WD-Mamba: A W-Mamba Diffusion Model for Enhancing Image Synthesis

Anonymous CVPR submission

Paper ID *****

Abstract

State Space Models (SSMs) have recently been integrated into diffusion models to enhance efficiency, yet challenges remain in balancing expressiveness with computational demands, particularly for high-resolution image synthesis. In this work, we introduce WD-Mamba, a novel W-Net-based diffusion model that leverages the structured efficiency of Mamba blocks within a W-shaped (W-Net) architecture. By extending Mamba to the W-Net configuration, W-Mamba captures both local and global dependencies across image scales, enhancing feature representation and spatial continuity. Key design innovations include bi-directional state space transitions, a hierarchical downsampling and upsampling pathway, and adaptive skip connections that reinforce multi-scale feature synthesis. Experimental results show that W-Mamba achieves competitive FID scores on standard image generation benchmarks while substantially reducing computational costs. This model bridges the gap between efficient SSM-based architectures and expressive diffusion models, offering a scalable solution for high-quality, high-resolution image synthesis. The source code is available at https://github.com/LakshmikarPolamreddy/Mamba4Diffusion.

1. Introduction

In recent years, diffusion models have emerged as a powerful tool for high-quality image generation, offering notable advantages over generative adversarial networks (GANs) in generating realistic and detailed images across various domains. The backbone architectures of these models have evolved significantly, moving from the widely-adopted U-Net structure to transformer-based models. In contrast to convolutional neural networks (CNNs), the key advantage of ViTs lies in their ability to provide each image patch with context that is dependent on both the data and the specific patch, thanks to the self-attention mechanism. Another notable strength of ViTs is their modality-agnostic nature, as they treat an image as a sequence of patches without relying on 2D inductive biases, making them a preferred choice for multimodal tasks. However, despite the remarkable capacity of transformers to model complex data structures, their quadratic complexity in terms of token interactions poses a significant challenge, especially for high-resolution image synthesis. This computational demand has driven a search for alternative backbones that offer both efficiency and expressive power.

State Space Models (SSMs) have recently garnered attention as an efficient solution to high-resolution generation tasks. Originally designed for sequence modeling, SSM-based backbones, such as Mamba, provide a promising alternative due to their linear scaling with the number of tokens, addressing the limitations of transformers in handling high-dimensional data. Mamba, in particular, has demonstrated considerable success across various modalities, including language, audio, and genomics, by efficiently capturing long-range dependencies. However, adapting SSM-based architectures to image generation presents unique challenges. Images, unlike sequential data, require two-dimensional spatial continuity, which is difficult to maintain in a sequential model like Mamba. Previous approaches have experimented with adapting 1-D SSMs for 2-D image data, but often encounter limitations in preserving spatial coherence across both local and global image regions.

In this paper, we introduce WD-Mamba, a versatile and efficient network designed to enhance generative modeling in image synthesis tasks. Inspired by the ViT and Mamba, WD-Mamba integrates a hybrid structure that combines the strengths of both convolutional and self-attention mechanisms to effectively capture localized fine-grained features and long-range dependencies. This innovation distinguishes WD-Mamba from traditional models, offering superior performance with linear scaling in feature size, unlike the quadratic complexity commonly encountered in Transformer-based models. Furthermore, WD-Mamba's adaptability allows it to seamlessly integrate with different datasets, making it highly scalable and flexible across a range of image generation tasks. Through extensive experiments on the CelebA dataset [9], we demonstrate that WD-Mamba not only achieves state-of-the-art results, as shown in Fig. 1 but also provides a robust framework for future ad-

vancements in generative models, paving the way for more efficient and scalable solutions in image synthesis.

