ZigMa: A DiT-style Zigzag Mamba Diffusion Model

Vincent Tao Hu Stefan Andreas Baumann Ming Gui
Olga Grebenkova Pingchuan Ma Johannes Fischer Björn Ommer
CompVis @ LMU Munich, MCML
https://taohu.me/zigma

Abstract

The diffusion model has long been plagued by scalability and quadratic complexity issues, especially within transformer-based structures. In this study, we aim to leverage the long sequence modeling capability of a State-Space Model called Mamba to extend its applicability to visual data generation. Firstly, we identify a critical oversight in most current Mamba-based vision methods, namely the lack of consideration for spatial continuity in the scan scheme of Mamba. Secondly, building upon this insight, we introduce Zigzag Mamba, a simple, plug-and-play, minimal-parameter burden, DiT style solution, which outperforms Mamba-based baselines and demonstrates improved speed and memory utilization compared to transformer-based baselines. Lastly, we integrate Zigzag Mamba with the Stochastic Interpolant framework to investigate the scalability of the model on large-resolution visual datasets, such as FacesHQ 1024×1024 and UCF101, MultiModal-CelebA-HQ, and MS COCO 256×256. Long version is at https://taohu.me/zigma/.

Figure 1: **Motivation.** Our Zigzag Mamba method improves the network's position-awareness by arranging and rearranging the scan path of Mamba in a heuristic manner.

1. Introduction

Diffusion models have demonstrated significant advancements across various applications, including image processing (Rombach et al., 2022), video analysis (Ho et al., 2022), point cloud processing (Wu et al., 2023), and human pose estimation (Gong et al., 2023). Many of these models are built upon Latent Diffusion Models (LDM)(Rombach et al., 2022), which are typically based on the UNet backbone. However, scalability remains a significant challenge in LDMs(Huang et al., 2024). Recently, transformer-based structures have gained popularity due to their scalability (Peebles & Xie, 2022; Bao et al., 2023a) and effectiveness in multi-modal training (Bao et al., 2023b). Notably, the transformer-based structure DiT (Peebles & Xie, 2022) has even contributed to enhancing the high-fidelity video generation model SORA (OpenAI, 2024) by OpenAI. Despite efforts to alleviate the quadratic complexity of the attention mechanism through techniques such as windowing (Liu et al., 2021), sliding (Beltagy et al., 2020), sparsification (Child et al., 2019; Kitaev et al., 2020), hashing (Choromanski et al., 2020; Sun et al., 2021), Ring Attention (Liu et al., 2023a; Brandon et al., 2023), Flash Attention (Dao et al., 2022) or a combination of them (Ao et al., 2024; zhuzilin, 2024), it remains a bottleneck for diffusion models.

On the other hand, State-Space Models (Gu et al., 2021a; Gupta et al., 2022; Gu et al., 2022) have demonstrated significant potential for long sequence modeling, rivaling transformer-based methods. Their biological similarity (Tikochinski et al., 2024) and efficient memory state also advocate for the use of the State-Space model over the transformer. Several methods (Gu & Dao, 2023; Gu et al., 2021a; Fu et al., 2022; Smith et al., 2022) have been proposed to

enhance the robustness (Yu et al., 2023), scalability (Gu & Dao, 2023), and efficiency (Gu et al., 2021a;b) of State-Space Models. Among these, a method called Mamba (Gu & Dao, 2023) aims to alleviate these issues through work-efficient parallel scanning and other data-dependent innovations. However, the advantage of Mamba lies in 1D sequence modeling, and extending it to 2D images is a challenging question. Previous works (Zhu et al., 2024; Liu et al., 2024b) have proposed flattening 2D tokens directly by computer hierarchy such as row-and-column-major order, but this approach neglects *Spatial Continuity*, as shown in Figure 1. Other works (Liu et al., 2024a; Ma et al., 2024a) consider various directions in a single Mamba block, but this introduces additional parameters and GPU memory burden. In this paper, we aim to emphasize the importance of *Spatial Continuity* in Mamba and propose several intuitive and simple methods to enable the application of Mamba blocks to 2D images by incorporating continuity-based inductive biases in images. We also generalize these methods to 3D with spatial-temporal factorization on 3D sequence.

