

# 000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 RESPONES-TO-REVIEWER-NC9H

## Anonymous authors

Paper under double-blind review

Dear Reviewer nc9h, 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

Limitation: The reliance on DFT-based frequency-domain modulation may introduce challenges in scenarios where frequency-domain information is less effective, such as highly noisy or irregular data.

We understand your concern that the DFT-based frequency modulation in **GSSM** may not perform optimally in certain complex scenarios. However, we argue that frequency-domain information offers unique advantages for handling such data. Since noise often manifests as high-frequency interference while semantic structures reside in low-frequency components, our GSSM can adaptively filter out irrelevant noise through its *Frequency Encoding Module* and *Frequency-Guided Modulation Module*, effectively acting as a soft frequency filter. To this end, We conducted additional robustness and cross-domain generalization experiments, and the corresponding results and analyses are provided in **Appendix K.4** of the revised manuscript.

First, we conducted controlled perturbation experiments using Gaussian noise, Gaussian blur, and partial occlusion. Our GMamba model consistently achieved the best performance across all perturbation types—improving over the baseline by 5.23% under Gaussian noise and by 5.35% under Gaussian blur. These results demonstrate that the frequency-guided modulation mechanism is not hindered by such perturbations; instead, it enhances the model’s robustness to noise. Second, to further evaluate the model’s performance under more challenging domain shifts, we performed cross-domain experiments on the LoveDA dataset (urban → rural and rural → urban). GMamba achieved the best performance in all settings, improving mIoU by an average of 3.76% and mF1 by 3.13% compared to the baseline. This indicates that GMamba not only preserves semantic cues under frequency variations but also generalizes effectively to previously unseen domains.

Moreover, the proposed GSSM exhibits inherent adaptivity in both the spatial and frequency domains. In scenarios where frequency-domain cues contribute less, our model can automatically rely more heavily on the spatial pathway, ensuring that performance does not degrade when frequency information is insufficient. This dual-domain adaptivity enables GSSM to maintain consistently strong performance across diverse datasets, without requiring manual consideration of the relative importance of spatial versus frequency cues.

054      Appendix K.4 Robustness and Cross-Domain Generalization Analysis  
 055

056      To further evaluate the robustness and cross-domain generalization capability of the GMamba model  
 057      under complex environments, we adopt UNet-ConvNeXt (S) as the baseline and apply typical input  
 058      perturbations including Gaussian noise (0.01), Gaussian blur (k=3), and partial occlusion (5%). As  
 059      shown in Table 1, GMamba consistently achieves the best performance under all perturbation set-  
 060      tings, with a 5.23% improvement under Gaussian noise and a 5.35% gain under blur, highlighting its  
 061      strong noise resistance and stability. In addition, to assess the model’s adaptability to cross-domain  
 062      scenarios, we conduct domain generalization experiments on the LoveDA dataset. As reported in  
 063      Table 2, GMamba not only enhances semantic representation but also significantly alleviates domain  
 064      shift, maintaining superior and stable segmentation performance across different domains.

065      **Table 1: Robustness Evaluation under Typical Input Perturbations**

Method	Clean (%)	Noise (0.01) (%)	Blur (k=3) (%)	Occlusion (5%) (%)
UNet-ConvNext(S)	83.11	77.55	79.15	77.00
+Swin	84.82	79.10	80.30	78.20
+SwinV2	84.36	79.25	80.45	78.35
+ViM	84.24	79.00	80.20	78.10
+VMamba	84.56	79.40	80.60	78.50
+TinyViM	84.38	79.35	80.55	78.40
+Mamba Version	84.80	79.70	81.10	78.90
+Spatial Mamba	84.50	79.45	80.85	78.60
+FreqMamba	84.60	80.10	81.50	79.20
<b>+GMamba (Ours)</b>	<b>86.00</b>	<b>82.78</b>	<b>84.50</b>	<b>81.00</b>

