ZigMa: A DiT-style Zigzag Mamba Diffusion Model

Vincent Tao Hu, Stefan Andreas Baumann, Ming Gui,
Olga Grebenkova, Pingchuan Ma, Johannes Fischer, and Björn Ommer

CompVis @ LMU Munich, MCML
https://compvis.github.io/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, also this heterogeneous layerwise scan enables zero memory and speed burden when we consider more scan paths. 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.

Keywords: Diffusion Model · State-Space Model · Stochastic Interpolants

1 Introduction

Diffusion models have demonstrated significant advancements across various applications, including image processing [45,48,84], video analysis [44], point cloud processing [109], representation learning [30] and human pose estimation [32]. Many of these models are built upon Latent Diffusion Models (LDM) [84], which are typically based on the UNet backbone. However, scalability remains a significant challenge in LDMs [50]. Recently, transformer-based structures have gained popularity due to their scalability [9, 80] and effectiveness in multi-modal training [10]. Notably, the transformer-based structure DiT [80] has even contributed to enhancing the high-fidelity video generation model SORA [78] by OpenAI. Despite efforts to alleviate the quadratic complexity of the attention mechanism through techniques such as windowing [71], sliding [13], sparsification [19, 56],

hashing [20,93], Ring Attention [15,66], Flash Attention [23] or a combination of them [8,124], it remains a bottleneck for diffusion models.

On the other hand, State-Space Models [34,35,39] have demonstrated significant potential for long sequence modeling, rivaling transformer-based methods. Their biological similarity [95] and efficient memory state also advocate for the use of the State-Space model over the transformer. Several methods [29,33,35,88] have been proposed to enhance the robustness [116], scalability [33], and efficiency [35,36] of State-Space Models. Among these, a method called Mamba [33] 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 [70,123] 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 [67,73] 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 [3] provides a more generalized framework that can uniform various generative models including, Normalizing Flow [17], diffusion model [43,89,91], Flow matching [4,64,69], and Schrödinger Bridge [65]. Previously, some works [74] 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 heterogeneous layerwise scan 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   Related Works

Mamba. Several works [102,103,103] have demonstrated that the State-Space Model possesses universal approximation ability under certain conditions. Mamba, as a new State-Space Model, has superior potential for modeling long sequences efficiently, which has been explored in various fields such as medical imag-
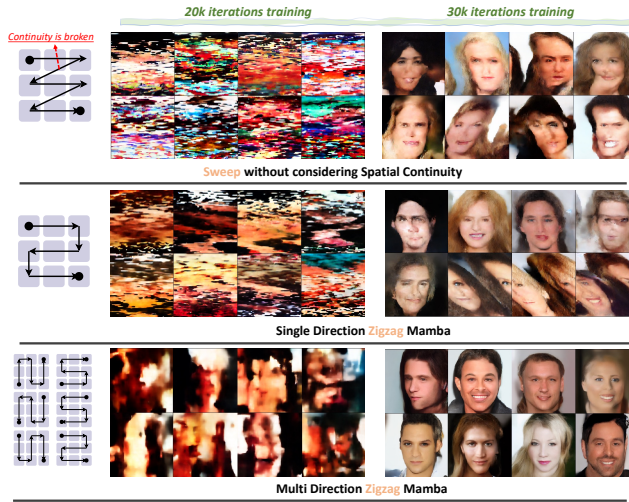
Figure 1: Motivation. Our Zigzag Mamba method improves the network's position-awareness by arranging and rearranging the scan path of Mamba in a heuristic manner.

ing [73, 86, 108, 111], video [58, 79], image restoration [38, 122], graphs [12], NLP word byte [100], tabular data [2], point clouds [61], human motion [106, 120], multi-task [62] and image generation [27]. Among them, the most related to us are VisionMamba [70, 123], S4ND [77] and Mamba-ND [59]. VisionMamba [70, 123] uses a bidirectional SSM in discriminative tasks which incurs a high computational cost. Our method applies a simple alternative mamba diffusion in generative models. S4ND [77] introduces local convolution into Mamba's reasoning process, moving beyond the use of only 1D data. Mamba-ND [59] takes multi-dimensionality into account in discriminative tasks, making use of various scans within a single block. In contrast, our focus is on distributing scan complexity across every layer of the network, thus maximizing the incorporation of inductive bias from visual data with zero parameter burden. Scan curve is an important direction in SSM, PointMamba [61] is a representative work that employs SSM with space curves (e.g., Hilbert) for point cloud analysis, achieving remarkable performance. In contrast with them, our preliminary results show that the Hilbert curve doesn't work well with our method (see Appendix), while our method can be regarded as the simplest Peano curve. For more information related to Mamba's work, please refer to the survey [105].

Backbones in Diffusion Models. Diffusion models primarily employ UNet-based [43, 84] and ViT-based [9, 80] backbones. While UNet is known for high memory demands [84], ViT benefits from scalability [18, 24] and multi-modal learning [10]. However, ViT's quadratic complexity limits visual token processing, prompting studies towards mitigating this issue [13, 23, 104]. Our work, inspired by Mamba [33], explores an SSM-based model as a generic diffusion backbone, retaining ViT's modality-agnostic and sequential modeling advantages.

Concurrently, DiffSSM [112] concentrates on unconditional and class conditioning within the S4 model [35]. DIS [27] mainly explores the state-space model on a relatively small resolution, which is not the exact focus of our work. Our work significantly differs from theirs as it primarily focuses on the backbone design using the Mamba block and extends it to text conditioning. Furthermore, we apply our method to more complex visual data.