In the end, Stochastic Interpolant (Albergo et al., 2023) provides a more generalized framework that can uniform various generative models including, Normalizing Flow (Chen et al., 2018), diffusion model (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021), Flow matching (Lipman et al., 2023; Liu et al., 2023b; Albergo & Vanden-Eijnden, 2022), and Schrödinger Bridge (Liu et al., 2022). Previously, some works (Ma et al., 2024b) explore the Stochastic Interpolant on relatively small resolutions, e.g., 256×256, 512×512. In this work, we aim to explore it in further more complex scenarios e.g., 1024×1024 resolution and even in videos.

In summary, our contributions are as follows: Firstly, we identify the critical issue of *Spatial Continuity* in generalizing the Mamba block from 1D sequence modeling to 2D image and 3D video modeling. Building on this insight, we propose a simple, plug-and-play, zero-parameter paradigm named *Zigzag Mamba (ZigMa)* that leverages spatial continuity to maximally incorporate the inductive bias from visual data. Secondly, we extend the methodology from 2D to 3D by factorizing the spatial and temporal sequences to optimize performance. Secondly, we provide comprehensive analysis surrounding the Mamba block within the regime of diffusion models. Lastly, we demonstrate that our designed *Zigzag Mamba* outperforms related Mamba-based baselines, representing the first exploration of Stochastic Interpolants on large-scale image data (1024×1024) and videos.

2. Method

In this section, we begin by providing background information on State-Space Models (Gu et al., 2021a; Gupta et al., 2022; Gu et al., 2022), with a particular focus on a special case known as Mamba (Gu & Dao, 2023). We then highlight the critical issue of *Spatial Continuity* within the Mamba framework, and based on this insight, we propose the Zigzag Mamba. This enhancement aims to improve the efficiency of 2D data modeling by incorporating the continuity inductive bias inherent in 2D data. Furthermore, we design a basic cross-attention block upon Mamba block to achieve text-conditioning. Subsequently, we suggest extending this approach to 3D video data by factorizing the model into spatial and temporal dimensions, thereby facilitating the modeling process. Finally, we introduce the theoretical aspects of stochastic interpolation for training and sampling, which underpin our network architecture.

2.1. Diffusion Backbone: Zigzag Mamba

Zigzag Scanning in Mamba. Previous studies (Wang et al., 2022; Yan et al., 2023) have used bidirectional scanning within the SSM framework. This approach has been expanded to include additional scanning directions (Liu et al., 2024a;b; Yang et al., 2024b) to account for the characteristics of 2D image data. These approaches unfold image patches along four directions, resulting in four distinct sequences. Each of these sequences is subsequently processed together through every SSM. However, since each direction may have different SSM parameters ($\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$), scaling up the number of directions could potentially lead to memory issues. In this work, we investigate the potential for amortizing the complexity of the Mamba into each layer of the network.

Our approach centers around the concept of token rearrangement before feeding them into the Forward Scan block. For a given input feature $\mathbf{z}_i$ from layer i, the output feature $\mathbf{z}_{i+1}$ of the Forward Scan block after the rearrangement can be expressed as:

$$\mathbf{z}_{\Omega_i} = \text{arrange}(\mathbf{z}_i, \Omega_i), \tag{1}$$

$$\bar{\mathbf{z}}_{\Omega_i} = \text{scan}(\mathbf{z}_{\Omega_i}), \tag{2}$$

$$\mathbf{z}_{i+1} = \text{arrange}(\bar{\mathbf{z}}_{\Omega_i}, \bar{\Omega}_i), \tag{3}$$

Ω_i represents the 1D permutation of layer i, which rearranges the order of the patch tokens by Ω_i, and Ω_i and $\overline{\Omega}_i$ represent the reverse operation. This ensures that both $\mathbf{z}_i$ and $\mathbf{z}_{i+1}$ maintain the sample order of the original image tokens.

