# SUPPLEMENTARY MATERIAL: WD-MAMBA: A W-MAMBA DIFFUSION MODEL FOR ENHANCING IMAGE SYNTHESIS

*Lakshmikar Reddy Polamreddy*[1] *Jialu Li*[2] *Youshan Zhang*[3]

[1]PhD in Mathematics, Yeshiva University, New York, USA
[2]MS in Artificial Intelligence, Yeshiva University, New York, USA
[3]Department of Computer Science and Engineering, Yeshiva University, New York, USA
yz945@cornell.edu

## 1. TRAINING ALGORITHM

The following algorithm describes how our WD-Mamba works to predict noise effectively.

**Algorithm 1** Image Generation with WD-Mamba Model

```
Input: Training images $\{x_i\}$, noisy image $y$, true noise $\epsilon_i$
Output: Predicted noise $\epsilon_{\text{pred}}$, Generated image
for each training iteration do
    Sample a batch of images $\{x_i\}$
    Generate noisy patches $P$ from $x_i$
    Compute input embedding as shown in Eq. (3):
        $x \leftarrow \textsc{Embed}(P, T, C)$
    for each stage in {Encoder, Bottleneck, Decoder} do
        for each Mamba block do
            $x' \leftarrow \textsc{MambaBlock}(x)$
        end for
    end for
    Obtain W-Net output: $\hat{y}$
    Predict noise: $\epsilon_{\text{pred}} \leftarrow \text{Conv2D}(\hat{y})$
    Compute synchronization loss as shown in Eq. (11):
        $\mathcal{L}_{\text{sync}} = \alpha \mathcal{L}_{\text{mag}}(\epsilon_i, \epsilon_{\text{pred}}) + \beta \mathcal{L}_{\text{dir}}(\epsilon_i, \epsilon_{\text{pred}})$
    Update model parameters
end for
Sampler substracts Predicted noise from noisy image:
Generated image $\leftarrow (y - \epsilon_{\text{pred}})$
```

## 2. GFLOPS AND WALL-CLOCK SPEED COMPARISON.

Table 1 presents a detailed comparison of compute and efficiency metrics for generating 256×256 images. All evaluations are conducted on $2\times$ A100 GPUs with a batch size of 128. Notably, DiT-XL/2 requires 3 GPUs due to its high memory consumption (196 GB), making it less efficient for limited hardware environments. We essentially compare our WD-Mamba model with U-ViT-H/2 and DiT-XL/2, both of which utilize flash attention mechanisms. Our WD-Mamba maintains lower GFLOPs and fewer parameters compared to these ViT and DiT models. Moreover, WD-Mamba achieves faster training and inference speeds compared to DiT, offering a strong balance between efficiency and performance.

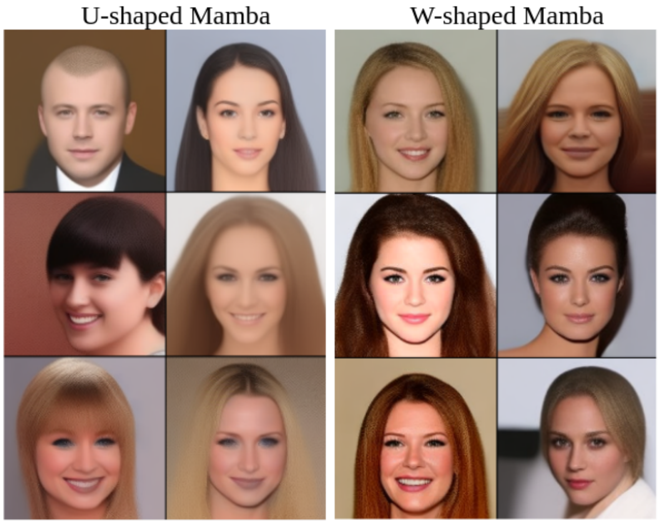

**Fig. 1.** Comparison of U-shaped Mamba and W-shaped Mamba generated images on Celeba-HQ dataset after training for 100K iterations.

