# Supplementary Material

In this section, we further include more results and analysis to complement the main paper. We provide additional details on the following topics:

- Architectural Details (Sec. 7)
- Ablations (Sec. 8)
- Qualitative Results (Sec. 9)
- Discussion (Sec. 10)
- Limitations (Sec. 11)

## 7. Architectural Details

We develop three variants of our GroupMamba backbones, each tailored to different performance and efficiency requirements: GroupMamba-T (Tiny), GroupMamba-S (Small), and GroupMamba-B (Base), with 23M, 34M, and 57M parameters, respectively. These variants differ in their channel dimensions and the number of layers per stage, as detailed in Tab. 6.

## 8. Ablations

In Tab. 3, we provide additional ablation results regarding the distillation training objective. For the GroupMamba-T and GroupMamba-S variants, the distilled loss improves performance by an absolute gain of 0.8% and 0.9%, respectively. For the largest variant, GroupMamba-B, the distilled loss improves performance by 1.3%. This demonstrates that larger Mamba-based models with MLP tend to saturate and struggle to converge effectively without distillation. Incorporating distillation for the large model boosts its performance from 83.2% to 84.5%.

We also visualize the training loss curves with and without our proposed distilled loss for GroupMamba-S in Fig. 5. The shaded areas indicate the standard deviation of loss across the training epochs. As shown, incorporating the distilled loss (green curve) consistently leads to lower training losses and less loss variability throughout the training process, leading to improved stability.

We compare in Tab. 4 the performance of different scanning directions with respect to the number of groups for GroupMamba-T. In the first row, we use Direction 1. In the second row, we use Direction 1 and Direction 2. In the last row, we use the four scanning scanning directions (As visualized in Fig. 2 (d). Four groups with four directions capture richer spatial cues, which provide comprehensive feature representation and lead to higher top-1 accuracy with comparable throughput.

We also conduct an ablation study to evaluate efficiency with varying numbers of groups. While utilizing two groups reduces parameters by 15% and four groups achieves a reduction of 26%, employing eight groups yields only a marginally greater reduction of 28% due to the nonlinear scaling of MLP parameters. In addition, using eight groups (with eight scanning directions) decreases throughput, negatively impacting model efficiency. Hence, four groups have the optimal trade-off between parameter reduction and high throughput.

In Tab. 5, we present an additional ablation study and a fair comparison between GroupMamba-T and VMamba-T without distillation alongside another variant of GroupMamba-T designed to match the parameter count of VMamba-T for balanced evaluation. Remarkably, GroupMamba-T achieves equivalent performance to VMamba-T with 26% fewer parameters. When parameter counts are matched, the enhanced variant, GroupMamba-T$^{\dagger}$, outperforms VMamba-T, achieving a top-1 accuracy of 83.1% on ImageNet-1K, compared to 82.5% for VMamba-T, without using any distillation.

| Method | #Param. | FLOPs | Top-1 acc. |
|---|---|---|---|
| GroupMamba-T w/o Distilled Loss | 23M | 4.6G | 82.5 |
| GroupMamba-T with Distilled Loss | 23M | 4.6G | 83.3 (+0.8) |
| GroupMamba-S w/o Distilled Loss | 34M | 7.0G | 83.0 |
| GroupMamba-S with Distilled Loss | 34M | 7.0G | 83.9 (+0.9) |
| GroupMamba-B w/o Distilled Loss | 57M | 14G | 83.2 |
| GroupMamba-B with Distilled Loss | 57M | 14G | 84.5 (+1.3) |

Table 3. Ablation study on GroupMamba variants with and without the Distilled Loss.

| Scanning Directions | Throughput (ms) ↑ | #Param ↓ | Top-1 (%) ↑ |
|---|---|---|---|
| D1 | 1096 | 23M | 82.9 |
| D1, D2 | 1087 | 23M | 83.1 |
| **D1, D2, D3, D4** | **1069** | **23M** | **83.3** |

Table 4. Comparison of different scanning directions in terms of throughput, parameters, and top-1 accuracy for GroupMamba-T.

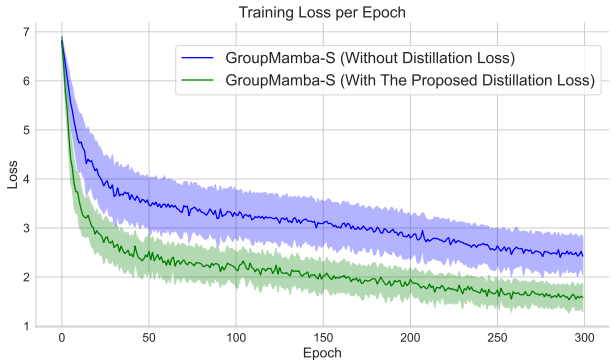


Figure 5. Training loss visualization for GroupMamba-S with and without the proposed distilled loss.

| Method | #Param. | FLOPs | Top-1 acc. |
|---|---|---|---|
| VMamba-T | 31M | 4.9G | 82.50% |
| GroupMamba-T | 23M | 4.5G | 82.50% |
| GroupMamba-T$^{\dagger}$ | 31M | 5.2G | 83.10% |

Table 5. Comparison of VMamba-T and GroupMamba-T without distillation. The number of channels is increased in GroupMamba-T$^{\dagger}$ to match the same parameters of VMamba-T

## 9. Qualitative Results

In Fig. 6, we show additional qualitative results of GroupMamba-T on samples from the ADE20K [69] validation set for semantic segmentation. The first row shows the ground truth masks, while the second row displays the predicted masks. Our model consistently has sharp and accurate delineations, effectively capturing fine details and complex object boundaries, further emphasizing its robustness in semantic segmentation. Similarly, we present in Fig. 7 additional qualitative results of GroupMamba-T on samples from the COCO validation set [33], showcasing its strong performance in both instance segmentation and object detection tasks. The model excels at accurately localizing objects and producing precise segmentations, even in complex scenes with varying scales, multiple instances, and challenging backgrounds. The quantitative and qualitative results of GroupMamba demonstrate the robust generalization capability of our GroupMamba backbones across diverse downstream tasks, including semantic segmentation, object detection, and instance segmentation.