2. Related Work

Mamba Models. Mamba models have emerged as a highly efficient alternative to traditional attention-based architectures like Transformers. These models aim to balance computational efficiency with expressive power by employing structured state space models (SSMs) as a foundational component. Unlike Transformers, which rely on computationally expensive attention mechanisms, Mamba models leverage the inherent capabilities of SSMs to capture long-range dependencies while maintaining a lightweight and efficient architecture. This innovative approach reduces inference time, making Mamba models a competitive choice for tasks requiring real-time or high-resolution processing. The development of Mamba models has seen contributions from various researchers proposing specialized adaptations. Gu et al. [6] introduced the original Mamba model by integrating SSMs into a simple end-to-end neural network architecture, eliminating the need for attention mechanisms and even MLP blocks. This design prioritizes speed and computational efficiency, positioning Mamba as a faster alternative to traditional Transformers. Building on this foundation, Zhu et al. [18] proposed Vision Mamba (ViM), a generic vision backbone utilizing bidirectional Mamba blocks. ViM uses position embeddings for image sequences and compresses visual representations through bidirectional SSMs, further enhancing the architecture's ability to capture spatial and temporal dependencies. Despite its advantages, Mamba models initially lacked mechanisms to effectively account for spatial continuity. To address this, Hu et al. [8] introduced Zigzag Mamba, a plug-and-play solution with minimal parameter overhead. This approach employs a zigzag scanning scheme, outperforming baseline Mamba models in tasks that require stronger spatial continuity. Additionally, Fu et al. [5] introduced Local Attentional Mamba (LaMamba) blocks, which integrate self-attention mechanisms with Mamba blocks to achieve linear complexity while effectively capturing both global context and local details. In the medical imaging domain, researchers have proposed several Mamba variants to address segmentation challenges. Ma et al. [13] presented U-Mamba, a hybrid network combining convolutional layers with SSMs. This design captures both local features and long-range dependencies, making it ideal for biomedical image segmentation. Liu et al. [11] extended this concept with Swin-UMamba, leveraging ImageNet-based pretraining for enhanced segmentation performance. The versatility of Mamba models extends to text-to-image generation as well. Fei et al.[4] introduced Dimba, a hybrid text-to-image diffusion model combining Transformer and Mamba elements. Dimba leverages the strengths of both architectures for more effective text-to-image synthesis. Similarly, Teng et al. [16] developed Diffusion Mamba (DiM), which integrates the efficiency of Mamba with the expressive capabilities of diffusion models. DiM enables efficient high-resolution image synthesis, merging the advantages of SSMs and diffusion frameworks.

Diffusion Models. Recent advancements in diffusion models have led to significant improvements in image synthesis quality, efficiency, and flexibility. Ho et al. [7] Ho et al. (2020) introduced Denoising Diffusion Probabilistic Models (DDPMs), a class of latent variable models inspired by non-equilibrium thermodynamics, which generate high-quality images by reversing a process that gradually adds noise to data. This groundbreaking approach allowed for the generation of diverse and high-fidelity images, setting the stage for further developments. Building on this foundation, Dhariwal et al. [3] introduced classifier guidance, a technique that enhances image generation by incorporating gradients from a pre-trained classifier during the denoising process. This method provides a balance between diversity (variability in generated images) and fidelity (how accurately the images represent the target distribution) in a computationally efficient manner, offering more control over the generated samples. To address the challenge of computational resource limitations, Rombach et al. [15] proposed Latent Diffusion Models (LDMs), which apply the diffusion process in a compressed latent space. This reduces the computational burden while still maintaining high-quality outputs, making it possible to train these models on less powerful hardware and scale them more effectively. In terms of architectural advancements, Bao et al. [1] introduced U-ViT, a Vision Transformer-based model specifically designed for diffusion tasks. This architecture offers a simpler, yet more flexible framework compared to traditional convolutional networks, taking advantage of the powerful attention mechanisms in transformers to capture long-range dependencies in the data. Peebles et al. [14] explored the potential of Diffusion Transformers by replacing the conventional U-Net backbone with transformer-based architectures. This shift highlights the growing importance of transformer models in diffusion frameworks, offering improved scalability and performance for image generation tasks. Furthermore, Tian et al. [17] introduced U-DiTs, a model that optimizes the transformer's self-attention mechanism by applying token downsampling to the query-key-value tuple, significantly reducing computational costs while maintaining performance. Together, these contributions represent rapid progress in diffusion models, each approach enhancing the overall efficiency or quality of image synthesis in innovative ways, further solidifying diffusion models as a state-of-the-art method for generative image modeling.