## 3. ABLATION STUDIES

### 3.1. Why W-shape instead of U-shape?

We adopt the WD-Mamba model within a W-Net architecture, which consistently outperforms the traditional U-Net design due to its distinctive W-shaped structure. This architecture introduces additional intermediate stages between the encoder and decoder, enabling the model to retain and refine richer feature representations across multiple levels of abstraction. Furthermore, the multiple intermediate connections in

| Model | Params (M) | GPU Memory (GB) | GFLOPs | Training Speed (s/iter) | Inference Speed (s/1000 samples) |
|---|---|---|---|---|---|
| U-ViT-H/2 | 501 | 39.5 | 133 | 2.5 | 510 |
| DiT-XL/2 | 675 | 196 | 119 | 3.9 | 980 |
| DiM-Huge | 860 | - | 210 | - | - |
| DiffuSSM-XL | 660 | - | 280 | - | - |
| LDM-4 | 400 | - | 104 | - | - |
| WD-Mamba | 462 | 67 | 106 | 2.9 | 750 |

**Table 1**. Comparison of different models based on model parameters, GPU memory, GFLOPs, training, and inference speeds.

**Table 2**. Comparison of scan patterns performance on FFHQ dataset

| Scan Pattern | FID |
|---|---|
| Sweep [1] | 8.23 |
| Zigzag [1] | 6.17 |
| Spiral | **3.61** |

the W-shaped design facilitate improved gradient flow during backpropagation, resulting in more stable training and faster convergence.

Experimentally, we evaluate the impact of U-shaped and W-shaped Mamba on image quality using the CelebA-HQ dataset. Fig. 1 shows that while U-shaped Mamba generates visually acceptable images, it struggles with details like hair texture, facial expressions, and lighting. In contrast, W-shaped Mamba produces higher-resolution images with sharper textures, more accurate facial features, and better lighting, demonstrating superior detail capture. Quantitative results further confirm this, with W-shaped Mamba achieving a lower FID score of 3.52 compared to 7.19 for U-shaped Mamba, indicating closer alignment with real image distributions.

### 3.1.1. *why Spiral Scan Patterns?*

We conduct an ablation study with existing scan patterns—sweep and zigzag—and compare the results with the proposed spiral scan pattern. We trained the WD-Mamba model using each scan pattern on the FFHQ dataset and evaluated the Fréchet Inception Distance (FID) on 50K generated images. As shown in Table 2, the spiral scan pattern achieves the lowest FID, indicating superior image quality compared to the other patterns. So, we introduce this spiral pattern in our work to enhance the overall image synthesis performance.

## 3.2. Why do we design a new loss function?

While Mean of Squares (MoS) loss is effective, it overlooks the alignment of orientations, leading to slower convergence. To overcome this, we propose synchronization loss, which incorporates both magnitude and directional components. Results from WD-Mamba generated images on the FFHQ and CIFAR10 datasets demonstrate that synchronization loss consistently achieves better FID scores and faster convergence, as shown in Fig. 2. At 60K iterations on FFHQ, synchronization loss achieves an FID of 18.3, while MoS loss results in 20.6. After 100K iterations, synchronization loss improves it to 3.6, while MoS loss to 4.8. On CIFAR10, at 120K iterations, synchronization loss has an FID of 14.6, compared to 15.1 with MoS loss. After 200K iterations, synchronization loss achieves 2.6 compared to 3.1 of MoS loss.

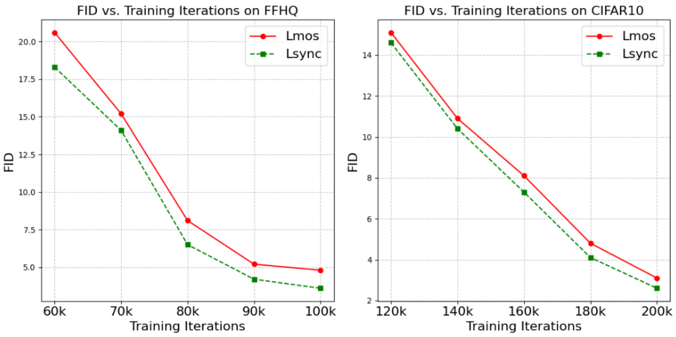

**Fig. 2**. FID scores of WD-Mamba generated images showing the superior performance of synchronization loss (Lsync) over MoS loss (Lmos).