SDE and ODE in Diffusion models. The realm of Score-based Generative Models encompasses significant contributions from foundational works such as Score Matching with Langevin Dynamics (SMLD) by Song et al. [90], and the advent of Diffusion Models with Denoising Score Matching (DDPMs) proposed by Ho et al. [43]. These methodologies operate within the framework of Stochastic Differential Equations (SDEs), a concept further refined in the research of Song et al. [91]. Recent research strides, as exemplified by Karras et al. [52] and Lee et al. [57], have showcased the efficacy of employing Ordinary Differential Equation (ODE) samplers for diffusion SDEs, offering significant reductions in sampling costs compared to traditional approaches that entail discretizing diffusion SDEs. Furthermore, within the domain of Flow Matching [64] and Rectified Flow [68], both SMLD and DDPMs emerge as specialized instances under distinct paths of the Probability Flow ODE framework [91], with broad applications in vision [22,28,49], depth [37], human motion [47], even language [46]. These models typically utilize velocity field parameterizations employing the linear interpolant, a concept that finds broader applications in the Stochastic Interpolant framework [3], with subsequent generalizations extending to manifold settings [14]. The SiT model [74] scrutinizes the interplay between interpolation methods in both sampling and training contexts, albeit in the context of smaller resolutions such as 512×512. Our research endeavors to extend these insights to a larger scale, focusing on the generalization capabilities for 2D images of 1024×1024 and 3D video data.

3   Method

In this section, we begin by providing background information on State-Space Models [34,35,39], with a particular focus on a special case known as Mamba [33]. 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.

3.1   Background: State-Space Models

State Space Models (SSMs) [34, 35, 39] have been proven to handle long-range dependencies theoretically and empirically [36] with linear scaling w.r.t sequence length. In their general form, a linear state space model can be written as follows:

$$x'(t) = \mathbf{A}(t)x(t) + \mathbf{B}(t)u(t)$$
$$y(t) = \mathbf{C}(t)x(t) + \mathbf{D}(t)u(t),$$

mapping a 1-D input sequence $u(t) \in \mathbb{R}$ to a 1-D output sequence $y(t) \in \mathbb{R}$ through an implicit N-D latent state sequence $x(t) \in \mathbb{R}^n$. Concretely, deep SSMs seek to use stacks of this simple model in a neural sequence modeling architecture, where the parameters $\mathbf{A}, \mathbf{B}, \mathbf{C}$ and $\mathbf{D}$ for each layer can be learned via gradient descent.
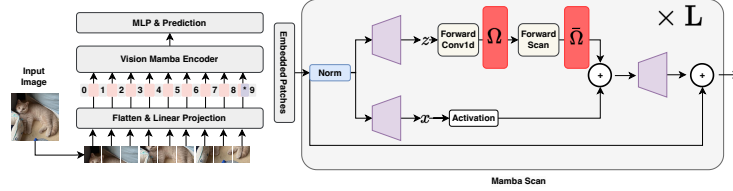
Figure 2: ZigMa. Our backbone is structured in L layers, mirroring the style of DiT [80]. We use the single-scan Mamba block as the primary reasoning module across different patches. To ensure the network is positionally aware, we've designed an arrange-rearrange scheme based on the single-scan Mamba. Different layers follow pairs of unique rearrange operation Ω and reverse rearrange $\bar{\Omega}$, optimizing the position-awareness of the method.

Recently, Mamba [33] largely improved the flexibility of SSMs in Language Modelling by relaxing the time-invariance constraint on SSM parameters, while maintaining computational efficiency. Several studies [70, 123] have been conducted to adapt the use of Mamba from unidimensional language data to multidimensional visual data. While most of these studies try to duplicate the $\mathbf{A}$ to facilitate the new (reversed) direction, this approach can lead to additional parameters and an increased memory burden. In this paper, we focus on exploring the scanning scheme of Mamba in diffusion models to efficiently maximize the use of inductive-bias from multi-dimensional visual data with zero parameter and memory burden.

3.2   Diffusion Backbone: Zigzag Mamba

DiT-Style Network. We opt to use the framework of DiT by AdaLN [80] rather than the skip-layer focused U-ViT structure [9], as DiT has been validated as a

scalable structure in literature [10, 18, 78]. Additionally, the Hourglass structure with downsampling [76, 85] requires selecting the depth and width based on the complexity of the dataset and task. This requirement limits the flexibility of the solution. Considering the aforementioned points, it informs our Mamba network design depicted in Figure 4. The core component of this design is the Zigzag Scanning, which will be explained in the following paragraph.

Zigzag Scanning in Mamba. Previous studies [101, 112] have used bidirectional scanning within the SSM framework. This approach has been expanded to include additional scanning directions [67, 70, 115] 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.

**(a)** sweep-scan   **(b)** zigzag-scan   **(c)** zigzag-scan with 8 schemes
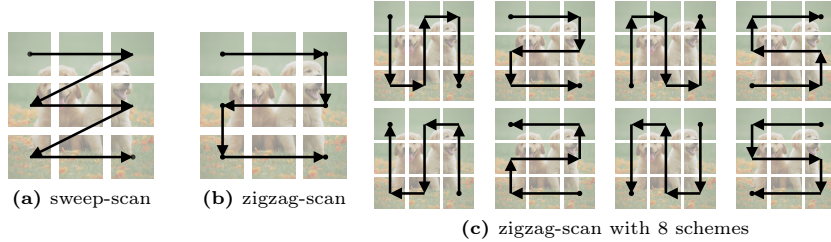
Figure 3: The 2D Image Scan. Our mamba scan design is based on the sweep-scan scheme shown in subfigure (a). From this, we developed a zigzag-scan scheme displayed in subfigure (b) to enhance the continuity of the patches, thereby maximizing the potential of the Mamba block. Since there are several possible arrangements for these continuous scans, we have listed the eight most common zigzag-scans in subfigure (c).

Now we explore the design of the Ω_i operation, considering additional inductive biases from 2D images. We propose one key properties: *Spatial Con-*

tinuity. Regarding Spatial Continuity, current innovations of Mamba in images [67, 70, 123] 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 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.