Now we explore the design of the Ω_i operation, considering additional inductive biases from 2D images. We propose one key properties: *Spatial Continuity*. Regarding Spatial Continuity, current innovations of Mamba in images (Zhu et al., 2024; Liu et al., 2024b;a) often squeeze 2D patch tokens directly following the computer hierarchy, such as row-and-column-major order. However, this approach may not be optimal for incorporating the inductive bias with neighboring tokens, as illustrated in Figure 3. To address

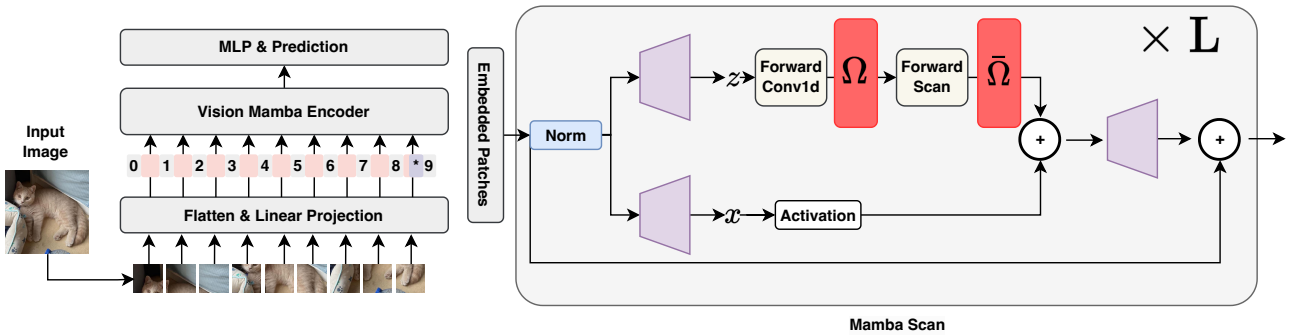


Figure 2: **ZigMa.** Our backbone is structured in L layers, mirroring the style of DiT (Peebles & Xie, 2022). We use the single-scan Mamba block as the primary reasoning module across different patches. To ensure the network is positionally aware, we've designed an arrange-rearrange scheme based on the single-scan Mamba. Different layers follow pairs of unique rearrange operation Ω and reverse rearrange $\bar{\Omega}$, optimizing the position-awareness of the method.

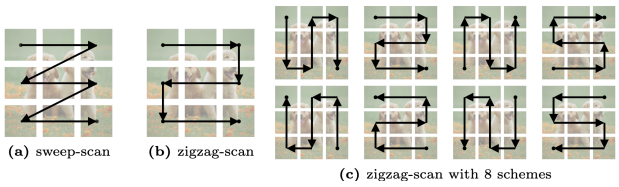


Figure 3: **The 2D Image Scan.** Our mamba scan design is based on the sweep-scan scheme shown in subfigure (a). From this, we developed a zigzag-scan scheme displayed in subfigure (b) to enhance the continuity of the patches, thereby maximizing the potential of the Mamba block. Since there are several possible arrangements for these continuous scans, we have listed the eight most common zigzag-scans in subfigure (c).

this, we introduce a novel scanning scheme designed to maintain spatial continuity during the scan process. Additionally, we consider space-filling, which entails that for a patch of size $N \times N$, the length of the 1D continuous scanning scheme should be N^2. This helps to efficiently incorporate tokens to maximize the potential of long sequence modeling within the Mamba block.

To achieve the aforementioned property, we heuristically design eight possible space-filling continuous schemes[1], denoted as $\mathbf{S}_j$ (where $j \in [0, 7]$), as illustrated in Figure 3. While there may be other conceivable schemes, for simplicity, we limit our usage to these eight. Consequently, the scheme for each layer can be represented as $\Omega_i = \mathbf{S}_{\{i\%8\}}$, where $\%$ denotes the modulo operator.

[1]We also experimented with more complex continuous space-filling paths, such as the Hilbert space-filling curve (McKenna, 2019). However, empirical findings indicate that this approach may lead to deteriorated results. For further detailed comparisons, please refer to the Appendix.

Table 1: **Ablation of Scanning Scheme Number**. We evaluate various zigzag scanning schemes. Starting from a simple "Sweep" baseline, we consistently observe improvements as more schemes are implemented.

	MultiModal-CelebA256			MultiModal-CelebA512		
	FID^{5k}	FDD^{5k}	KID^{5k}	FID^{5k}	FDD^{5k}	KID^{5k}
Sweep	158.1	75.9	0.169	162.3	103.2	0.203
Zigzag-1	65.7	47.8	0.051	121.0	78.0	0.113
Zigzag-2	54.7	45.5	0.041	96.0	59.5	0.079
Zigzag-8	45.5	26.4	0.011	34.9	29.5	0.023

3. Experiment

In this section, we begin by detailing the experimental setup concerning image and video datasets, as well as our training details. Subsequently, we delve into several in-depth analyses aimed at elucidating the rationale behind our method design across various resolutions. Finally, we present our results obtained from higher-resolution, we defer more results on video in long version.