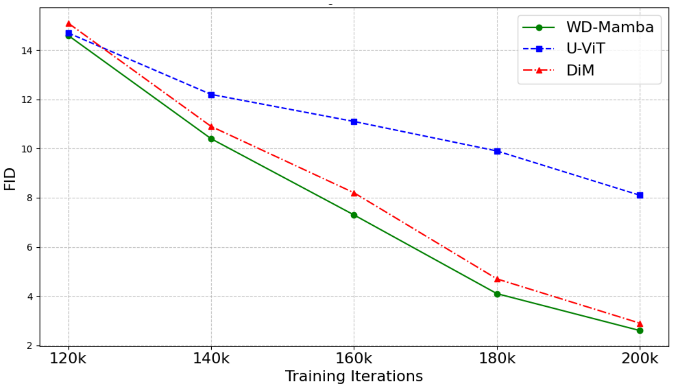

**Fig. 3**. Comparison of FID versus training iterations on CIFAR10 dataset.

## 4. COMPARISON OF CONVERGENCE SPEED

We compare WD-Mamba, U-ViT, and DiM based on their FID scores over training iterations on the CIFAR10 dataset [2].

As training progresses from 120K to 200K iterations, WD-Mamba consistently achieves lower FID values, reaching 2.6 at 200K iterations, outperforming DiM (2.9) and significantly surpassing U-ViT (8.1). These results demonstrate the superior convergence speed and image quality of WD-Mamba compared to the other models. Notably, while the reported FID of U-ViT in its original paper [3] is 2.9, this was achieved only after 500K training iterations.

## 5. MORE GENERATED IMAGES

Fig. 4, Fig. 5, Fig. 6, Fig. 7, and Fig. 8 present additional images generated by the WD-Mamba model.

## 6. REFERENCES

[1] Vincent Tao Hu, Stefan Andreas Baumann, Ming Gui, Olga Grebenkova, Pingchuan Ma, Johannes Fischer, and Bjorn Ommer, "Zigma: Zigzag mamba diffusion model," *arXiv preprint arXiv:2403.13802*, 2024.

[2] Alex Krizhevsky, Geoffrey Hinton, et al., "Learning multiple layers of features from tiny images," 2009.

[3] Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu, "All are worth words: A vit backbone for diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 22669–22679.

[4] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[5] Tero Karras, Samuli Laine, and Timo Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[6] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao, "Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop," *arXiv preprint arXiv:1506.03365*, 2015.