078      **Table 2: Cross-Domain Evaluation on LoveDA Dataset**

Method	Urban → Rural		Rural → Urban	
	mIoU (%)	mF1-score (%)	mIoU (%)	mF1-score (%)
UNet-ConVNext(S)	39.13	53.48	52.15	68.12
+Swin	41.05	55.25	54.10	70.05
+SwinV2	41.18	55.40	54.22	70.12
+TinyViM	40.50	54.90	53.70	69.80
+ViM	40.20	54.70	53.45	69.50
+VMamba	39.80	54.30	53.10	69.10
+Mamba Version	40.00	54.50	53.30	69.30
+Spatial Mamba	40.40	54.85	53.55	69.60
+FreqMamba	41.00	55.10	54.00	69.95
<b>+GMamba (Ours)</b>	<b>42.89</b>	<b>58.38</b>	<b>55.63</b>	<b>71.00</b>

095      **Comment 2**

096      Impact: While the frequency-domain approach enhances global perception, it may struggle  
 097      in cases where spatial features dominate or where frequency information is less relevant.

100      We appreciate your raising this important consideration. We fully agree that, in certain scenarios,  
 101      spatial features may be more critical than frequency information. This is precisely the motivation  
 102      behind the design of our Frequency-Guided Modulation Module (FGMM).

104      **Adaptive Modulation:** Our GSSM module does not simply “inject” frequency information. In-  
 105      stead, it employs an adaptive modulation mechanism. As shown in Equation  $X_{\text{modulated}} =$   
 106       $\mathcal{G}(X, \mathbf{F}_{\text{global}}) = \alpha_1(\mathbf{F}_{\text{global}}) \odot X + \alpha_2(\mathbf{F}_{\text{global}}) \odot \mathbf{F}_{\text{global}}$ , the model balances the contributions of  
 107      the original spatial features ( $X$ ) and the global frequency features ( $\mathbf{F}_{\text{global}}$ ) through two dynamically  
 108      generated coefficients  $\alpha_1$  and  $\alpha_2$ .

108  
**Data-Driven Learning:** These coefficients,  $\alpha_1$  and  $\alpha_2$ , are learned from the features themselves via  
 109 a small convolutional block. This means that if a task or a particular image region relies primarily  
 110 on spatial features, the model can learn to increase  $\alpha_1$  (spatial weight) and decrease  $\alpha_2$  (frequency  
 111 weight), effectively “turning off” or reducing the influence of frequency information.

112 In the original manuscript, we conducted detailed experiments and analyses on different frequency  
 113 information and integration strategies, as shown in Table 3 in **Appendix K.3**, further demonstrating  
 114 the necessity of this adaptive mechanism. Compared to a simple addition strategy (mIoU 85.53%),  
 115 our adaptive modulation approach achieves a significant improvement (mIoU 86.00%), indicating  
 116 that the model actively learns how to optimally fuse these two types of information.

117 Therefore, GSSM does not rely on frequency information blindly; rather, it possesses adaptive learning  
 118 capabilities that dynamically adjust the weights of spatial and frequency information according  
 119 to task requirements, thereby maintaining robustness across a wide range of scenarios.  
 120

123 **Table 3: Ablation Study on Frequency-domain Information**

Model Variant	mIoU (%)	mF1 (%)	OA (%)
w/o Frequency (NoFreq)	84.01	91.01	93.38
+ High-Frequency Only (HF Only)	85.20	91.83	93.74
+ Low-Frequency Only (LF Only)	85.36	91.94	93.78
+ HF + LF (Simple Addition)	85.53	92.03	93.89
+ HF + LF (Adaptive Modulation, Ours)	<b>86.00</b>	<b>92.31</b>	<b>93.99</b>

**Comment 3**

135 Although GMamba is more efficient than self-attention mechanisms, it still introduces ad-  
 136 dditional computational overhead compared to simpler SSM-based methods like Vim or  
 137 TinyViM.

142 We appreciate your attention to the computational efficiency of GMamba. Our approach is compu-  
 143 tationally competitive—even superior—compared to Vim/TinyViM.