Heterogeneous Layerwise Scan. 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.
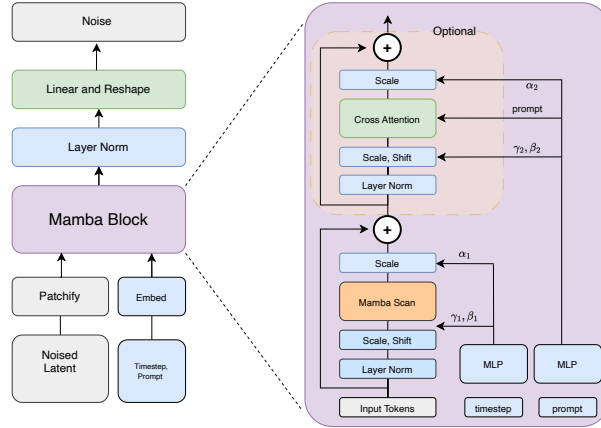
Figure 4: The Detail of our Zigzag Mamba block. The detail of Mamba Scan is shown in Figure 2. The condition can include a timestep and a text prompt. These are fed into an MLP, which separately modulates the Mamba scan for long sequence modeling and cross-attention for multi-modal reasoning.

Deploying text-condition on Zigzag Mamba. While Mamba offers the advantage of efficient long sequence modeling, it does so at the expense of the attention mechanism. As a result, there has been limited exploration into incorporating text-conditioning in Mamba-based diffusion models. To address this

---

[1] We also experimented with more complex continuous space-filling paths, such as the Hilbert space-filling curve [75]. However, empirical findings indicate that this approach may lead to deteriorated results. For further detailed comparisons, please refer to the Appendix.

gap, we propose a straightforward cross-attention block with skip layers built upon the Mamba block, as illustrated in Figure 4. This design not only enables long sequence modeling but also facilitates multi-token conditioning, such as text-conditioning. Furthermore, it has the potential to provide interpretability [16, 42, 94], as cross-attention has been utilized in diffusion models.

Generalizing to 3D videos by factorizing spatial and temporal information. In previous sections, our focus has been on the spatial 2D Mamba, where we designed several spatially continuous, space-filling 2D scanning schemes. In this section, we aim to leverage this experience to aid in designing corresponding mechanisms for 3D video processing. We commence our design process by extrapolating from the conventional directional Mamba, as depicted in Figure 5. Given a video feature input $\mathbf{z} \in \mathbb{R}^{B \times T \times C \times W \times H}$, we propose three variants of the Video Mamba Block for facilitating 3D video generation.

(a) Sweep-scan: In this approach, we directly flatten the 3D feature $\mathbf{z}$ without considering spatial or temporal continuity. It's worth noting that the flattening process follows the computer hierarchy order, meaning that no continuity is preserved in the flattened representation.

(b) 3D Zigzag: Compared with the formulation of the 2D zigzag in previous subsections, we follow the similar design to generalize it to 3D Zigzag to keep the continuity in 2D and 3D simultaneously. Potentially, the scheme has much more complexity. We heuristically list 8 schemes as well. However, we empirically find that this scheme will lead to suboptimal optimization.

(c) Factorized 3D Zigzag = 2D Zigzag + 1D Sweep: To address the suboptimal optimization issue, we propose to factorize the spatial and temporal correlations as separate Mamba blocks. The order of their application can be adjusted as desired, for example, "sstt" or "ststst", where "s" represents the spatial-zigzag Mamba and "t" represents the temporal-zigzag Mamba. For a 1D temporal sweep, we simply opt for forward and backward scanning, since there is only one dimension on the time axis.

Computation Analysis. For a visual sequence $\mathbf{T} \in \mathbb{R}^{1 \times M \times D}$, the computation complexity of global self-attention and k-direction mamba and our zigzag mamba are as follows:

$$\zeta(\text{self-attention}) = 4\text{MD}^2 + 2\text{M}^2\text{D}, \tag{4}$$

$$\zeta(\text{k-mamba}) = k \times [3\text{M}(2\text{D})\text{N} + \text{M}(2\text{D})\text{N}^2], \tag{5}$$

$$\zeta(\text{zigzag}) = 3\text{M}(2\text{D})\text{N} + \text{M}(2\text{D})\text{N}^2, \tag{6}$$

where self-attention exhibits quadratic complexity with respect to sequence length M, while Mamba exhibits linear complexity (N is a fixed parameter, set to 16 by default). Here, k represents the number of scan directions in a single Mamba block. Therefore, k-mamba and zigzag share linear complexity with respect to self-attention. Moreover, our zigzag method can eliminate the k series, further reducing the overall complexity.
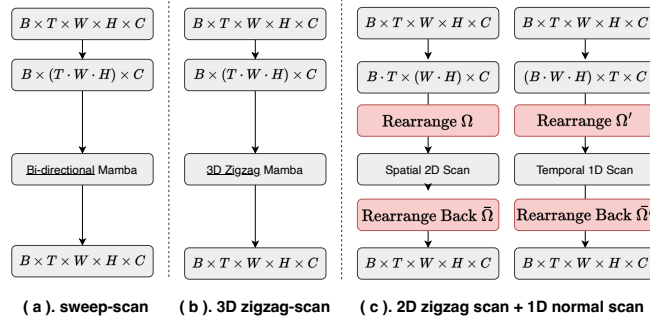
Figure 5: The 3D Video Scan. (a) We illustrate the bidirectional Mamba with the sweep scan, where the spatial and temporal information is treated as a set of tokens with a computer-hierarchy order. **(b)** For the 3D zigzag-scan, we aim to maximize the potential of Mamba by employing a spatial continuous scan scheme and adopting the optimal zigzag scan solution, as depicted in Figure 3. **(c)** We further separate the reasoning between spatial and temporal information, resulting in a factorized combination of 2D spatial scan (Ω) plus a 1D temporal scan (Ω') scheme.

Upon completing the design of the Zigzag Mamba network for improved visual inductive-bias integration, we proceed to combine it with a new diffusion framework, as illustrated below.