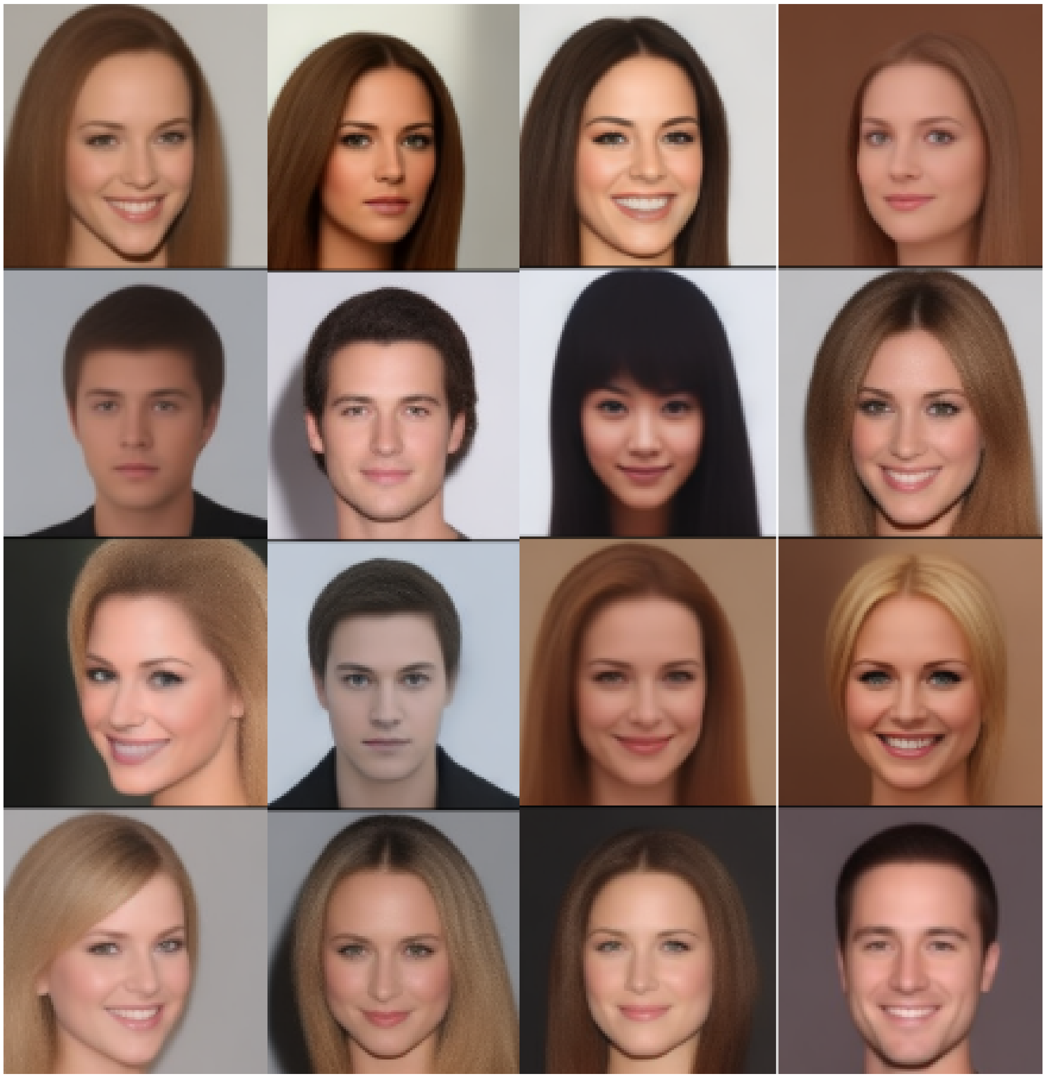

**Fig. 4**. WD-Mamba generated images of 256x256 resolution on CelebA-HQ [4] dataset.

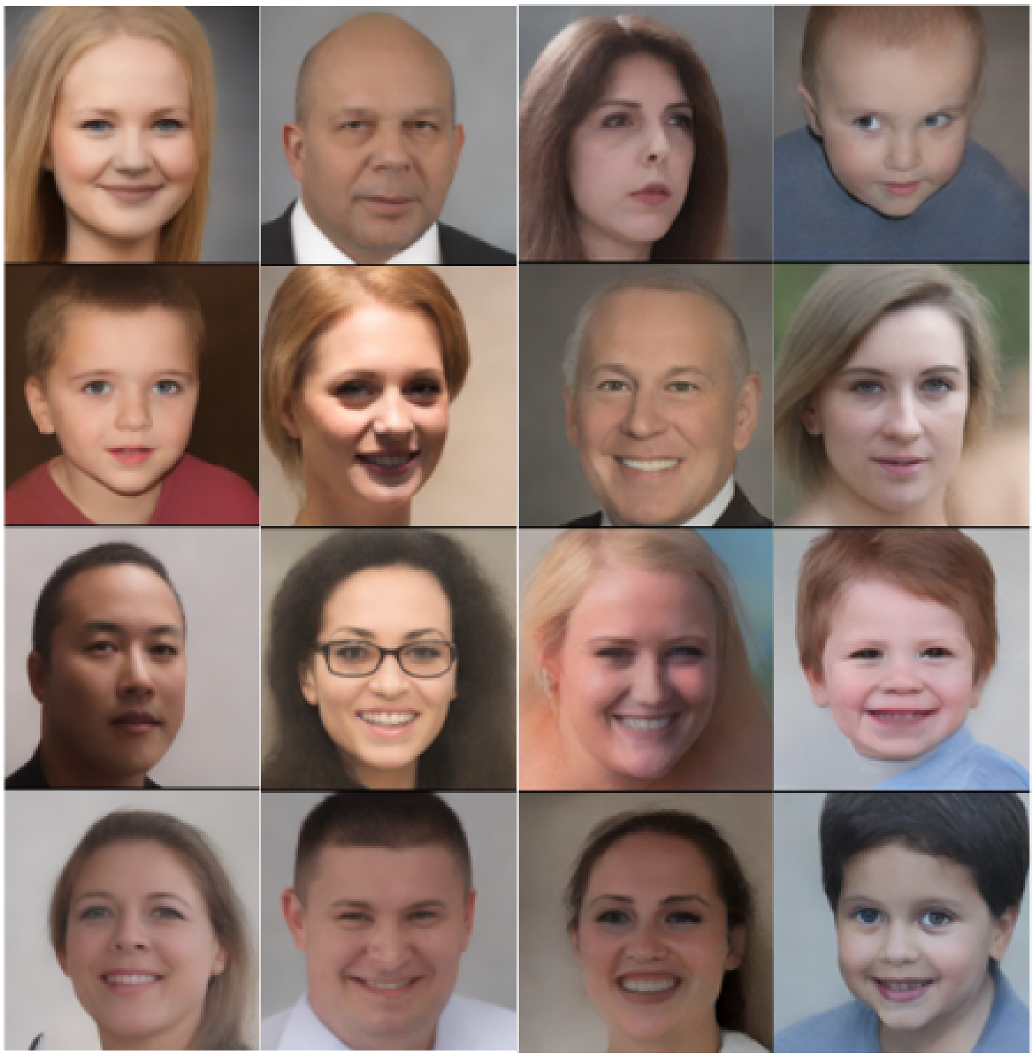