While Mamba blocks and traditional diffusion models each possess distinct strengths—Mamba blocks offering

Figure 1. WD-Mamba generated images on CelebA dataset [9] after training for 100K iterations.

computational efficiency and structured feature representation, and diffusion models excelling in generating high-quality outputs—their integration has been relatively underexplored, with only a few studies addressing this intersection. [4, 16]. To bridge this gap, we introduce WD-Mamba, a novel model that unites the strengths of Mamba blocks and Transformer architectures. WD-Mamba combines the structured efficiency of Mamba blocks with the contextual modeling capabilities of Transformers, aiming to enhance both image synthesis quality and computational efficiency. By leveraging the complementary benefits of these approaches, WD-Mamba aspires to advance performance in image generation tasks significantly.

3. Methods

3.1. Preliminaries

State Space Models (SSMs) are sequence-to-sequence architectures that map an input sequence $x(t) \in \mathbb{R}$ to an output sequence $y(t) \in \mathbb{R}$ via a hidden state $h(t) \in \mathbb{R}^N$. The continuous-time formulation of a linear time-invariant (LTI) SSM is expressed as:

$$h'(t) = Ah(t) + Bx(t), \quad y(t) = Ch(t) + Dx(t), \quad (1)$$

where $A \in \mathbb{R}^{N\times N}$ is the state transition matrix, $B \in \mathbb{R}^{N\times 1}$ represents the input matrix, $C \in \mathbb{R}^{1\times N}$ is the output matrix, and $D \in \mathbb{R}$ provides a direct path from the input to the output. Here, N denotes the dimension of the hidden state.

In practical applications, data is typically discrete rather than continuous, necessitating the discretization of the continuous-time model. A common approach for this is the Zero-Order Hold (ZOH) method, which discretizes the model as follows:

$$h_t = \overline{A}h_{t-1} + \overline{B}x_t, \quad y_t = Ch_t + Dx_t, \quad (2)$$

where

$$\overline{A} = e^{\Delta A} \quad \text{and} \quad \overline{B} = (\Delta A)^{-1}\left(e^{\Delta A} - I\right)\Delta B.$$

Here, Δ is a timescale parameter controlling the discretization interval, and I is the identity matrix.

Recent advancements, such as Mamba [6], have enhanced the flexibility of SSMs by introducing the Selective Scan mechanism (S6). This mechanism allows dynamic context selection based on input sequences, thereby relaxing the time-invariant constraints on parameters B, C, and Δ. Specifically, S6 enables these parameters to become time-dependent, capturing richer contextual information from input tokens. Given an input $x_t \in \mathbb{R}^D$ with D channels, the S6 mechanism dynamically selects context for each channel, producing a hidden state $h \in \mathbb{R}^N$ for an SSM state dimension N.

Our proposed WD-Mamba model builds upon the strengths of Mamba by leveraging its ability to capture global contextual information efficiently in linear time. Additionally, WD-Mamba addresses the limitation of losing fine-grained details by incorporating local attention mechanisms. This combination ensures that the model maintains

both global coherence and local precision in high-resolution image synthesis tasks.

3.2. WD-Mamba Model Architecture

Our WD-Mamba architecture, as shown in Fig. 2, consists of a series of interconnected encoders and decoders that progressively extract, refine, and enhance features, ensuring that crucial information flows throughout the network. This structure enables effective feature extraction across multiple levels of abstraction, contributing to the high quality of the generated output.

Input and Embedding: The process begins with the input image being divided into smaller patches, capturing essential spatial information from local regions of the image. Each patch undergoes an embedding transformation, where spatial and positional details are encoded to create a feature-rich representation of each patch. This embedding stage is critical as it prepares the input for deeper processing and encodes a foundational spatial understanding necessary for the downstream encoding and decoding steps.

Mamba Encoders: The architecture includes two Mamba encoders, each consisting of multiple transformation blocks that progressively refine the embedded patches. The first encoder receives the embedded patches and processes them through a series of transformations that extract hierarchical features from different spatial and semantic levels. This encoder creates a rich set of intermediate features, representing low-level to mid-level patterns, which are essential for accurate image synthesis. To preserve essential details across the network and prevent information loss, skip connections link each Mamba encoder to the decoders. These connections allow intermediate outputs from earlier layers to bypass certain transformations and connect directly with the decoders, providing a direct path for detailed feature information. This helps maintain fine-grained information that may otherwise be diminished through deeper processing. The first encoder's output is passed both to the first decoder and simultaneously to the second encoder. The second encoder continues feature extraction, processing the refined information to produce more abstract representations. By stacking two encoders with skip connections, the architecture captures a broader range of feature hierarchies, supporting both local and global structure comprehension.