145 To support this, we have added new ablation experiments in the main text of the revised manuscript  
 146 (see Table 4), as well as efficiency comparison experiments and analyses for GSSM in **Appendix**  
 147 **K.6** (see Table 5 and Fig. 1).

148 Compared to VSSM (Mamba), our GSSM (Ours) introduces only an additional 6.5 ms of inference  
 149 latency and 1.42 GB of GPU memory overhead, while achieving a significant improvement of 2.0%  
 150 in mIoU. Although there is a slight increase in the number of parameters and FLOPs, this cost is very  
 151 limited, demonstrating that the proposed method offers a high marginal gain and excellent trade-off  
 152 between accuracy improvement and computational cost. Fig. 1 illustrates the inference latency and  
 153 peak GPU memory of GSSM compared to other global modeling methods at different resolutions.  
 154 Our GSSM exhibits approximately linear growth in both inference latency and peak GPU memory  
 155 as the resolution increases.

156 In Tables 2, 3, and 4 of the revised manuscript, we provide a comprehensive comparison of various  
 157 global modeling modules across different tasks and backbone networks. Although GMamba intro-  
 158 duces slightly more parameters and FLOPs than Vim and TinyViM, the increase is minimal while  
 159 achieving the best performance. This demonstrates substantial marginal gains and an excellent bal-  
 160 ance between accuracy and computational cost. The results further show that GMamba remains  
 161 highly competitive in terms of computational complexity.

162  
163  
164  
165  
166  
167  
168  
169  
170  
171  
172

Table 4: Effect of Replacing Global Modeling Module in GMamba

GMamba Variant	Params (M)	FLOPs (G)	mIoU (%)	mF1 (%)	OA (%)
GMamba w/ Self-Attention	74.87	91.20	84.78	91.57	93.62
GMamba w/ VSSM (Mamba)	68.31	82.00	84.01	91.01	93.38
GMamba w/ GSSM (Ours)	71.06	85.66	86.00	92.31	93.99

173  
174  
175  
176  
177  
178

Table 5: 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

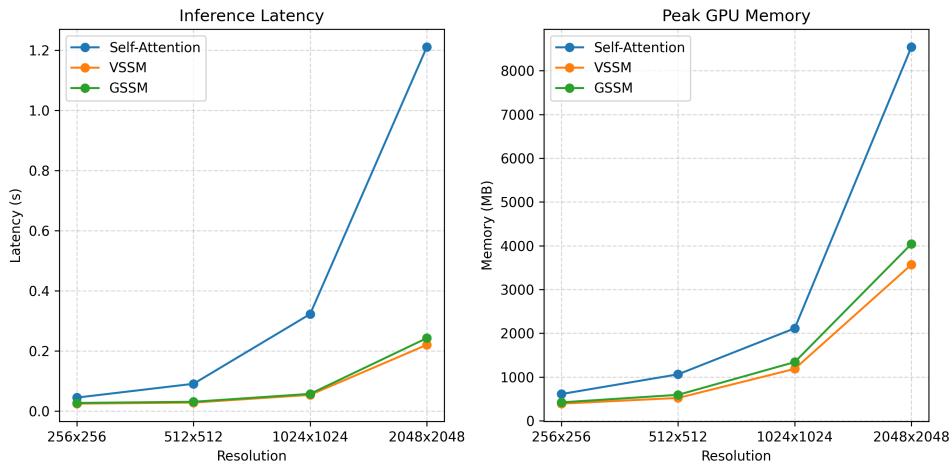
179  
180  
181  
182  
183  
184  
185  
186  
187  
188  
189  
190  
191  
192  
193  
194195  
196  
197  
198  
199  
200  
201  
202

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

**Comment 4**

While GMamba is described as "plug-and-play," its integration requires careful tuning of parameters such as modulation coefficients and frequency-domain weights. This could increase the complexity of implementation and training.