**Fig. 5**. WD-Mamba generated images of 256x256 resolution on FFHQ [5] dataset.

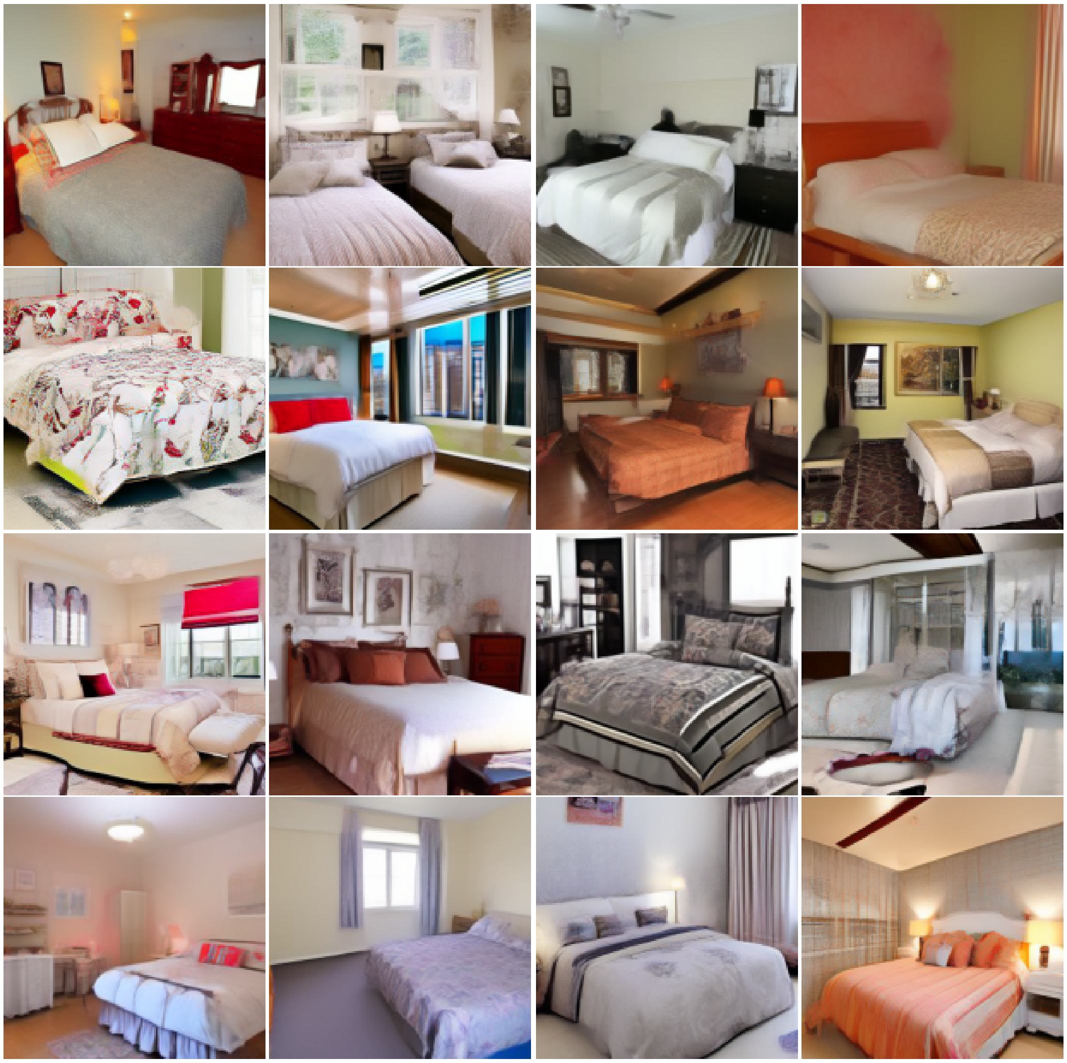

**Fig. 6**. WD-Mamba