Mamba Decoders: The decoding phase follows a two-stage approach, with each stage progressively refining and reconstructing the feature map toward the final output. The first decoder receives the output from the first encoder, incorporating intermediate features through skip connections. By reintroducing this detailed information, the decoder enhances the reconstruction quality and maintains fine spatial details. The first decoder refines the feature map through multiple transformations, producing an initial decoded representation that is then sent to the second encoder for further abstraction and detail preservation. The second decoder takes the output of the second encoder, which contains deeply abstracted feature representations. By incorporating both refined features from the second encoder and direct information from skip connections, the second decoder further enhances the feature map. This dual input provides the decoder with a comprehensive understanding of both fine and coarse details, enabling a more accurate final reconstruction. Once the final decoding is complete, the model passes the feature map through a concluding transformation layer, which is responsible for generating the final output. In the context of diffusion models, this output represents the predicted noise, which is crucial for guiding the synthesis process and refining the generated image. The model's ability to predict noise accurately directly impacts the quality and realism of the generated samples, as it fine-tunes the details and enhances the final image quality. The predicted noise is further utilized in sampling to generate images.

The Mamba block, which processes embedding features as input, is designed to refine data representations through a series of sequential transformations. It begins with a linear layer that applies a transformation to the input features, followed by a Structured State Space Model (SSM) component, which captures temporal or sequential dependencies within the data. Next, a Conv 1D layer captures local dependencies across the sequence or patch dimension, enhancing spatial feature extraction. Two additional linear layers follow, further refining the transformed features, and a final normalization layer stabilizes the feature values, ensuring consistent scaling and improving the training process. By stacking multiple Mamba blocks within the Mamba Encoder and Decoder, the model progressively enhances the data representation, which ultimately aids in predicting noise effectively. Algorithm. 1 describes how our WD-Mamba works to predict noise effectively.

Memory Efficiency: The Mamba block in the WD-Mamba architecture significantly enhances memory efficiency by leveraging recomputation and modular design. Each block processes only a subset of tokens at a time, reducing the intermediate storage requirements compared to transformer-based models that often operate on the entire token set at once. Furthermore, the state-space model (SSM) within the Mamba block allows for efficient long-range dependency modeling without requiring explicit token-to-token interactions, which are memory-intensive. By replacing global self-attention with localized operations like 1D convolutions and linear layers, the Mamba block minimizes memory IO operations, particularly in high-resolution tasks, where saving memory is critical. This design is especially advantageous for tasks involving high-dimensional inputs, such as images, as it enables processing

on devices with limited GPU memory without compromising performance. The memory efficiency of the Mamba block in the WD-Mamba architecture is primarily determined by the reduced reliance on global operations (e.g., self-attention) and the use of recomputation. The memory usage equation is:

$$\text{Memory Usage (M)} = O(N \times D) \tag{3}$$

where:

- N = Number of tokens (sequence length or patches from the input image).
- D = Embedding dimension.

This is more efficient compared to transformer models, where memory scales quadratically ($O(N^2 \times D)$) due to self-attention mechanisms.

Computational Efficiency: From a computational perspective, the Mamba block reduces the overall complexity by optimizing operations. The linear layers and 1D convolutions in the block scale efficiently with the number of tokens and embedding dimensions, avoiding the quadratic growth of complexity typically associated with self-attention mechanisms. The SSM further enhances efficiency by processing sequences bidirectionally in a streamlined manner, capturing spatial dependencies with fewer operations. As a result, the WD-Mamba architecture achieves a balance between computational speed and expressive power, allowing for faster inference and training times compared to traditional transformer-based architectures. This efficiency makes the architecture well-suited for resource-constrained environments or applications requiring rapid model execution.

The computational complexity for the Mamba block comes from the linear transformations, state-space models (SSM), and convolution operations. The total complexity is given by:

$$\text{Computational Complexity (C)} = O(N \times D^2 + N \times K) \tag{4}$$

where, K = Kernel size of the 1D convolution.

Key contributors to this complexity are:

- $O(N \times D^2)$: Complexity for the linear transformations in the block.
- $O(N \times K)$: Complexity for the 1D convolution operation.