207  
208  
209  
210  
211  
212  
213

We sincerely appreciate your attention to the practical implementation and training of GMamba. The "modulation coefficients" and "frequency weights" you mentioned are not hyperparameters that require manual tuning; rather, they are automatically learned during training.

**Frequency Weights (FEM):** The low- and high-frequency component weights in FEM ( $\theta_{\text{low}}$ ,  $\theta_{\text{high}}$ ) are learnable parameters, optimized end-to-end via standard backpropagation together with the rest of the network.

214  
215

**Modulation Coefficients (FGMM):** The modulation coefficients  $a_1$  and  $a_2$  are not fixed hyperparameters. They are dynamically generated by a convolutional block (ConvBlock) whose input is the current feature map.

216 We think that this design actually reduces complexity. Users do not need to manually adjust these  
 217 weights for new tasks; the module adapts automatically. The plug-and-play nature of GMamba al-  
 218 lows it to be inserted like a standard Residual Block and fully optimized through end-to-end training  
 219 without any module-specific manual tuning. Furthermore, across all experiments, we strictly main-  
 220 tain a consistent number of training epochs, and standard training schedules are sufficient to achieve  
 221 optimal performance, introducing no additional training difficulty or convergence burden.

222 **Comment 5**

224 The robustness of GMamba in such challenging scenarios remains unclear, which could  
 225 affect its reliability in real-world applications.

227 We sincerely appreciate your concern regarding the robustness of our model, as it is directly related  
 228 to its reliability in practical applications. Thus, we have added a systematic and comprehensive  
 229 analysis of robustness and cross-domain generalization in **Appendix K.4** of the revised manuscript.

230 Furthermore, in the experimental section of the main text, we have thoroughly validated GMamba  
 231 across multiple tasks (semantic segmentation, object detection, and instance segmentation), vari-  
 232 ous datasets (remote sensing and natural scene datasets), different models, and multiple backbone  
 233 networks. Our GMamba not only demonstrates significant improvements in accuracy over exist-  
 234 ing methods but also maintains a high level of robustness and reliability under complex, noisy, and  
 235 cross-domain scenarios.

238 **Appendix K.4 Robustness and Cross-Domain Generalization Analysis**

239 To further evaluate the robustness and cross-domain generalization capability of the GMamba model  
 240 under complex environments, we adopt UNet-ConvNeXt (S) as the baseline and apply typical input  
 241 perturbations including Gaussian noise (0.01), Gaussian blur (k=3), and partial occlusion (5%). As  
 242 shown in Table 6, GMamba consistently achieves the best performance under all perturbation set-  
 243 tings, with a 5.23% improvement under Gaussian noise and a 5.35% gain under blur, highlighting its  
 244 strong noise resistance and stability. In addition, to assess the model's adaptability to cross-domain  
 245 scenarios, we conduct domain generalization experiments on the LoveDA dataset. As reported in  
 246 Table 7, GMamba not only enhances semantic representation but also significantly alleviates domain  
 247 shift, maintaining superior and stable segmentation performance across different domains.

249 **Table 6: Robustness Evaluation under Typical Input Perturbations**

Method	Clean (%)	Noise (0.01) (%)	Blur (k=3) (%)	Occlusion (5%) (%)
UNet-ConvNext(S)	83.11	77.55	79.15	77.00
+Swin	84.82	79.10	80.30	78.20
+SwinV2	84.36	79.25	80.45	78.35
+ViM	84.24	79.00	80.20	78.10
+VMamba	84.56	79.40	80.60	78.50
+TinyViM	84.38	79.35	80.55	78.40
+Mamba Version	84.80	79.70	81.10	78.90
+Spatial Mamba	84.50	79.45	80.85	78.60
+FreqMamba	84.60	80.10	81.50	79.20
<b>+GMamba (Ours)</b>	<b>86.00</b>	<b>82.78</b>	<b>84.50</b>	<b>81.00</b>