generated images of 256x256 resolution on LSUN Bedrooms [6] dataset.

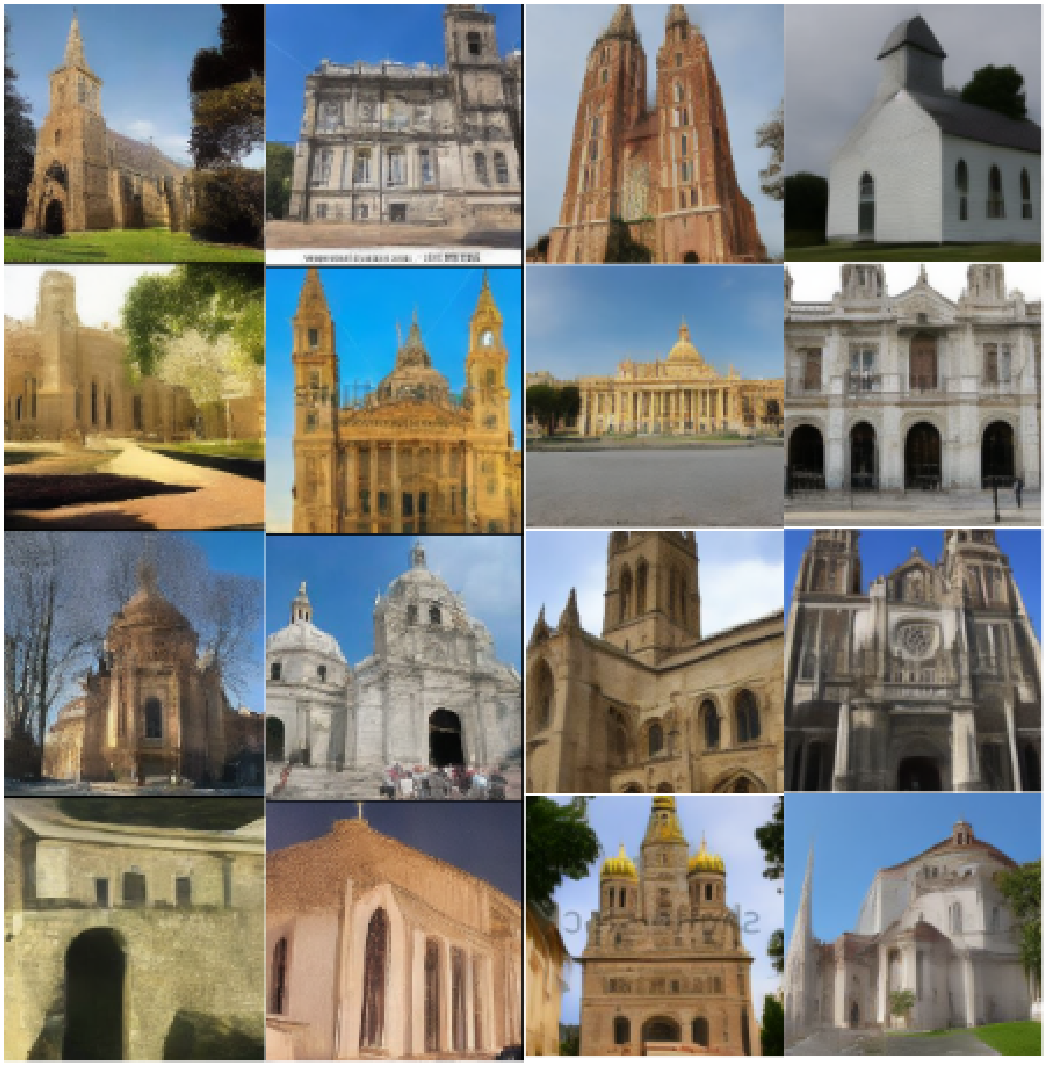

**Fig. 7**. WD-Mamba generated images of 256x256 