These equations highlight the Mamba block's efficiency by avoiding the $O(N^2 \times D)$ complexity of self-attention, while retaining the capacity for long-range dependency modeling through SSM and convolution.

4. Results and Discussion

4.1. Datasets

In our experiments, we utilize an unconditional dataset, CelebAHQ [9], and a class-conditional dataset, ImageNet [2].

Algorithm 1 Noise Prediction with WD-Mamba Model

Input: Noisy Image Latent I
Output: Prediction $\hat{y}$

Patch Extraction:
Divide I into patches and combine with Time token and Class token ($L + T + C$).

Embedding:
$x \leftarrow \text{Conv2d}(I)$

Normalization:
$x' \leftarrow \text{RMSNorm}(x)$

Mamba Block Iteration:
For each stage (First Encoder, First Bottleneck, First Decoder, Second Encoder, Second Bottleneck, Second Decoder):
  For each Mamba block:
    $x' \leftarrow \text{Norm}(x')$
    $x' \leftarrow \text{Linear}(x')$
    $z \leftarrow \text{Conv1D}(x')$
    $z \leftarrow \text{SSM}(x')$
    $z \leftarrow \text{Linear}(z)$
    $x' \leftarrow \text{RMSNorm}(z)$
  End for each Mamba block.
Prediction:
$\hat{y} \leftarrow \text{Conv2d}(x')$

Return $\hat{y}$

4.2. Implementation details

We trained WD-Mamba model for 100K iterations with a batch size of 128, using 2 A100 GPUs. We utilized the same DDPM [7] scheduler, a pre-trained image autoencoder [10], and DPM-Solver [12]. The learning rate is set to 2×10^{-4}, and we employed the Adam optimizer along with EMA (Exponential Moving Average) with a decay rate of 0.9999.

4.3. Results

Qualitative Analysis. Fig. 3 provides a visual comparison of images generated by three different models—U-ViT, DiM, and WD-Mamba (proposed)—after 100K training iterations on the CelebA dataset [9]. The WD-Mamba model distinctly outperforms the other two models, producing images with finer details, improved facial features, and more realistic textures. This enhanced performance highlights WD-Mamba's ability to capture subtle details in the dataset, making it highly effective for applications demanding high-