270  
271  
272

273

Table 7: Cross-Domain Evaluation on LoveDA Dataset

274

Method	Urban → Rural		Rural → Urban	
	mIoU (%)	mF1-score (%)	mIoU (%)	mF1-score (%)
UNet-ConVNext(S)	39.13	53.48	52.15	68.12
+Swin	41.05	55.25	54.10	70.05
+SwinV2	41.18	55.40	54.22	70.12
+TinyViM	40.50	54.90	53.70	69.80
+ViM	40.20	54.70	53.45	69.50
+VMamba	39.80	54.30	53.10	69.10
+Mamba Version	40.00	54.50	53.30	69.30
+Spatial Mamba	40.40	54.85	53.55	69.60
+FreqMamba	41.00	55.10	54.00	69.95
<b>+GMamba (Ours)</b>	<b>42.89</b>	<b>58.38</b>	<b>55.63</b>	<b>71.00</b>

285

286

287

288

289

290

291

292

293

294

## QUESTIONS:

295

296

297

298

**Comment 1**

299

300

301

302

The paper claims to provide the first rigorous mathematical definition of global image modeling, which is a significant contribution. However, it does not compare this definition with existing heuristic approaches in detail, leaving room for further exploration of how it improves interpretability and theoretical support.

303

304

305

306

307

We sincerely appreciate your recognition of the importance of the mathematical definition we proposed. We agree that it is crucial to compare this definition with existing heuristic methods; accordingly, we have added a detailed discussion at the end of **Section 2.1** in the main text and in **Appendix K.1** (Table 8) of the revised manuscript.

308

Our definition contains two core criteria:

310

1. **Global Gradient Dependence (Equation 1):** The norm of the gradient of the output with respect to any input pixel must be greater than a positive lower bound  $\tau$ .

311

2. **Non-Sequential Constraint:** Due to the non-causal nature of images, the model should not impose strict sequential dependencies on the input.

312

313

314

**Self-Attention:** Although self-attention can model global dependencies, its architecture does not guarantee this. Attention weights are learned dynamically, and the model may focus only on local pixels, causing the gradient with respect to certain distant pixels to approach zero ( $\tau \rightarrow 0$ ). As discussed in the revised manuscript, its globality is an “unstable, empirically observed emergent property,” rather than an architectural guarantee.

315

316

317

318

319

**Recurrent SSMs (Vanilla Mamba, ViM, etc.):** These methods suffer from structural conflicts. They rely on causal (recurrent) order to model dependencies, which violates the non-sequential constraint of images. Furthermore, as analyzed in Section 2.2.1, their influence on long-range dependencies decays exponentially, making it structurally impossible to satisfy the gradient lower bound requirement ( $\tau > 0$ ).

Through this definition, “global modeling” is transformed from a vague empirical concept into a theoretical property that can be rigorously analyzed and guaranteed during architectural design. We provide a detailed theoretical comparison in [Appendix K.1](#). In addition, in the experimental section, we conduct comprehensive comparisons between existing global modeling modules (e.g., Swin Transformer, VMamba, ViM) and our GMamba module, which is designed to satisfy this definition. The results show that our method not only provides stricter theoretical guarantees of globality but also achieves significant performance improvements, while enhancing the interpretability of module design. This further validates the necessity and soundness of the proposed theoretical definition.

Table 8: Comparison of Global Modeling Methods with Respect to Theoretical Properties

Method	Globality	Gradient Dependence	Positional Consistency	Theoretical Guarantee
Self-Attention	Full token-wise interaction	Partially satisfied (depends on attention weights)	Yes	Weak
Vanilla SSM	Recursive accumulation	No (exponential decay)	No (depends on sequence order)	Weak
VMamba	Multi-directional scanning	No (still decays)	No	Weak
DFT	Frequency-domain transform	Yes	Yes	Strong
GSSM (Ours)	DFT-guided SSM	Yes	Yes	Strong

**Comment 2**

How does the frequency-domain transfer function derived for SSMs compare to other global modeling techniques, such as attention mechanisms?

We sincerely thank you for raising this insightful theoretical question. SSMs and Attention mechanisms differ fundamentally, which is directly reflected in their “transfer function” or “filtering” characteristics.