3.1. Ablation Study

Scan Scheme Ablation. We provide several important findings based on our ablation studies on MultiModal-CelebA dataset in various resolutions in Table 1. Firstly, switching the scanning scheme from sweep to zigzag led to some gains. Secondly, as we increased the zigzag scheme from 1 to 8, we saw consistent gains. This indicates that alternating the scanning scheme in various blocks can be beneficial. Finally, the relative gain between Zigzag-1 and Zigzag-8 is more prominent at higher resolutions (512×512, or longer sequence token number) compared to lower resolutions (256×256, or shorter sequence token number), this shows the great potential and more efficient inductive-bias incorporation in longer sequence number.

Table 2: **Main result on FacesHQ-1024 dataset with 4,094 tokens in latent space.** Our method can outperform the baseline and can achieve even better results when the training scale is increased.

Method	FID^{5k}	FDD^{5k}
Bidirection Mamba-16GPU (Zhu et al., 2024)	51.1	66.3
Zigzag-Mamba -16GPU	37.8	50.5
Zigzag-Mamba -32GPU	26.6	31.2

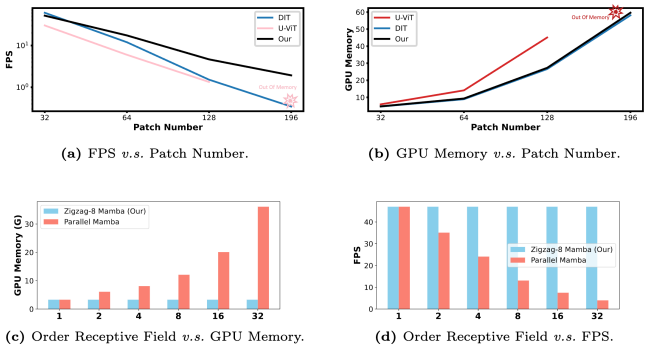
Ablation study about the Network and FPS/GPU-Memory. In Figure 4 (a,b),we analyze the forward speed and GPU memory usage while varying the global patch dimensions from 32×32 to 196×196. For the speed analysis, we report Frame Per Second (FPS) instead of FLOPS, as FPS provides a more explicit and appropriate evaluation of speed. For simplicity, we uniformly apply the zigzag-1 Mamba scan scheme and use batch size=1 and patch size=1 on an A100 GPU with 80GB memory. It's worth noting that all methods share nearly identical parameter numbers for fair comparison. We primarily compare our method with two popular transformer-based Diffusion backbones, U-ViT (Bao et al., 2023a) and DiT (Peebles & Xie, 2022). It is evident that our method achieves the best FPS and GPU utilization when gradually increasing the patching number. U-ViT demonstrates the worst performance, even exceeds the memory bounds when the patch number is 196. Surprisingly, DiT's GPU utilization is close to our method, which supports our backbone choice of DiT from a practical perspective.

Order Receptive Field. We propose a new concept in Mamba-based structure for multidimensional data. Given that various spatially-continuous zigzag paths may exist in multidimensional data, we introduce the term *Order Receptive Field* which denotes the number of zigzag paths *explicitly* employed in the network design.

Ablation study about the Order Receptive Field and FPS/GPU-Memory. As depicted in Figure 4 (c,d), Zigzag Mamba consistently maintains its GPU memory consumption and FPS rate, even with a gradually increasing Order Receptive Field. In contrast, our primary baseline, Parallel Mamba, along with variants like Bidirectional Mamba and Vision Mamba (Liu et al., 2024b; Zhu et al., 2024), experience a consistent decrease in FPS due to increased parameters. Notably, Zigzag Mamba, with an Order Receptive Field of 8, can perform faster without altering parameters.

3.2. Main Result

Main Result on 1024×1024 FacesHQ. To elaborate on the scalability of our method within the Mamba and Stochastic Interpolant framework, we provide comparisons on a high-resolution dataset (1024×1024 FacesHQ) in Table 2.



(a) FPS *v.s.* Patch Number. (b) GPU Memory *v.s.* Patch Number.

(c) Order Receptive Field *v.s.* GPU Memory. (d) Order Receptive Field *v.s.* FPS.