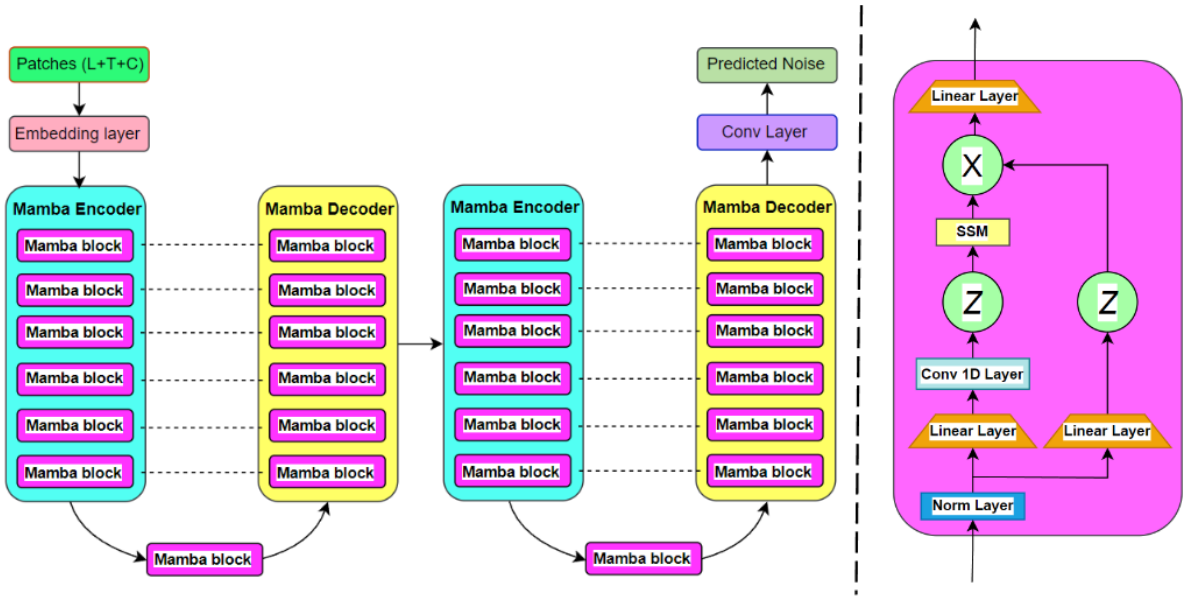

Figure 2. Our WD-Mamba architecture (left) consists of two Mamba encoders and two Mamba decoders, each containing six Mamba blocks. After the first and second encoders, the extracted features are refined and compressed through dedicated Mamba bottleneck layers. These refined features are then further processed in the corresponding decoders, with skip connections linking the encoder and decoder blocks. The final output is passed through a convolutional layer to predict the noise for image generation. Mamba block architecture is shown on the right side, where Z refers to activation function and X to concatenation.

fidelity image generation. The visual evidence underscores the robustness and adaptability of WD-Mamba for generating realistic images with precision.

Quantitative Analysis. The superiority of the WD-Mamba model is further supported by its Fréchet Inception Distance (FID) scores, as presented in Table 1. Despite having fewer parameters compared to the DiM model, WD-Mamba achieves a significantly lower FID score, indicating better alignment between the generated and real image distributions. This efficiency is particularly noteworthy, demonstrating that WD-Mamba not only excels in quality but also in computational efficiency. The quantitative and qualitative results together establish WD-Mamba as a powerful tool for fine-grained image generation tasks.

Table 1. Comparison of FID metric achieved by state-of-the-art methods on CelebA [9] dataset.

Model	# Parameters	# Iterations	FID
DiM	862M	100K	65.66
U-ViT	44M	100K	87.51
WD-Mamba (Ours)	463M	100K	68.59

4.4. Ablation Study

To evaluate the impact of different architectural configurations on image generation quality, we perform an ablation study using three variations of the Mamba-based architecture: U-Mamba, HalfU-Mamba, and WD-Mamba. Each of these architectures incorporates Mamba blocks as their foundational components, but they differ in layer depth and parameter optimization. The experiments are conducted on the CelebA dataset over 100,000 training iterations to ensure a fair comparison.

The HalfU-Mamba architecture, a reduced version with significantly fewer layers, fails to converge effectively during training. The images it generates are predominantly noisy, lacking any recognizable features. This result highlights the importance of sufficient network depth for capturing the complex distributions in the dataset. In contrast, the U-Mamba architecture, composed of 13 layers, produces images that are recognizable and considerably better than those generated by HalfU-Mamba. However, the quality of these images falls short compared to those generated by the WD-Mamba model. While U-Mamba can replicate basic structural elements, the finer details and overall realism remain suboptimal. Quantitatively, the FID metric further substantiates these findings. U-Mamba achieves an FID score of 77.19, indicating a moderate level of alignment between the distributions of generated and real images. On the other hand, WD-Mamba, with its optimized architecture, achieves a significantly better FID score of 68.59, showcasing its superior ability to generate high-quality images with