3.3   Diffusion Framework: Stochastic Interpolant

Sampling based on vector v and score s. Following [3, 96], the time-dependent probability distribution $p_t(\mathbf{x})$ of $\mathbf{x}_t$ also coincides with the distribution of the reverse-time SDE [6]:

$$d\mathbf{X}_t = \mathbf{v}(\mathbf{X}_t, t)dt + \frac{1}{2}w_t\mathbf{s}(\mathbf{X}_t, t)dt + \sqrt{w_t}d\bar{\mathbf{W}}_t, \tag{7}$$

where $\bar{\mathbf{W}}_t$ is a reverse-time Wiener process, $w_t > 0$ is an arbitrary time-dependent diffusion coefficient, $\mathbf{s}(\mathbf{x}, t) = \nabla \log p_t(\mathbf{x})$ is the score, and $\mathbf{v}(\mathbf{x}, t)$ is given by the conditional expectation

$$\begin{aligned}
\mathbf{v}(\mathbf{x}, t) &= \mathbb{E}[\dot{\mathbf{x}}_t | \mathbf{x}_t = \mathbf{x}], \\
&= \dot{\alpha}_t \mathbb{E}[\mathbf{x}_* | \mathbf{x}_t = \mathbf{x}] + \dot{\sigma}_t \mathbb{E}[\boldsymbol{\varepsilon} | \mathbf{x}_t = \mathbf{x}],
\end{aligned} \tag{8}$$

where α_t is a decreasing function of t, and σ_t is an increasing function of t. Here, $\dot{\alpha}_t$ and $\dot{\sigma}_t$ denote the time derivatives of α_t and σ_t, respectively.

As long as we can estimate the velocity $\mathbf{v}(\mathbf{x}, t)$ and/or score $\mathbf{s}(\mathbf{x}, t)$ fields, we can utilize it for the sampling process either by probability flow ODE [91] or the reverse-time SDE (7). Solving the reverse SDE (7) backwards in time from $\mathbf{X}_T = \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \mathbf{I})$ enables generating samples from the approximated data distribution $p_0(\mathbf{x}) \sim p(\mathbf{x})$. During sampling, we can perform direct sampling

from either ODE or SDEs to balance between sampling speed and fidelity. If we choose to conduct ODE sampling, we can achieve this simply by setting the noise term $\mathbf{s}$ to zero.

In [3], it shows that one of the two quantities $\mathbf{s}_\theta(\mathbf{x},t)$ and $\mathbf{v}_\theta(\mathbf{x},t)$ needs to be estimated in practice. This follows directly from the constraint

$$\begin{aligned}\mathbf{x} &= \mathbb{E}[\mathbf{x}_t|\mathbf{x}_t=\mathbf{x}],\\ &= \alpha_t\mathbb{E}[\mathbf{x}_*|\mathbf{x}_t=\mathbf{x}] + \sigma_t\mathbb{E}[\varepsilon|\mathbf{x}_t=\mathbf{x}],\end{aligned} \tag{9}$$

which can be used to re-express the score $\mathbf{s}(\mathbf{x},t)$ in terms of the velocity $\mathbf{v}(\mathbf{x},t)$ as

$$\mathbf{s}(\mathbf{x},t) = \sigma_t^{-1}\frac{\alpha_t\mathbf{v}(\mathbf{x},t) - \dot{\alpha}_t\mathbf{x}}{\dot{\alpha}_t\sigma_t - \alpha_t\dot{\sigma}_t}. \tag{10}$$

Thus, $\mathbf{v}(\mathbf{x},t)$ and $\mathbf{s}(\mathbf{x},t)$ can be mutually conversed. We illustrate how to compute them in the following.

Estimating the score s and the velocity v. It has been shown in score-based diffusion models [91] that the score can be estimated parametrically as $\mathbf{s}_\theta(\mathbf{x},t)$ using the loss

$$\mathcal{L}_\mathrm{s}(\theta) = \int_0^T \mathbb{E}[\|\sigma_t\mathbf{s}_\theta(\mathbf{x}_t,t) + \varepsilon\|^2]\mathrm{d}t. \tag{11}$$

Similarly, the velocity $\mathbf{v}(\mathbf{x},t)$ can be estimated parametrically as $\mathbf{v}_\theta(\mathbf{x},t)$ via the loss

$$\mathcal{L}_\mathrm{v}(\theta) = \int_0^T \mathbb{E}[\|\mathbf{v}_\theta(\mathbf{x}_t,t) - \dot{\alpha}_t\mathbf{x}_* - \dot{\sigma}_t\varepsilon\|^2]\mathrm{d}t, \tag{12}$$

where θ represents the Zigzag Mamba network that we described in the previous section, we adopt the linear path for training, due to its simplicity and relatively straight trajectory:

$$\alpha_t = 1 - t, \qquad \sigma_t = t. \tag{13}$$

We note that any time-dependent weight can be included under the integrals in both (11) and (12). These weight factors play a crucial role in score-based models when T becomes large [54, 55]. Thus, they provide a general form that considers both the time-dependent weight and the stochasticity.

4   Experiment

4.1   Dataset and Training Detail

Image Dataset. To explore the scalability in high resolution, we conduct experiments on the FacesHQ 1024×1024. The general dataset that we use for training and ablations is FacesHQ, a compilation of CelebA-HQ [110] and FFHQ [53], as employed in previous work such as [26, 28].

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-CelebA-256			MultiModal-CelebA-512		
	$FID^{5k} \downarrow$	$FDD^{5k} \downarrow$	$KID^{5k} \downarrow$	$FID^{5k} \downarrow$	$FDD^{5k} \downarrow$	$KID^{5k} \downarrow$
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

Video Dataset. UCF101 dataset consists of 13,320 video clips, which are classified into 101 categories. The total length of these video clips is over 27 hours. All these videos are collected from YouTube and have a fixed frame rate of 25 FPS with the resolution of 320×240. We randomly sample continuous 16 frames and resize the frames to 256×256.
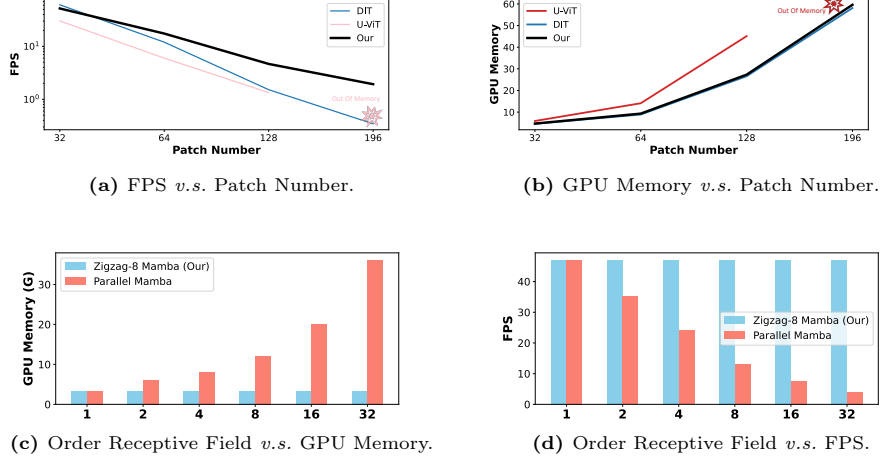
Training Details. We uniformly use AdamW [72] optimizer with $1e-4$ learning rate. For extracting latent features, we employ off-the-shelf VAE encoders. To mitigate computational costs, we adopted a mixed-precision training approach. Additionally, we applied gradient clipping with a threshold of 2.0 and a weight decay of 0.01 to prevent NaN occurrences during Mamba training. Most of our experiments were conducted on 4 A100 GPUs, with scalability exploration extended to 16 and 32 A100 GPUs. For sampling, we adopt the ODE sampling for speed consideration. For further details, please refer to the Appendix 8.8.

