RESPONES-TO-REVIEWER-YRTS

Anonymous authors
Paper under double-blind review

Dear Reviewer YRTs, 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

> The paper's theoretical claims rest heavily on its new definition of global perception—"global gradient dependency". This definition requires that the Frobenius norm of the gradient of the output with respect to any input pixel is bounded by a non-zero constant τ . This is a very low bar. A gradient that is infinitesimally small but non-zero would satisfy this definition, but this may not align with an intuitive or practical understanding of "global influence." The claim of "true" global perception is therefore only as strong as the acceptance of this new, and debatable, definition.

We thank you for this excellent question, which concerns the mathematical rigor and practical significance of our proposed definition. We understand the reviewer's concern that a non-zero constant $\tau > 0$ may appear numerically to set a "low bar," but we respectfully point out that from topological and architectural perspectives, this is a rigorous discriminative criterion. Our definition is not intended to constrain the strength of influence (which should be learned by the weights), but rather to guarantee the structural accessibility of information. We clarify this through three points:

1. The definition distinguishes between "truly global" and "effectively local"

Self-Attention: Its "globality" is emergent and data-dependent, not inherently guaranteed. As we noted in the paper, the self-attention architecture does not enforce $\tau > 0$; attention weights are learned dynamically, and the learned weights may potentially focus only on local pixels.

Standard SSMs: As shown in Equation (3) of the paper, standard discrete-time SSMs rely on recurrent updates $x_t = \bar{A}x_{t-1} + \bar{B}u_t$. To ensure stability, the spectral radius $\rho(\bar{A}) < 1$, leading to exponential decay of influence with distance as $|\bar{A}|^d$. While τ is mathematically non-zero for finite sequences, in the context of flattened high-resolution images (with large sequence length L), this decay forces long-range gradients into numerical underflow, and moreover, SSMs do not satisfy our proposed constraint.

Therefore, the existence of a uniform non-zero constant τ is actually a very high structural threshold.

2. GSSM ensures spatially uniform potential

The core contribution of our frequency-domain modulated SSM theoretical framework lies in proving that GSSM satisfies this definition through 2D DFT modulation, while standard SSMs cannot. As detailed in Appendix E, the gradient norm of DFT components with respect to input pixels is $\sqrt{HW}$ (or $1/\sqrt{HW}$ for the inverse norm). This is not an infinitesimal ϵ; it is a constant determined by resolution, entirely independent of spatial distance between pixels. This demonstrates that GSSM possesses ”information accessibility”: every pixel has an equal and non-decaying theoretical path to influence all other pixels. This contrasts sharply with the inherent ”distance bias” in recurrent scanning.

3. Why a ”higher standard” is not advisable

The definition ensures the *capability* for global interaction, not the certainty of high-intensity interaction everywhere. If we were to set a high numerical threshold for τ (e.g., requiring strong influence across all regions), we would force the architecture to ”over-smooth” features, thereby destroying the local structure (edges, textures) essential for visual tasks. The condition $\tau > 0$ is the minimum necessary requirement for a model to be theoretically capable of global perception without structural blind spots. Once a model is structurally guaranteed to have $\tau > 0$, learnable weights can adaptively determine the *strength* of that influence based on data.

Our proposed definition is a binary test of architectural topology, not a test of signal strength. Its purpose is not to constrain the magnitude of gradients (which are typically learned adaptively from data), but rather to constrain the receptive field structure of the architecture itself.

Comment 2

Although the authors claim that the proposed GSSM enables efficient global image modeling, there is no related experimental data on practical inference latency or GPU memory cost. Relying only on FLOPs may not accurately reflect the module’s true efficiency, as operations like 2D-DFT can have different hardware utilization profiles than standard convolutions.

We thank you for pointing out that FLOPs do not directly reflect hardware efficiency, particularly for operators such as the DFT. To address this, we have added **Appendix K.6**, which provides a detailed analysis of the performance and efficiency of GSSM compared to other global modeling modules under practical scenarios.

All experiments were conducted on the Vaihingen dataset using a single NVIDIA RTX 4090 (24GB) GPU with an input resolution of 1024×1024, and tested on the UNet-ConvNeXt(S) backbone (Table 1). We measured the average inference latency (ms/image) and peak GPU memory usage (GB) with a batch size of 1 to provide a realistic assessment of hardware efficiency. Our proposed GSSM module achieves a significant improvement in mIoU (86.00%) while maintaining competitive latency (58.1 ms/image) and moderate GPU memory usage (10.13 GB), demonstrating a superior balance between efficiency and accuracy.

We also compared the inference performance of GSSM with other global modeling modules on a single NVIDIA RTX 4090 (24 GB) GPU across different image resolutions (Fig.1). As the resolution increases, both the inference latency and peak GPU memory usage of GSSM exhibit approximately linear growth.

These results provide strong evidence that GSSM is not only a theoretical improvement but also achieves substantial effectiveness in practical tasks.