resolution on LSUN Churches [6] dataset.

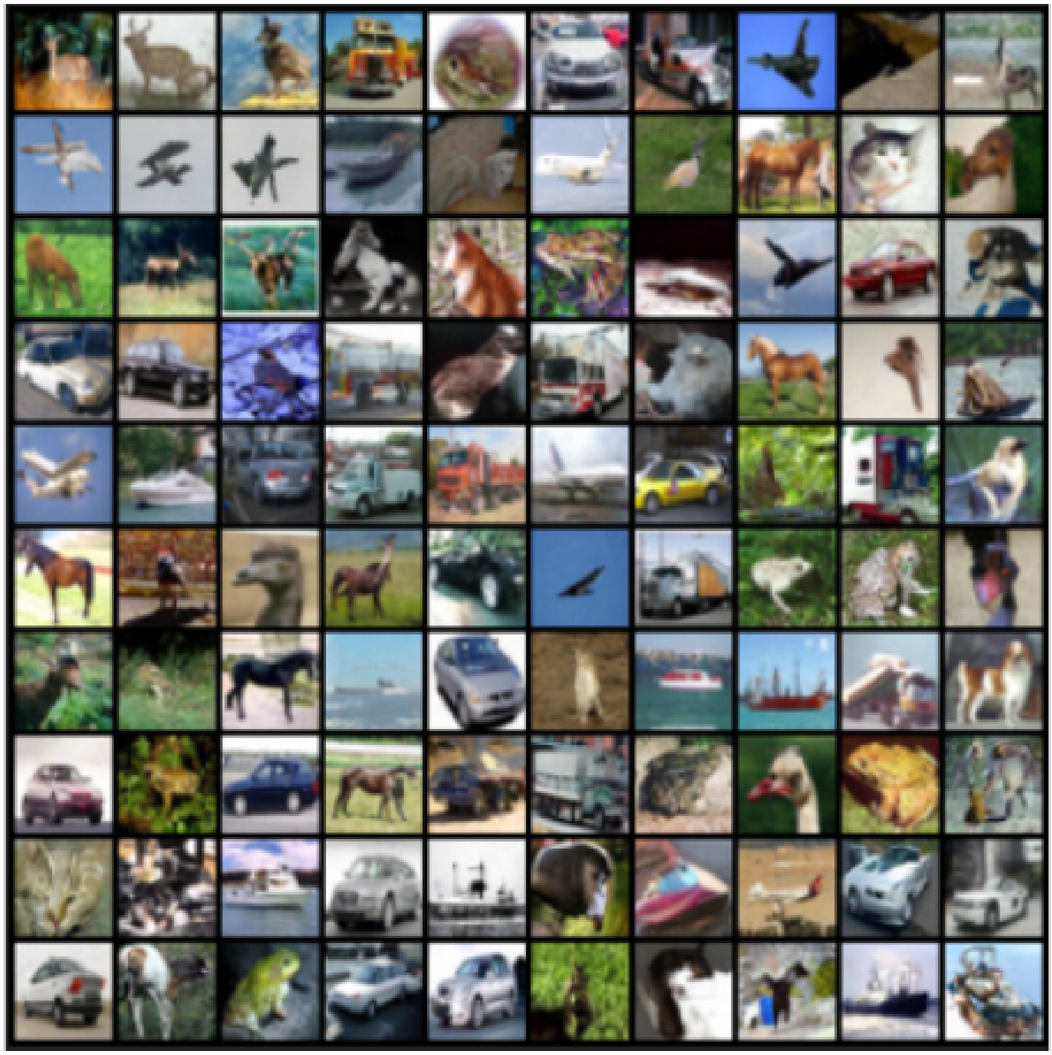

**Fig. 8**. WD-Mamba generated images of 32x32 resolution on CIFAR10 [2] dataset.