4.2   Ablation Study

Table 2: Ablation about Position Embedding (PE) on unconditional CelebA dataset (256^2). To better abate PE and eliminate the conditional signal's influence, we use an unconditional dataset.

FID/FDD $\downarrow$	No PE	Cosine PE	Learnable PE
VisionMamba [123]	21.33/21.00	18.47/19.90	16.38/18.20
ZigMa	14.27/18.00	14.04/17.91	13.32/17.40

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)

(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 6: (a, b).GPU Memory usage and FPS between our method and transformer-based methods(U-VIT [9] and DiT [80]). (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.

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.

Ablation about Position Embedding. As shown in Table 2, the learnable embedding performs better than the Sinusoidal embedding, which in turn performs better than no position embedding. In various cases, our zigzag method surpasses the baselines. Notably, our performance remains almost unchanged whether we use the Sinusoidal position embedding or no position embedding. This suggests that our method can better incorporate spatial inductive-bias compared to our baseline. Finally, using the learnable position embedding provides further, albeit marginal, gains suggesting that better position embedding exists even within our zigzag scan scheme. We find that [79] shares the same conclusion as us in video-related tasks.

Ablation study about the Network and FPS/GPU-Memory. In Figure 6 (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 [2]. 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 [9] and DiT [80]. It is evident that our method achieves the best FPS and GPU

[2] https://github.com/state-spaces/mamba/issues/110#issuecomment-1916464012

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.

Table 3: Main result on FacesHQ-1024 dataset with 4,094 tokens in latent space and bs=512. Our method can outperform the baseline and can achieve even better results when the training scale is increased.

Method	$\mathbf{FID}^{5k}\downarrow$	$\mathbf{FDD}^{5k}\downarrow$
VisionMamba [123]	51.1	66.3
ZigMa	37.8	50.5
ZigMa bs $\times$ 2	26.6	31.2

Table 4: Main Results on MS-COCO dataset with bs=256. Our method consistently outperforms the baseline. ZigMa with 8 scans performs much better compared with the baseline.

Method	$\mathbf{FID}^{5k}\downarrow$
Sweep	195.1
Zigzag-1	73.1
VisionMamba [123]	60.2
Zigzag-8	41.8

Table 5: Transformer-based methods comparison on unconditional CelebA256.

Method	FID$\downarrow$	Memory(G) $\downarrow$	FLOPS(G) $\downarrow$
U-ViT	14.50	35.10	12.5
DiT	14.64	29.20	5.5
ZigMa	14.27	17.80	5.2

Table 6: Video Scan Scheme on UCF101 dataset with bs=32.

Method	Frame-FID$^{5k}\downarrow$	FVD$^{5k}\downarrow$
Bidirection [123]	256.1	320.2
3D Zigzag	238.1	282.3
Our	216.1	210.2
Bidirection [123] bs$\times$4	146.2	201.1
ZigMa bs$\times$4	121.2	140.1

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 Fig. 6 (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 [70, 123], 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.

Comparison with transformer-based methods. We show the result in Table 5 on unconditional generation task.Our method achieves performance comparable to Transformer-based methods, with significantly less memory consumption and fewer FLOPS.

4.3   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 3. Our primary comparison is against Bidirectional Mamba, a commonly used solution for applying Mamba to 2D image data [70, 123]. 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, FID curves, and visualization can be found in the Appendix. While constrained by GPU resource limitations, preventing longer training duration, we anticipate consistent outperformance of Bidirectional Mamba with extended training duration.

COCO dataset. To further compare the performance of our method, we also evaluate it on the more complex and common dataset MS COCO. We compare with the Bidirection Mamba as the baseline in Table 4. It should be noted that all methods share nearly identical parameter numbers for fair comparison. We trained all methods using 16 A100 GPUs. please check Appendix 8.8 for details. As depicted in Table 4, our Zigzag-8 method outperforms Bidirectional Mamba as well as Zigzag-1. This suggests that amortizing various scanning schemes can yield significant improvements, attributed to better incorporation of the inductive bias for 2D images in Mamba.

UCF101 dataset. In Table 6, we present our results on the UCF101 dataset, training all methods using 4 A100 GPUs, with further scalability exploration conducted using 16 A100 GPUs. We mainly compare our method consistantly with Vision Mamba [123]. For the choice of the 3D Zigzag Mamba, please refer to Appendix 8.8. For Factorized 3D Zigzag Mamba in video processing, we deploy the *sst* scheme for factorizing spatial and temporal modeling. This scheme prioritizes spatial information complexity over temporal information, hypothesizing that redundancy exists in the temporal domain. Our results consistently demonstrate the superior performance of our method across various scenarios, underscoring the intricacy and effectiveness of our approach.

5   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 with heterogeneous layerwise scan 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.