Figure 4: (a, b).GPU Memory usage and FPS between our method and transformer-based methods(U-VIT (Bao et al., 2023a) and DiT (Peebles & Xie, 2022)). (c). Order Receptive Field and GPU memory (d). Order Receptive Field and FPS. Order Receptive Field denotes how many scan paths we consider in our network design.

Our primary comparison is against Bidirectional Mamba, a commonly used solution for applying Mamba to 2D image data (Liu et al., 2024b; Zhu et al., 2024). With the aim of investigating Mamba's scalability in large resolutions up to 1,024, we employ the diffusion model on the latent space of 128×128 with a patch size of 2, resulting in 4,096 tokens. The network is trained on 16 A100 GPUs. Notably, our method demonstrates superior results compared to Bidirectional Mamba. Details regarding loss and FID curves can be found in long version. While constrained by GPU resource limitations, preventing longer training duration, we anticipate consistent outperformance of Bidirectional Mamba with extended training duration.

4. Conclusion

In this paper, we present the Zigzag Mamba Diffusion Model, developed within the Stochastic Interpolant framework. Our initial focus is on addressing the critical issue of spatial continuity. We then devise a Zigzag Mamba block to better utilize the inductive bias in 2D images. Further, we factorize the 3D Mamba into 2D and 1D Zigzag Mamba to facilitate optimization. We empirically design various ablation studies to examine different factors. This approach allows for a more in-depth exploration of the Stochastic Interpolant theory. We hope our endeavor can inspire further exploration in the Mamba network design. We anticipate that our scan path will be suitable for other linear attention models such as RWKV (Peng et al., 2024), xLSTM (Beck et al., 2024), HGRN (Qin et al., 2024), GLA (Yang et al., 2024a), and several others listed at FLA (Yang & Zhang, 2024)[2].

[2]https://github.com/sustcsonglin/flash-linear-attention

References

Albergo, M. S. and Vanden-Eijnden, E. Building normalizing flows with stochastic interpolants. *arXiv*, 2022.

Albergo, M. S., Boffi, N. M., and Vanden-Eijnden, E. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv*, 2023.

Ao, S., Zhao, W., Han, X., Yang, C., Liu, Z., Shi, C., Sun, M., Wang, S., and Su, T. Burstattention: An efficient distributed attention framework for extremely long sequences. *arXiv*, 2024.

Bao, F., Li, C., Cao, Y., and Zhu, J. All are worth words: a vit backbone for score-based diffusion models. *CVPR*, 2023a.

Bao, F., Nie, S., Xue, K., Li, C., Pu, S., Wang, Y., Yue, G., Cao, Y., Su, H., and Zhu, J. One transformer fits all distributions in multi-modal diffusion at scale. *arXiv*, 2023b.

Beck, M., Pöppel, K., Spanring, M., Auer, A., Prudnikova, O., Kopp, M., Klambauer, G., Brandstetter, J., and Hochreiter, S. xlstm: Extended long short-term memory, 2024.

Beltagy, I., Peters, M. E., and Cohan, A. Longformer: The long-document transformer. *arXiv*, 2020.

Brandon, W., Nrusimha, A., Qian, K., Ankner, Z., Jin, T., Song, Z., and Ragan-Kelley, J. Striped attention: Faster ring attention for causal transformers. *arXiv preprint arXiv:2311.09431*, 2023.

Chen, R. T., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. *NeurIPS*, 2018.

Child, R., Gray, S., Radford, A., and Sutskever, I. Generating long sequences with sparse transformers. *arXiv*, 2019.

Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, L., et al. Rethinking attention with performers. *arXiv*, 2020.

Dao, T., Fu, D., Ermon, S., Rudra, A., and Ré, C. Flashattention: Fast and memory-efficient exact attention with io-awareness. *NeurIPS*, 2022.

Fu, D. Y., Dao, T., Saab, K. K., Thomas, A. W., Rudra, A., and Ré, C. Hungry hungry hippos: Towards language modeling with state space models. *arXiv*, 2022.

Gong, J., Foo, L. G., Fan, Z., Ke, Q., Rahmani, H., and Liu, J. Diffpose: Toward more reliable 3d pose estimation. In *CVPR*, 2023.

Gu, A. and Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv*, 2023.

Gu, A., Goel, K., and Ré, C. Efficiently modeling long sequences with structured state spaces. 2021a.

Gu, A., Johnson, I., Goel, K., Saab, K., Dao, T., Rudra, A., and Ré, C. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *NeurIPS*, 2021b.