## 10. Discussion

Our main contributions include introducing the Modulated Group Mamba layer, which enhances computational efficiency and interaction in state-space models through a multi-direction scanning method. We also introduce the Channel Affinity Modulation (CAM) operator to improve feature aggregation across channels, addressing limitations in grouping operations. Additionally, we employ a distillation-based training objective to stabilize the training of models with a large number of parameters. These contributions enable us to achieve competitive performance with recent state-space models in image classification, object detection, instance segmentation, and semantic segmentation with fewer number of parameters.

This can further facilitate the development of vision foundation models based on Mamba that can be scaled to a large number of parameters efficiently and stably. The Modulated Group Mamba layer and CAM operator enhance computational efficiency and feature interaction, allowing models to manage more extensive and complex datasets without excessive resource demands. The distillation-based training objective ensures stability during training, which is crucial for maintaining performance as model sizes increase. Together, these advancements enable the creation of scalable, reliable vision models that can be deployed effectively in various real-world applications.

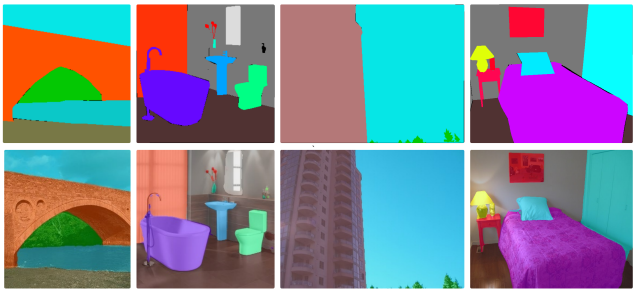

Figure 6. Qualitative results of GroupMamba-T for semantic segmentation on ADE20K validation set. The first row shows the ground truth for the masks, while the second and second show the corresponding predictions of our model.

## 11. Limitations

Despite demonstrating clear improvements in efficiency, stability, and accuracy for image classification tasks and fewer parameters for dense prediction tasks, our proposed Modulated Group Mamba layer shows relatively comparable performance on downstream tasks such as object detection and segmentation to VMamba. This minor improvement can be attributed to the more complex nature and diverse requirements of these dense prediction tasks, where the accuracy relies heavily not only on effective global dependency capture but also on more localized spatial feature aggregation and specialized detection or segmentation heads. The proposed model architecture enhances parameter efficiency and global feature modeling through SSM mechanisms, but addressing the intricacies inherent to localization-sensitive tasks may require additional targeted modules or task-specific optimizations.

Although the incorporation of knowledge distillation has successfully improved training stability and yielded performance gain for large-scale models, investigating more efficient or self-guided stabilization approaches would help enhance the model training practicality without requiring auxiliary external teacher models.

| Stage | Resolution | Type | Config | GroupMamba | | |
|---|---|---|---|---|---|---|
| | | | | T | S | B |
| Stem | $\frac{H}{2} \times \frac{W}{2}$ | Patch Embedding | Patch Size | k=3x3, s=2 | | |
| | | | Embed. Dim. | 32 | 64 | 64 |
| | $\frac{H}{4} \times \frac{W}{4}$ | Patch Embedding | Patch Size | k=3x3, s=2 | | |
| | | | Embed. Dim. | 96 | 96 | 128 |
| 1 | $\frac{H}{4} \times \frac{W}{4}$ | Modulated Group Mamba | Layers | 2 | 2 | 2 |
| | $\frac{H}{8} \times \frac{W}{8}$ | Down-Sampling | Patch Size | k=3x3, s=2 | | |
| | | | Embed. Dim. | 192 | 192 | 256 |
| 2 | $\frac{H}{8} \times \frac{W}{8}$ | Modulated Group Mamba | Layers | 2 | 2 | 2 |
| | $\frac{H}{16} \times \frac{W}{16}$ | Down-Sampling | Patch Size | k=3x3, s=2 | | |
| | | | Embed. Dim. | 368 | 384 | 496 |
| 3 | $\frac{H}{16} \times \frac{W}{16}$ | Modulated Group Mamba | Layers | 9 | 20 | 20 |
| | $\frac{H}{32} \times \frac{W}{32}$ | Down-Sampling | Patch Size | k=3x3, s=2 | | |
| | | | Embed. Dim. | 760 | 768 | 1012 |
| 4 | $\frac{H}{32} \times \frac{W}{32}$ | Modulated Group Mamba | Layers | 2 | 2 | 2 |
| Parameters<br>FLOPs | | | | 23M<br>4.5G | 34M<br>7.0G | 57M<br>14.0G |

Table 6. GroupMamba architectures. Description of the configurations of the model variants for the embedding size, the number of layers, and the model's GFLOPs and Parameters. Between two consecutive stages, we incorporate a downsampling layer to increase the number of channels and reduce the resolution by two.


Figure 7. Qualitative results of GroupMamba-T for object detection and instance segmentation on the COCO validation set.