Table 1: Ablation of GMamba with Different Global Modeling Modules

Global Module	Latency (ms / image)	GPU Memory (GB)	mIoU (%)
Self-Attention (Transformer)	68.9	12.78	84.78
VSSM (Mamba)	51.6	8.71	84.01
GSSM (GMamba, Ours)	58.1	10.13	86.00

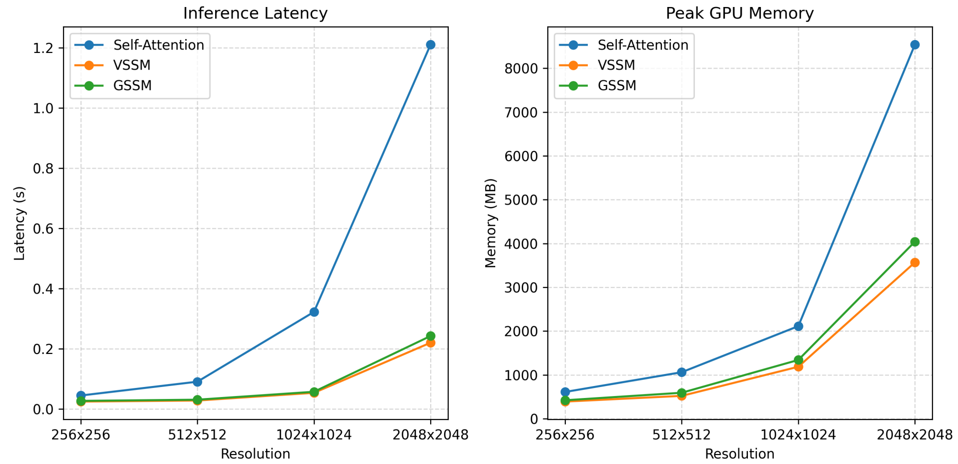

Figure 1: Performance comparison of three global modeling modules under different input resolutions.

Comment 3

The experimental validation is extensive on dense prediction tasks (segmentation and detection). However, a standard benchmark for vision backbones is ImageNet classification. The absence of this benchmark makes it slightly more difficult to assess the module's generalizability as a fundamental building block for all vision tasks.

We thank you for raising this important point. We understand that ImageNet classification serves as a standard benchmark for validating general-purpose visual backbone networks. However, we argue that for the specific scientific contribution of this paper—defining and achieving genuine global perception through frequency-domain modulation—dense prediction tasks (segmentation/detection) provide more rigorous and relevant validation than image classification. Moreover, our GMamba module is designed as a plug-and-play global modeling component, which cannot be directly evaluated under the conventional ImageNet backbone benchmarking protocol.

We did not constrain the scope to "dense prediction" in our title because the GMamba module is indeed a fundamental vision module. The extensive experiments we conducted on dense tasks, combined with fair baseline comparisons, sufficiently validate the module's generality and efficiency.

The core motivation of our work lies in addressing the lack of rigorous definition for global context modeling, which subsequently led to the design of the GMamba module to tackle the challenge of "efficiency and capability of global modeling at high resolution."

Classification is typically local: ImageNet classification is inherently object-centric. Convolutional Neural Networks (CNNs) and even Transformer models typically rely on local discriminative features (e.g., textures or specific object parts) rather than global structural understanding to solve classification tasks.

Dense prediction requires global context: In contrast, our chosen tasks—semantic segmentation and object detection—are context-centric. Distinguishing between "road" and "lane," or detecting small objects in cluttered backgrounds, necessitates understanding long-range dependencies and spatial relationships among pixels across the entire image.

We are not disregarding ImageNet; rather, we deliberately chose more challenging tasks that are more sensitive to spatial structural information as our core validation. Meanwhile, we designed GMamba as a plug-and-play module and conducted extensive experiments to demonstrate its efficiency, achieving strong performance on nearly every task, thereby validating its broad deployment capability. Since our central goal is to tackle the tension between efficiency and global modeling capability at high resolution, evaluations on high-resolution segmentation and detection tasks provide substantially stronger evidence than low-resolution (224 × 224) ImageNet classification.

> **Comment 4**
>
> Some illustrations of previous work in this paper appear to be incorrect and may mislead readers. The authors show the ViM scanning routes in Figure 1(c), but this depiction seems to be wrong (as noted in other work, e.g., [1]). This potential misrepresentation of a key baseline method is a concern.
> [1]. Visual mamba: A survey and new outlooks, 2024.

We sincerely thank you for pointing this out. We have carefully re-verified our illustration based on the original Vision Mamba (Vim) paper [1].

Our Fig. 1(c) was drawn strictly following the bidirectional SSM scanning mechanism proposed in the original Vim paper [1], where the image sequence is scanned in both forward and backward directions.

We note that the survey cited by the reviewer ([2]) may discuss various subsequent variants or generalizations of Mamba. Nevertheless, in order to ensure a fair benchmark comparison, we have faithfully adhered to the architectural design of the original Vim implementation. We hope that our illustration accurately reflects the specific benchmark (Vim) used in our experiments.

[1] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision Mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024a.

[2] Rui Xu, Shu Yang, Yihui Wang, Yu Cai, Bo Du, and Hao Chen. Visual Mamba: A Survey and New Outlooks. *arXiv preprint arXiv:2404.18861*, 2024. Available: https://arxiv.org/abs/2404.18861.