Gu, A., Goel, K., Gupta, A., and Ré, C. On the parameterization and initialization of diagonal state space models. *NeurIPS*, 2022.

Gupta, A., Gu, A., and Berant, J. Diagonal state spaces are as effective as structured state spaces. *NeurIPS*, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., and Fleet, D. J. Video diffusion models. In *ARXIV*, 2022.

Huang, Z., Zhou, P., Yan, S., and Lin, L. Scalelong: Towards more stable training of diffusion model via scaling network long skip connection. *NeurIPS*, 2024.

Kitaev, N., Kaiser, Ł., and Levskaya, A. Reformer: The efficient transformer. *arXiv*, 2020.

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. *ICLR*, 2023.

Liu, G.-H., Chen, T., So, O., and Theodorou, E. Deep generalized schrödinger bridge. *NeurIPS*, 2022.

Liu, H., Zaharia, M., and Abbeel, P. Ring attention with blockwise transformers for near-infinite context. *arXiv*, 2023a.

Liu, J., Yang, H., Zhou, H.-Y., Xi, Y., Yu, L., Yu, Y., Liang, Y., Shi, G., Zhang, S., Zheng, H., et al. Swin-umamba: Mamba-based unet with imagenet-based pretraining. *arXiv*, 2024a.

Liu, X., Gong, C., and Liu, Q. Flow straight and fast: Learning to generate and transfer data with rectified flow. *ICLR*, 2023b.

Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., and Liu, Y. Vmamba: Visual state space model. *arXiv*, 2024b.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021.

Ma, J., Li, F., and Wang, B. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv*, 2024a.

Ma, N., Goldstein, M., Albergo, M. S., Boffi, N. M., Vanden-Eijnden, E., and Xie, S. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. *arXiv*, 2024b.

McKenna, D. M. Hilbert curves: Outside-in and inside-gone. *Mathemaesthetics, Inc*, 2019.

OpenAI. Sora: Creating video from text, 2024. URL https://openai.com/sora.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. *arXiv*, 2022.

Peng, B., Goldstein, D., Anthony, Q., Albalak, A., Alcaide, E., Biderman, S., Cheah, E., Ferdinan, T., Hou, H., Kazienko, P., et al. Eagle and finch: Rwkv with matrix-valued states and dynamic recurrence. *arXiv preprint arXiv:2404.05892*, 2024.

Qin, Z., Yang, S., Sun, W., Shen, X., Li, D., Sun, W., and Zhong, Y. Hgrn2: Gated linear rnns with state expansion. *arXiv preprint arXiv:2404.07904*, 2024.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Smith, J. T., Warrington, A., and Linderman, S. W. Simplified state space layers for sequence modeling. *arXiv*, 2022.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021.

Sun, Z., Yang, Y., and Yoo, S. Sparse attention with learning to hash. In *ICLR*, 2021.

Tikochinski, R., Goldstein, A., Meiri, Y., Hasson, U., and Reichart, R. An incremental large language model for long text processing in the brain. 2024.

Wang, J., Yan, J. N., Gu, A., and Rush, A. M. Pretraining without attention. *arXiv*, 2022.

Wu, L., Wang, D., Gong, C., Liu, X., Xiong, Y., Ranjan, R., Krishnamoorthi, R., Chandra, V., and Liu, Q. Fast point cloud generation with straight flows. In *CVPR*, 2023.

Yan, J. N., Gu, J., and Rush, A. M. Diffusion models without attention. *arXiv*, 2023.

Yang, S. and Zhang, Y. Fla: A triton-based library for hardware-efficient implementations of linear attention mechanism, January 2024. URL https://github.com/sustcsonglin/flash-linear-attention.

Yang, S., Wang, B., Shen, Y., Panda, R., and Kim, Y. Gated linear attention transformers with hardware-efficient training. *ICML*, 2024a.

Yang, Y., Xing, Z., and Zhu, L. Vivim: a video vision mamba for medical video object segmentation. *arXiv*, 2024b.

Yu, A., Nigmetov, A., Morozov, D., Mahoney, M. W., and Erichson, N. B. Robustifying state-space models for long sequences via approximate diagonalization. *arXiv*, 2023.

Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., and Wang, X. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv*, 2024.

zhuzilin. Ring flash attention. https://github.com/zhuzilin/ring-flash-attention, 2024.