Acknowledgements

References

1. Agarwal, N., Suo, D., Chen, X., Hazan, E.: Spectral state space models. arXiv (2023) 28

2. Ahamed, M.A., Cheng, Q.: Mambatab: A simple yet effective approach for handling tabular data. arXiv (2024) 3, 28

3. Albergo, M.S., Boffi, N.M., Vanden-Eijnden, E.: Stochastic interpolants: A unifying framework for flows and diffusions. arXiv (2023) 2, 4, 9, 10

4. Albergo, M.S., Vanden-Eijnden, E.: Building normalizing flows with stochastic interpolants. arXiv (2022) 2

5. Ali, A., Zimerman, I., Wolf, L.: The hidden attention of mamba models. arXiv (2024) 28

6. Anderson, B.D.: Reverse-time diffusion equation models. Stochastic Processes and their Applications (1982) 9

7. Anthony, Q., Tokpanov, Y., Glorioso, P., Millidge, B.: Blackmamba: Mixture of experts for state-space models. arXiv (2024) 28

8. Ao, S., Zhao, W., Han, X., Yang, C., Liu, Z., Shi, C., Sun, M., Wang, S., Su, T.: Burstattention: An efficient distributed attention framework for extremely long sequences. arXiv (2024) 2

9. Bao, F., Li, C., Cao, Y., Zhu, J.: All are worth words: a vit backbone for score-based diffusion models. CVPR (2023) 1, 3, 5, 12, 23

10. Bao, F., Nie, S., Xue, K., Li, C., Pu, S., Wang, Y., Yue, G., Cao, Y., Su, H., Zhu, J.: One transformer fits all distributions in multi-modal diffusion at scale. arXiv (2023) 1, 3, 6

11. Beck, M., Pöppel, K., Spanring, M., Auer, A., Prudnikova, O., Kopp, M., Klambauer, G., Brandstetter, J., Hochreiter, S.: xlstm: Extended long short-term memory (2024) 22

12. Behrouz, A., Hashemi, F.: Graph mamba: Towards learning on graphs with state space models. arXiv (2024) 3, 28

13. Beltagy, I., Peters, M.E., Cohan, A.: Longformer: The long-document transformer. arXiv (2020) 1, 3

14. Ben-Hamu, H., Cohen, S., Bose, J., Amos, B., Grover, A., Nickel, M., Chen, R.T., Lipman, Y.: Matching normalizing flows and probability paths on manifolds. In: ICML (2022) 4

15. Brandon, W., Nrusimha, A., Qian, K., Ankner, Z., Jin, T., Song, Z., Ragan-Kelley, J.: Striped attention: Faster ring attention for causal transformers. arXiv preprint arXiv:2311.09431 (2023) 2

16. Chefer, H., Gur, S., Wolf, L.: Transformer interpretability beyond attention visualization. In: CVPR (2021) 8
17. Chen, R.T., Rubanova, Y., Bettencourt, J., Duvenaud, D.K.: Neural ordinary differential equations. NeurIPS (2018) 2
18. Chen, S., Xu, M., Ren, J., Cong, Y., He, S., Xie, Y., Sinha, A., Luo, P., Xiang, T., Perez-Rua, J.M.: Gentron: Delving deep into diffusion transformers for image and video generation. arXiv (2023) 3, 6
19. Child, R., Gray, S., Radford, A., Sutskever, I.: Generating long sequences with sparse transformers. arXiv (2019) 1
20. 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) 2
21. Crowson, K., Baumann, S.A., Birch, A., Abraham, T.M., Kaplan, D.Z., Shippole, E.: Scalable high-resolution pixel-space image synthesis with hourglass diffusion transformers. arXiv (2024) 29
22. Dao, Q., Phung, H., Nguyen, B., Tran, A.: Flow matching in latent space. arXiv (2023) 4
23. Dao, T., Fu, D., Ermon, S., Rudra, A., Ré, C.: Flashattention: Fast and memory-efficient exact attention with io-awareness. NeurIPS (2022) 2, 3
24. Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., Steiner, A.P., Caron, M., Geirhos, R., Alabdulmohsin, I., et al.: Scaling vision transformers to 22 billion parameters. In: ICML (2023) 3
25. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021) 23, 27
26. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: CVPR (2021) 10
27. Fei, Z., Fan, M., Yu, C., Huang, J.: Scalable diffusion models with state space backbone. arXiv (2024) 3, 4, 28
28. Fischer, J.S., Gui, M., Ma, P., Stracke, N., Baumann, S.A., Ommer, B.: Boosting latent diffusion with flow matching. ECCV (2024) 4, 10
29. Fu, D.Y., Dao, T., Saab, K.K., Thomas, A.W., Rudra, A., Ré, C.: Hungry hungry hippos: Towards language modeling with state space models. arXiv (2022) 2
30. Fuest, M., Ma, P., Gui, M., Fischer, J.S., Hu, V.T., Ommer, B.: Diffusion models and representation learning: A survey. arXiv preprint arXiv:2407.00783 (2024) 1
31. Gong, H., Kang, L., Wang, Y., Wan, X., Li, H.: nnmamba: 3d biomedical image segmentation, classification and landmark detection with state space model. arXiv (2024) 28
32. Gong, J., Foo, L.G., Fan, Z., Ke, Q., Rahmani, H., Liu, J.: Diffpose: Toward more reliable 3d pose estimation. In: CVPR (2023) 1
33. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. CoLM (2024) 2, 3, 4, 5
34. Gu, A., Goel, K., Gupta, A., Ré, C.: On the parameterization and initialization of diagonal state space models. NeurIPS (2022) 2, 4, 5
35. Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces (2021) 2, 4, 5
36. Gu, A., Johnson, I., Goel, K., Saab, K., Dao, T., Rudra, A., Ré, C.: Combining recurrent, convolutional, and continuous-time models with linear state space layers. NeurIPS (2021) 2, 5

37. Gui, M., Fischer, J.S., Prestel, U., Ma, P., Kotovenko, D., Grebenkova, O., Baumann, S.A., Hu, V.T., Ommer, B.: Depthfm: Fast monocular depth estimation with flow matching. arXiv preprint arXiv:2403.13788 (2024) 4

38. Guo, H., Li, J., Dai, T., Ouyang, Z., Ren, X., Xia, S.T.: Mambair: A simple baseline for image restoration with state-space model. arXiv (2024) 3, 28

