMamba" models. Specifically, for ResNet34, the channel dimensions were increased from $[64, 128, 256, 512]$ to $[72, 144, 280, 568]$; for Swin-T, from $[96, 192, 384, 768]$ to $[112, 224, 448, 888]$; and for ConvNeXt-S, from $[64, 128, 256, 512]$ to $[104, 208, 424, 848]$.

Although scaling the baseline models slightly improves mIoU and mF1, GMamba consistently outperforms such parameter-only increases. These results indicate that GMamba's advantages stem not from additional parameters, but from its enhanced global modeling and frequency-guided state-space mechanism, demonstrating its efficiency and generality across different backbone architectures.

Table 2: Fair Comparison: Scaled Baselines Matching Params of +GMamba

Model Variant	Params (M)	mIoU (%)	mF1 (%)
ResNet34 (Baseline)	25.33	81.65	89.24
ResNet34 (Scaled, matched params)	31.06	83.18	90.59
ResNet34 + GMamba (Ours, matched params)	30.96	**84.74**	**91.56**
Swin-T (Baseline)	36.48	82.44	89.75
Swin-T (Scaled, matched params)	49.50	83.76	90.95
Swin-T + GMamba (Ours)	49.13	**84.83**	**91.61**
ConvNeXt-S (Baseline)	58.42	83.11	90.19
ConvNeXt-S (Scaled, matched params)	71.07	84.46	91.38
ConvNeXt-S + GMamba (Ours)	71.06	**86.00**	**92.31**

QUESTIONS:

Comment 1

It is unclear whether the GMamba models used for object detection and instance segmentation are initialized with backbones pre-trained on ImageNet or not. If yes, what is the top-1 accuracy of GMamba on ImageNet?

We thank you for raising the question regarding the initialization strategy. To clarify, our GMamba module is designed as an efficient plug-and-play component.

In our object detection and semantic segmentation experiments, we adopt standard backbone networks (ResNet34, ResNet50, Swin-T, and ConvNeXt-S) and initialize them using their official ImageNet-pretrained weights. We then insert the GMamba module into specific stages of these backbones, following the insertion strategy detailed in **Appendix H (Experimental Details)**. The parameters of GMamba itself are randomly initialized.

We did not perform standalone ImageNet pre-training for the combined architecture of "backbone + GMamba." Therefore, we are unable to report ImageNet Top-1 accuracy for a specific GMamba-integrated model.

Nevertheless, extensive experiments across various tasks and backbones consistently demonstrate the strong generalization capability and broad deployment potential of GMamba.