As derived in Section 2.2.1 (Eq. 5), the transfer function of an SSM,  $H(\omega)$ , is the Fourier transform of its convolution kernel  $\bar{K}$ . This defines the SSM as a (dynamic) linear time-invariant (LTI) filter, which “filters” the input signal in the frequency domain. However, due to the recursive structure of conventional SSMs, their modeling is performed in a step-wise dynamic manner, which is an indirect mechanism based on a “causal” modeling paradigm that conflicts with the intrinsic characteristics of images. In contrast, Self-Attention is not an LTI system and thus does not possess a fixed, input-independent transfer function  $H(\omega)$ . It is a nonlinear, input-adaptive mechanism. The “filtering” behavior of Attention (i.e., attention weights) is content-based (Query-Key similarity) and dynamically computed. Effectively, it generates a unique “filter” for each output position in real time. As a result, the information obtained through this mechanism is difficult to guarantee a priori.

Our GSSM recognizes the limitations of SSMs for image modeling tasks. Our contribution is not to “fix”  $H(\omega)$  itself; rather, we leverage the theoretically established frequency-domain modulation framework for SSMs. After demonstrating the global properties of the DFT, we further employ DFT-based pre-modulation of the input  $u_t$ . This step injects global frequency-domain information before it enters the SSM, enabling the efficient SSM to process image information under the guidance of global context, thereby addressing the limitations of causal processing inherent in conventional SSMs for image tasks, and ultimately designing a more efficient and more interpretable global modeling module.

In the revised manuscript, we have added ablation experiments in the main text (Table 9) and in [Appendix K.6](#) (Table 10) that compare the performance of GSSM with VSSM and self-attention mechanisms. These results further demonstrate experimentally that our design outperforms the other two methods in both performance and efficiency.

Table 9: Effect of Replacing Global Modeling Module in GMamba

GMamba Variant	Params (M)	FLOPs (G)	mIoU (%)	mF1 (%)	OA (%)
GMamba w/ Self-Attention	74.87	91.20	84.78	91.57	93.62
GMamba w/ VSSM (Mamba)	68.31	82.00	84.01	91.01	93.38
GMamba w/ GSSM (Ours)	71.06	85.66	86.00	92.31	93.99

378  
379  
380  
381  
382  
383  
384  
385  
386  
387  
388  
389  
390  
391  
392  
393  
394  
395  
396  
397  
398  
399  
400  
401  
402  
403  
404  
405  
406  
407  
408  
409  
410  
411  
412  
413  
414  
415  
416  
417  
418  
419  
420  
421  
422  
423  
424  
425  
426  
427  
428  
429  
430  
431

Table 10: 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

**Comment 3**

How does GMamba’s linear-logarithmic complexity compare to the linear complexity of other SSM-based methods like Vim, VMamba or spatial mamba?

We appreciate your attention to the computational complexity of GMamba. Theoretically, GMamba has a complexity of  $\mathcal{O}(M \log M)$ , which differs from the  $\mathcal{O}(M)$  complexity of Vim/VMamba. However, in practice, our method is highly competitive in terms of computation.

In our experiments, the  $\mathcal{O}(M \log M)$  term is implemented using highly optimized FFT (DFT) algorithms, which incur very low computational overhead and run extremely fast. Moreover, since our proposed GSSM performs only a single scan and does not employ multi-directional scanning or other complex strategies, the overall computational complexity remains efficient.

This is further validated in the comparative experiments presented in Tables 2, 3, and 4 of the main text: on a UNet-ResNet34 backbone, GMamba requires 36.30G FLOPs, which is comparable to ViM (35.81G) and Spatial Mamba (35.72G), introducing almost no additional overhead.

Although the theoretical complexity is “linear-logarithmic,” thanks to the high efficiency of FFT, GMamba achieves practical computational performance (in terms of FLOPs and latency) comparable to, or even better than, “linear” methods such as Vim and VMamba.