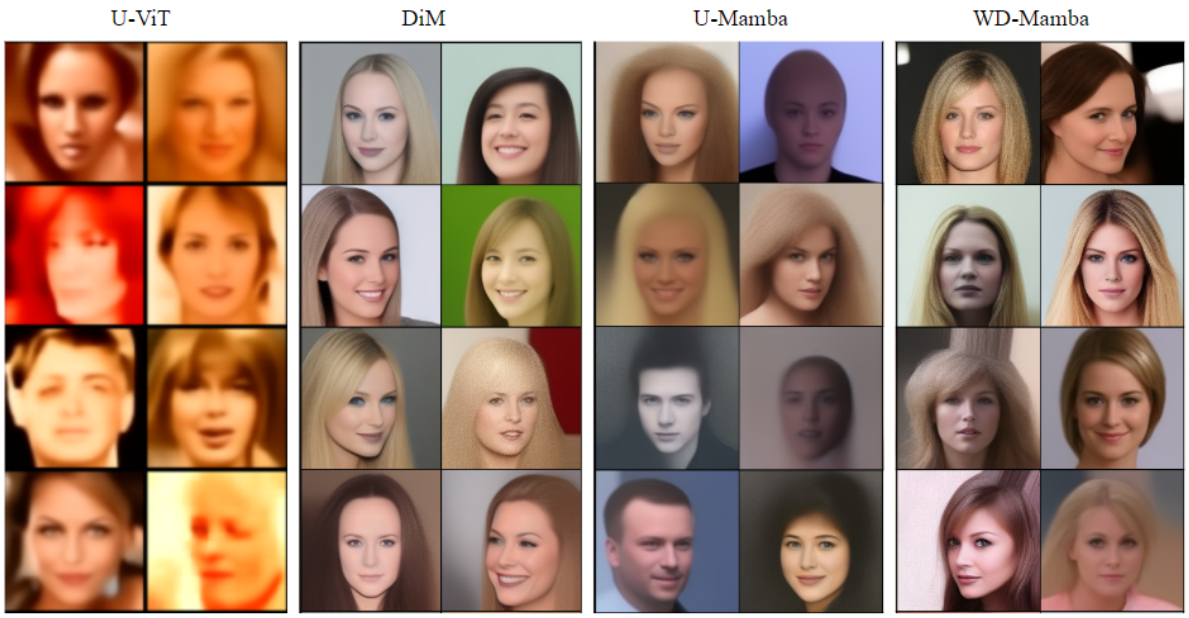

Figure 3. Comparison of Generated images on CelebA dataset [9] by four different models after training for 100K iterations.

finer details.

In future work, we will explore different configurations of WD-Mamba with respect to the number of layers, and hence with different numbers of training parameters. Based on this, we will categorize them into three sub-architectures: WD-Mamba Small, WD-Mamba Medium, and WD-Mamba Large. After assessing their performance qualitatively and quantitatively, with respect to the FID metric and computational cost, we will select the best configuration for optimal image generation quality and efficiency.

5. Conclusion

In this work, we introduce WD-Mamba, a novel W-Net-based diffusion model that combines the efficiency of Mamba blocks with the expressive power of self-attention to improve high-resolution image generation. WD-Mamba effectively captures both local and global dependencies through hierarchical downsampling and upsampling pathway, and adaptive skip connections. Our experiments on the CelebA dataset show that WD-Mamba outperforms existing models, achieving competitive FID scores while significantly reducing computational costs. In addition, WD-Mamba bridges the gap between efficient state space models and expressive diffusion models, providing a scalable and flexible solution for high-quality image synthesis.

References

[1] 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*, pages 22669–22679, 2023. 2

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[3] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[4] Zhengcong Fei, Mingyuan Fan, Changqian Yu, Debang Li, Youqiang Zhang, and Junshi Huang. Dimba: Transformer-mamba diffusion models. *arXiv preprint arXiv:2406.01159*, 2024. 2, 3

[5] Yunxiang Fu, Chaoqi Chen, and Yizhou Yu. Lamamba-diff: Linear-time high-fidelity diffusion models based on local attention and mamba. *arXiv preprint arXiv:2408.02615*, 2024. 2

[6] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 2, 3

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 5

[8] 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

[9] 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. 1, 3, 5, 6, 7

[10] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 5

[11] Jiarun Liu, Hao Yang, Hong-Yu Zhou, Yan Xi, Lequan Yu, Cheng Li, Yong Liang, Guangming Shi, Yizhou Yu, Shaoting Zhang, et al. Swin-umamba: Mamba-based unet with imagenet-based pretraining. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 615–625. Springer, 2024. 2

[12] C Lu, Y Zhou, F Bao, J Chen, and C Li. A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Proc. Adv. Neural Inf. Process. Syst., New Orleans, United States*, pages 1–31, 2022. 5

[13] Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024. 2

[14] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 2

[15] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[16] Yao Teng, Yue Wu, Han Shi, Xuefei Ning, Guohao Dai, Yu Wang, Zhenguo Li, and Xihui Liu. Dim: Diffusion mamba for efficient high-resolution image synthesis. *arXiv preprint arXiv:2405.14224*, 2024. 2, 3

[17] Yuchuan Tian, Zhijun Tu, Hanting Chen, Jie Hu, Chao Xu, and Yunhe Wang. U-dits: Downsample tokens in u-shaped diffusion transformers. *arXiv preprint arXiv:2405.02730*, 2024. 2

[18] 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*, 2024. 2