39. Gupta, A., Gu, A., Berant, J.: Diagonal state spaces are as effective as structured state spaces. NeurIPS (2022) 2, 4, 5

40. He, W., Han, K., Tang, Y., Wang, C., Yang, Y., Guo, T., Wang, Y.: Densemamba: State space models with dense hidden connection for efficient large language models. arXiv (2024) 28

41. He, X., Cao, K., Yan, K., Li, R., Xie, C., Zhang, J., Zhou, M.: Pan-mamba: Effective pan-sharpening with state space model. arXiv (2024) 28

42. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv (2022) 8

43. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020) 2, 3, 4

44. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. In: ARXIV (2022) 1

45. Hu, V.T., Chen, Y., Caron, M., Asano, Y.M., Snoek, C.G., Ommer, B.: Guided diffusion from self-supervised diffusion features. In: ARXIV (2023) 1

46. Hu, V.T., Wu, D., Asano, Y., Mettes, P., Fernando, B., Ommer, B., Snoek, C.: Flow matching for conditional text generation in a few sampling steps pp. 380–392 (2024) 4

47. Hu, V.T., Yin, W., Ma, P., Chen, Y., Fernando, B., Asano, Y.M., Gavves, E., Mettes, P., Ommer, B., Snoek, C.G.: Motion flow matching for human motion synthesis and editing. In: ARXIV (2023) 4

48. Hu, V.T., Zhang, D.W., Asano, Y.M., Burghouts, G.J., Snoek, C.G.M.: Self-guided diffusion models. In: CVPR (2023) 1

49. Hu, V.T., Zhang, D.W., Mettes, P., Tang, M., Zhao, D., Snoek, C.G.: Latent space editing in transformer-based flow matching. In: ICML 2023 Workshop, New Frontiers in Learning, Control, and Dynamical Systems (2023) 4

50. Huang, Z., Zhou, P., Yan, S., Lin, L.: Scalelong: Towards more stable training of diffusion model via scaling network long skip connection. NeurIPS (2024) 1

51. Huang, Z., Ben, Y., Luo, G., Cheng, P., Yu, G., Fu, B.: Shuffle transformer: Rethinking spatial shuffle for vision transformer. arXiv preprint arXiv:2106.03650 (2021) 29

52. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. In: NeurIPS (2022) 4

53. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: CVPR (2019) 10

54. Kingma, D., Salimans, T., Poole, B., Ho, J.: Variational diffusion models. In: NeurIPS (2021) 10

55. Kingma, D.P., Gao, R.: Understanding the diffusion objective as a weighted integral of elbos. arXiv (2023) 10

56. Kitaev, N., Kaiser, Ł., Levskaya, A.: Reformer: The efficient transformer. arXiv (2020) 1

57. Lee, S., Kim, B., Ye, J.C.: Minimizing trajectory curvature of ode-based generative models. ICML (2023) 4

58. Li, K., Li, X., Wang, Y., He, Y., Wang, Y., Wang, L., Qiao, Y.: Videomamba: State space model for efficient video understanding. ECCV (2024) 3

59. Li, S., Singh, H., Grover, A.: Mamba-nd: Selective state space modeling for multi-dimensional data. arXiv (2024) 3, 28, 29
60. Li, Y., Bornschein, J., Chen, T.: Denoising autoregressive representation learning. arXiv preprint arXiv:2403.05196 (2024) 29
61. Liang, D., Zhou, X., Wang, X., Zhu, X., Xu, W., Zou, Z., Ye, X., Bai, X.: Point-mamba: A simple state space model for point cloud analysis. arXiv preprint arXiv:2402.10739 (2024) 3, 27, 28
62. Lin, B., Jiang, W., Chen, P., Zhang, Y., Liu, S., Chen, Y.C.: Mtmamba: Enhancing multi-task dense scene understanding by mamba-based decoders. ECCV (2024) 3
63. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014) 30
64. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. ICLR (2023) 2, 4
65. Liu, G.H., Chen, T., So, O., Theodorou, E.: Deep generalized schrödinger bridge. NeurIPS (2022) 2
66. Liu, H., Zaharia, M., Abbeel, P.: Ring attention with blockwise transformers for near-infinite context. arXiv (2023) 2
67. 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 pre-training. arXiv (2024) 2, 6, 7
68. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv (2022) 4
69. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. ICLR (2023) 2
70. Liu, Y., Tian, Y., Zhao, Y., Yu, H., Xie, L., Wang, Y., Ye, Q., Liu, Y.: Vmamba: Visual state space model. arXiv (2024) 2, 3, 5, 6, 7, 13, 14, 28, 29
71. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV (2021) 1
72. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019) 11
73. Ma, J., Li, F., Wang, B.: U-mamba: Enhancing long-range dependency for biomedical image segmentation. arXiv (2024) 2, 3, 28
74. Ma, N., Goldstein, M., Albergo, M.S., Boffi, N.M., Vanden-Eijnden, E., Xie, S.: Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. arXiv (2024) 2, 4
75. McKenna, D.M.: Hilbert curves: Outside-in and inside-gone. Mathemaesthetics, Inc (2019) 7, 26
76. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: ECCV (2016) 6
77. Nguyen, E., Goel, K., Gu, A., Downs, G., Shah, P., Dao, T., Baccus, S., Ré, C.: S4nd: Modeling images and videos as multidimensional signals with state spaces. NeurIPS (2022) 3, 28, 29
78. OpenAI: Sora: Creating video from text (2024), https://openai.com/sora 1, 6
79. Park, J., Kim, H.S., Ko, K., Kim, M., Kim, C.: Videomamba: Spatio-temporal selective state space model. ECCV (2024) 3, 12
80. Peebles, W., Xie, S.: Scalable diffusion models with transformers. arXiv (2022) 1, 3, 5, 12, 23

81. 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) 22

82. Qin, Z., Yang, S., Sun, W., Shen, X., Li, D., Sun, W., Zhong, Y.: Hgrn2: Gated linear rnns with state expansion. arXiv preprint arXiv:2404.07904 (2024) 22

83. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021) 30

84. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022) 1, 3, 30

85. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015) 6

86. Ruan, J., Xiang, S.: Vm-unet: Vision mamba unet for medical image segmentation. arXiv (2024) 3, 28

87. Skorokhodov, I., Sotnikov, G., Elhoseiny, M.: Aligning latent and image spaces to connect the unconnectable. In: ICCV (2021) 34

88. Smith, J.T., Warrington, A., Linderman, S.W.: Simplified state space layers for sequence modeling. arXiv (2022) 2

89. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: ICML (2015) 2

90. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. arXiv (2019) 4

91. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: ICLR (2021) 2, 4, 9, 10

92. Stein, G., Cresswell, J., Hosseinzadeh, R., Sui, Y., Ross, B., Villecroze, V., Liu, Z., Caterini, A.L., Taylor, E., Loaiza-Ganem, G.: Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. NeurIPS (2023) 29

93. Sun, Z., Yang, Y., Yoo, S.: Sparse attention with learning to hash. In: ICLR (2021) 2

94. Tang, R., Liu, L., Pandey, A., Jiang, Z., Yang, G., Kumar, K., Stenetorp, P., Lin, J., Ture, F.: What the daam: Interpreting stable diffusion using cross attention. arXiv (2022) 8

95. Tikochinski, R., Goldstein, A., Meiri, Y., Hasson, U., Reichart, R.: An incremental large language model for long text processing in the brain (2024) 2

96. Tong, A., Malkin, N., Fatras, K., Atanackovic, L., Zhang, Y., Huguet, G., Wolf, G., Bengio, Y.: Simulation-free schr\" odinger bridges via score and flow matching. arXiv (2023) 9

97. Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Fvd: A new metric for video generation. ICLR Workshop (2019) 30

98. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017) 27

99. Wang, C., Tsepa, O., Ma, J., Wang, B.: Graph-mamba: Towards long-range graph sequence modeling with selective state spaces. arXiv (2024) 28

100. Wang, J., Gangavarapu, T., Yan, J.N., Rush, A.M.: Mambabyte: Token-free selective state space model. arXiv (2024) 3, 28

101. Wang, J., Yan, J.N., Gu, A., Rush, A.M.: Pretraining without attention. arXiv (2022) 6

102. Wang, S., Li, Q.: Stablessm: Alleviating the curse of memory in state-space models through stable reparameterization. arXiv (2023) 2, 28

103. Wang, S., Xue, B.: State-space models with layer-wise nonlinearity are universal approximators with exponential decaying memory. NeurIPS (2024) 2, 28

104. Wang, W., Ma, S., Xu, H., Usuyama, N., Ding, J., Poon, H., Wei, F.: When an image is worth 1,024 x 1,024 words: A case study in computational pathology. arXiv (2023) 3

105. Wang, X., Wang, S., Ding, Y., Li, Y., Wu, W., Rong, Y., Kong, W., Huang, J., Li, S., Yang, H., Wang, Z., Jiang, B., Li, C., Wang, Y., Tian, Y., Tang, J.: State space model for new-generation network alternative to transformers: A survey (2024) 3

106. Wang, X., Kang, Z., Mu, Y.: Text-controlled motion mamba: Text-instructed temporal grounding of human motion. arXiv preprint arXiv:2404.11375 (2024) 3

107. Wang, Z., Ma, C.: Semi-mamba-unet: Pixel-level contrastive cross-supervised visual mamba-based unet for semi-supervised medical image segmentation. arXiv (2024) 28

108. Wang, Z., Zheng, J.Q., Zhang, Y., Cui, G., Li, L.: Mamba-unet: Unet-like pure visual mamba for medical image segmentation. arXiv (2024) 3, 28

109. Wu, L., Wang, D., Gong, C., Liu, X., Xiong, Y., Ranjan, R., Krishnamoorthi, R., Chandra, V., Liu, Q.: Fast point cloud generation with straight flows. In: CVPR (2023) 1

110. Xia, W., Yang, Y., Xue, J.H., Wu, B.: Tedigan: Text-guided diverse face image generation and manipulation. In: CVPR (2021) 10, 30

111. Xing, Z., Ye, T., Yang, Y., Liu, G., Zhu, L.: Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. arXiv (2024) 3, 28

112. Yan, J.N., Gu, J., Rush, A.M.: Diffusion models without attention. arXiv (2023) 4, 6

113. Yang, S., Wang, B., Shen, Y., Panda, R., Kim, Y.: Gated linear attention transformers with hardware-efficient training. ICML (2024) 22

114. Yang, S., Zhang, Y.: Fla: A triton-based library for hardware-efficient implementations of linear attention mechanism (Jan 2024), https://github.com/sustcsonglin/flash-linear-attention 22

115. Yang, Y., Xing, Z., Zhu, L.: Vivim: a video vision mamba for medical video object segmentation. arXiv (2024) 6

116. Yu, A., Nigmetov, A., Morozov, D., Mahoney, M.W., Erichson, N.B.: Robustifying state-space models for long sequences via approximate diagonalization. arXiv (2023) 2

117. Yu, S., Sohn, K., Kim, S., Shin, J.: Video probabilistic diffusion models in projected latent space. In: CVPR (2023) 30

118. Zhang, T., Li, X., Yuan, H., Ji, S., Yan, S.: Point could mamba: Point cloud learning via state space model. arXiv (2024) 28

119. Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: CVPR (2018) 29

120. Zhang, Z., Liu, A., Reid, I., Hartley, R., Zhuang, B., Tang, H.: Motion mamba: Efficient and long sequence motion generation with hierarchical and bidirectional selective ssm. ECCV (2024) 3

121. Zhang, Z., Liu, A., Reid, I., Hartley, R., Zhuang, B., Tang, H.: Motion mamba: Efficient and long sequence motion generation with hierarchical and bidirectional selective ssm. arXiv (2024) 28

122. Zheng, Z., Wu, C.: U-shaped vision mamba for single image dehazing. arXiv (2024) 3, 28

123. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: Efficient visual representation learning with bidirectional state space model. ICML (2024) 2, 3, 5, 7, 11, 13, 14, 28

124. zhuzilin: Ring flash attention. https://github.com/zhuzilin/ring-flash-attention (2024) 2